DISSERTATION *submitted to the*

Combined Faculties for the
Natural Sciences and for Mathematics

of the

Ruperto-Carola University of Heidelberg,
Germany

for the degree of

Doctor of Natural Sciences

presented by

Lei Liu

born in: Hunan, P.R. China

oral examination date: 27.01.2015

Multiscale Modelling of CTCF and its Complexes

Referees: Prof. Dr. Dieter W. Heermann

Prof. Dr. Michael Hausmann

Abstract

More and more experiments show that the CCCTC-binding factor (CTCF), a multi-Cys_2His_2 (mC_2H_2) zinc finger protein, plays a key role in the spatial organization of chromatin and gene regulation in the nucleus of eukaryotic cells. In this context an important problem is to uncover the underlying mechanism of how CTCF shapes the chromatin structure. In this thesis, models on different scales, from atomistic scale to coarse-grained scale, are studied to better understand the conformational and dynamical properties of both the unbound CTCF and CTCF-DNA complexes.

Using homology modeling, an atomistic model of CTCF is constructed to study the conformational properties of unbound mC_2H_2 zinc finger proteins. To enhance the computing and sampling efficiency an atomistic pivoting algorithm and a mesoscale model for mC_2H_2 proteins is developed. It is shown that the conformations of unbound mC_2H_2 proteins, like CTCF, can be explained with a worm-like chain model. For proteins of a few zinc finger, an effective bending constraint favors an extended conformation, which is consistent with experimental findings. A self-avoiding chain model applies only to proteins of more than nine zinc fingers.

As a subsequent step, a mesoscale model is designed to study how a mC_2H_2 zinc finger protein binds to and searches for its target DNA loci. Statistical sequence-dependent interactions between the proteins and DNA are derived. Molecular dynamics simulations of this model reproduce several kinetic properties of mC_2H_2 zinc finger proteins, such as the rotation coupled sliding, the asymmetrical roles of different zinc fingers and the partial binding partial dangling mode. An application to CTCF in complexes with one of its target DNA

duplex shows that CTCF binds to DNA only by using its central zinc fingers. It asymmetrically bends the DNA duplex but does not form DNA loops. Other CTCF-assisted DNA looping mechanisms, like a bridged DNA loop organized by a CTCF homodimer, could be further studied with this model.

Motivated by the non-covalent binding of polypeptides to DNA, I study the adsorption of a flexible polymer to a rigid polymer with periodic binding sites, both in $2d$ and in $3d$. Analysis of Monte Carlo simulation results show that the phase transition, from non-adsorbed to adsorbed with increasing adsorbing strength, is a second order transition in $2d$, and higher order transition in $3d$. Compared to the adsorbed monomers, successive non-adsorbed monomers contribute more to the winding of the flexible polymer around a rigid polymer, showing the importance of the linkers in mC_2H_2 zinc finger proteins to wrap around DNA.

Zusammenfassung

Immer mehr Experimente zeigen, dass der CCCTC-Bindungsfaktor (CTCF), ein multi-Cys_2His_2 (mC_2H_2) Zinkfingerprotein, eine Schlüsselrolle in der räumlichen Anordung von Chromatin und der Genregulation im Zellkern eukaryotischer Zellen spielt. Ein bedeutendes Problem in diesem Zusammenhang besteht darin, herauszufinden, welchen Einfluss CTCF auf die Chromatinstruktur hat. In dieser Arbeit werden sowohl atomistische als auch sogenannte coarse-grained Simulationen verwendet, um die Konformationseigenschaften sowie die Dynamik von ungebundenem CTCF und CTCF-DNA Komplexen besser zu verstehen.

Mittels Homology Modelling wurde ein atomistisches Modell für CTCF erstellt, um die Konformationseigenschaften von ungebundenen mC_2H_2 Zinkfingerproteinen zu untersuchen. Um die Effizienz der Berechnung sowie des Samplings zu steigern, wurde sowohl ein Pivoting Algorithmus als auch ein Mesoskalenmodell für mC_2H_2-Proteine entwickelt. In einem weiteren Schritt wird gezeigt, dass sich die Konformationen von ungebundenen mC_2H_2-Proteinen, wie CTCF, durch ein „Worm-like Chain“ - Modell beschreiben lassen. Bei Proteinen mit nur wenigen Zinkfingern führt eine effektive Biegesteifigkeit zu gestreckten Konformationen, die auch in Experimenten beoachtet werden. Lediglich Proteine mit mehr als neun Zinkfingern lassen sich mit dem „Self-avoiding Chain“ - Modell beschreiben.

In einem weiteren Schritt wird mithilfe eines Mesoskalenmodells untersucht, wie ein mC_2H_2-Zinkfingerprotein seinen Ziel-DNA-Abschnitt findet und an diesen bindet. Dabei werden statistische, von der jeweiligen DNA-Sequenz abhängige,

Wechselwirkungen zwischen den Proteinen und der DNA abgeleitet. Mittels auf diesem Modell basierenden Molekulardynamik-Simulationen lassen sich wichtige kinetische Eigenschaften von mC_2H_2-Zinkfingerproteinen reproduzieren. Aus der Untersuchung von Komplexen, die aus CTCF und einem der Ziel-DNA-Loci bestehen, geht hervor, dass CTCF nur mittels der zentralen Zinkfinger an die DNA bindet. Es biegt den DNA-Doppelstrang asymmetrisch ohne Schleifen zu bilden. Mithilfe dieses Modells ist es möglich, auch andere Mechanismen zur DNA-Schleifenbildung, an denen CTCF beteiligt ist, zu analysieren.

Da Polypeptide nichtkovalente Bindungen mit der DNA ausbilden, studieren wir anhand von Monte-Carlo Simulationen die Adsorption eines flexiblen Polymers an ein steifes Polymer mit periodischen Bindungsstellen sowohl in zwei als auch in drei Dimensionen. Die Analyse der Ergebnisse dieser Simulationen zeigt, dass es sich bei dem Phasenübergang von nicht-adsorbiertem zu adsorbiertem Zustand bei sukzessivem Erhöhen der Adsorptionsstärke in zwei Dimensionen um einen Phasenübergang zweiter Ordnung und in drei Dimensionen um einen Phasenübergang höherer Ordnung handelt. Im Vergleich zu adsorbierten Monomeren, tragen aufeinanderfolgende, nicht-adsorbierte Monomere stärker zur Windung des flexiblen Polymers um das starre Polymer bei. Das unterstreicht die Bedeutung der Linker in mC_2H_2-Zinkfingerproteinen für das Umwickeln der DNA.

Acknowledgements

I would like to thank my advisor Prof. Dieter W. Heermann for his enduring support during my doctoral studying period. Without his invaluable advice and instructions, it would be impossible for me to finish this work.

I appreciate my co-advisor Prof. Rebecca C. Wade for her great patience and priceless suggestions.

I am also grateful to my former and current group members, Miriam Fritsche, Songling Li, Hansjörg Jerabek, Gabriell Máté, Wei Xiong, Fei Xing, Chu Min, David Schubert and Andreas Hoffmann. They not only help me a lot on my work, but also give me many unforgettable memories in Germany.

I am willing to acknowledge the fundings from the Heinz-Goetze-Foundation, the Institute for Theoretical Physics, and the Heidelberg Graduate School of Mathematical and Computational Method in Sciences.

Last but not least, many thanks to my parents.

Contents

1 Aim and structure of this thesis

1.1 Intention

How the genetic information carrier, DNA, is dynamically spatially organized in eukaryotic nucleus is a very important question for a better understanding of human genome. Recent experimental approaches have revealed that DNA forms topological domains and subdomains, or generally called loops, in size from several kilobases to megabases [1, 2]. These loops are not formed by randomly distributed DNA distal interactions, but instead are tightly related to the proper genome functions. The underlying molecular mechanism of the organization of these loops then becomes an interesting problem.

One of the most compelling candidates for organizing the genome in eukaryotes is the CCCTC-binding factor (CTCF). Since CTCF can bind to a wide range of long and variant DNA sequences, CTCF binding sites are ubiquitous in human genome and it was described as a "multivalent factor" [3]. After it was discovered that the topological domain boundaries are highly correlated with CTCF binding sites, more and more studies suggested a "master weaver" [4] or an "architectural protein" [5] role of CTCF. Its historical transcription regulation function, as an insulator, has also been reinterpreted as both a con-

sequence and an effector of its ability to organize DNA loops.

Current experiments provide few insight into how CTCF associates with DNA on the molecular scale. By using molecular dynamics and Monte Carlo simulations on different scales, we attempt to shed more light on the conformation and dynamic properties of unbound CTCF, and more importantly, CTCF in complex with DNA.

1.2 Structure of this thesis

In this thesis, due to the limitation of the computing power, the unbound CTCF and CTCF binding to DNA are studied with different models on different length scales.

In *chapter 2*, our current understanding of the genome packing in eukaryotes is introduced. Characteristics of the genome organization on variant length scales are briefly described. We review some recent experimental tools and results about CTCF, focusing on the correlation between CTCF binding sites and genome topological functional domains, as well as the CTCF binding motifs. We summarize this chapter with an introduction of intrinsically disordered proteins to which CTCF belongs, and point out the difficulty in determining the structure of this kind of proteins in experiments.

In *chapter 3*, the basic principles of Monte Carlo simulation and molecular dynamics simulation, both employed in this thesis, are explained with formulas and simple examples. In addition, since in principle DNA and proteins are polymers, we introduce some most commonly used polymer models and concepts for general purposes and for biological macromolecules.

In *chapter 4*, we focus on the conformational properties of unbound CTCF. We construct an atomistic model of CTCF, which contains ten Cys_2His_2 (C_2H_2) zinc finger domains and one C_2HC zinc finger domain, one by one connected by

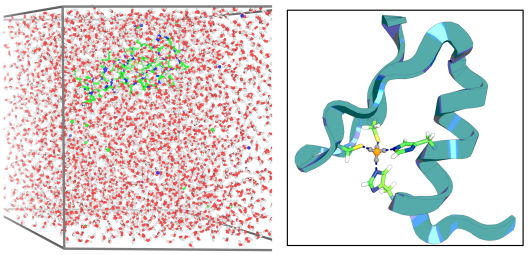
short linkers. Atomistic molecular dynamic simulations of proteins with fewer C_2H_2 zinc fingers confirm that the linkers are flexible and disordered, while the zinc fingers are structured. To improve computation efficiency, we develop an atomistic pivoting algorithm and a mesoscale model. They show that the conformation of unbound multi-C_2H_2 (mC_2H_2), like CTCF, can be explained using a worm-like chain model. For proteins of a few zinc fingers, an effective bending constraint prefers an extended conformation. A self-avoiding chain model applies only to proteins containing more than nine zinc fingers.

In *chapter 5*, we study how a mC_2H_2 zinc finger protein binds to and searches for its target DNA loci. The interactions between mC_2H_2 zinc finger proteins and DNA in a mesoscale model are derived by using a top-down scheme. Molecular dynamics simulations of this model present several interesting kinetic properties of the proteins, such as the rotation coupled sliding, the asymmetrical roles of different zinc fingers and the partial binding partial dangling mode. Our model shows proper DNA sequence specificities. An application to CTCF in complex with its target DNA duplex finds that only the central five zinc fingers of CTCF contact DNA. The DNA duplex is asymmetrically bent, but no DNA loop forms by a single CTCF binding.

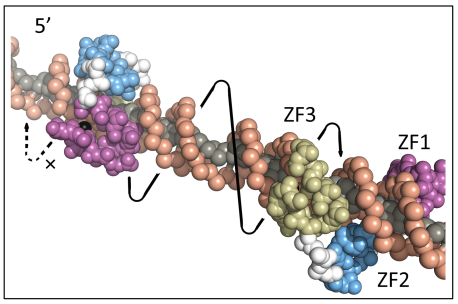
In *chapter 6*, we study our problem by using a more coarse-grained polymer model. The non-covalent binding of polypeptides to DNA is mapped to the adsorption of a flexible polymer to a rigid polymer with periodic distributed binding sites, both in $2d$ and in $3d$. Analysis of the fraction of adsorbed monomers, the specific heat and the Binder cummulant show that the phase transition, from completely non-adsorbed states to adsorbed states with increasing adsorbing strength, is a second order transition in $2d$, and higher order transition in $3d$. We also find that compared to the adsorbed monomers, successive non-adsorbed monomers contribute more to the winding of the flexible polymer around the

rigid one, which reminds us the importance of the linkers in mC_2H_2 zinc finger proteins to wrap around DNA.

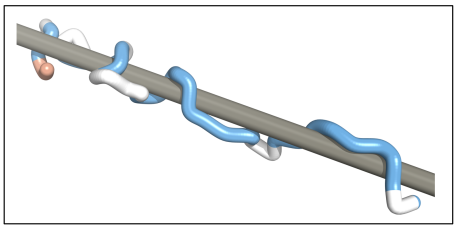
In *chapter 7*, all results are summarized. Some limitations about our work, as well as possible extensions in the future, are discussed. Figure 1.1 gives an visual summary of different models utilized in this thesis.



(a) Chapter 4: Atomistic



(b) Chapter 4/5 : Mesoscale



(c) Chapter 6 : Coarse-grained

Figure 1.1: Multiscale models employed in this thesis. (a) An atomistic unbound C_2H_2 zinc finger with explicit solvent molecules. A cartoon style rendering of the zinc finger is in the right half panel. (b) A mesoscale model of a mC_2H_2 zinc finger protein in complex with double-strand DNA. (c) A coarse-grained polymer model.

While part of chapter 4 (the atomistic pivoting algorithm) has been published in section 3 (A Model for CTCF) in *p1*, the whole chapter 4, as another article *p2*, is in preparation. The results in chapter 5 have been accepted as *p3*, and the contents of chapter 6 are formatted in *p4* that is under peer review.

- *p1.* Feinauer C J, Hofmann A, Goldt S, Liu L, Mate G and Heermann D W. 2013. Zinc finger proteins and the 3D organization of chromosomes. *Organization of Chromosomes (Advances in Protein Chemistry and Structural Biology vol 90)* ed Donev R (Academic Press) pp 67-117. My contribution is a model for CTCF.

- *p2.* Liu L, Wade R C and Heermann D W. 2014. A multiscale approach to simulating the conformational properties of unbound multi-Cys_2His_2 zinc finger proteins. In preparation

- *p3.* Liu L and Heermann D W. 2014. The interaction of DNA with multi-Cys_2His_2 zinc finger proteins. Journal of Physics: Condensed Matter. Accepted

- *p4.* Liu L, Schubert D, Chu M and Heermann D W. 2014. Phase transition and winding properties of a flexible polymer adsorbed to a rigid periodic copolymer. Physical Review E. Under review. Lei Liu performed most simulations and wrote the manuscript. David Schubert and Min Chu contributed other materials to this work.

2 Introduction

In this chapter, biological background about the spatial organization of the genome in eukaryotes and the architectural role of the CCCTC-binding factor (CTCF) is introduced. Some recent experimental tools and results, which show how CTCF is involved in bridging genome packing and function, are discussed. It provides the motivation of the whole thesis, i.e., to study how CTCF interacts with chromatin via modelling. This chapter can be skipped for those who are familiar with these topics.

2.1 Genome Organization in Eukaryotes

The control center of an eukaryotic cell is the cell *nucleus*, as shown in Figure 2.1. It is enclosed by a lipid bilayer membrane (*nucleus envelope*), which is mechanically supported by a dense fibrillar network on the internal face of the envelope (*nucleus lamina*). In mammalian cells, the average diameter of the nucleus is approximately 6 μm, and the nucleus lamina has a thickness of 30 to 100 nm. Most space inside the nucleus is occupied by the most familiar biological macromolecule, *deoxyribonucleic acid* (DNA).

Since the basic mechanism by which DNA carries genetic information was

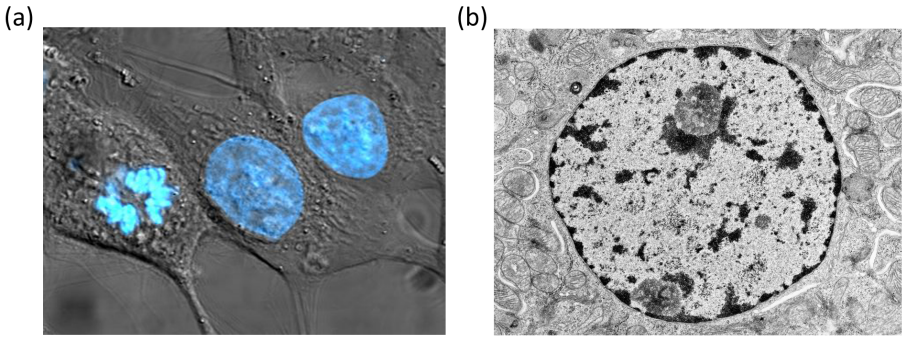


Figure 2.1: Cell nucleus. (a) A microscopy image of HeLa cells in which DNA is stained with blue dye. The central and right cells are in interphase, while the left cell is in mitosis. (b) A microscopy image of cell nucleus. Two types of chromatin appear differently. Heterochromatin is darkly stained and euchromatin is not easily stainable. While euchromatin occupies most part of the cell nucleus, heterochromatin is dispersed all over the nucleus or accumulated near the nuclear envelope. Figures adapted from [6, 7].

discovered by Waston and Crick in 1950's [8], it has become the central icon of molecular biology. As shown in 2.2(a), most DNA molecules are made of two biopolymer strands, which coil around each other and form a double-strand helix. Each strand is composed of four kinds of basic units called nucleotides. They are consisted of the same sugar group, the same phosphate group and different nucleobases, namely adenine (A), cytosine (C), guanine (G) and thymine (T). Along a DNA single strand, nucleotides are connected into a chain by covalent bonds between the sugar group of one nucleotide and the phosphate group of the next, which results in a sugar-phosphate alternating backbone and gives the strand a direction of 3' hydroxyl and 5' phosphate ends. Two strands then are anti-parallelly aligned, i.e., with opposite directions, and bind together via hydrogen bonds between the complementary nucleobase pairs, namely A-T and C-G. Genetic information is stored as the nucleobase sequence on each strand, which can be copied to messenger RNA, used by ribosomes to build proteins

and hence guides the construction and proper functions of entire body.

A naked double-strand DNA has a radius of 10 Å and a pitch of 34 Å. In eukaryotic cells, DNA seldom exists in a naked form. Instead, *histone* proteins bind to the naked DNA and form a DNA-histone complex called *nucleosome* (see Figure 2.2(b)). Two of each of the core histones (H2A, H2B, H3 and H4) make up a octameric nucleosome core with a diameter of about 63 Å. 147 base pairs (bps) of DNA wrap around this core particle 1.65 times in a left-handed super-helical turn, which results in the nucleosome of around 100 Å (or 10 nm) in diameter (see also Figure 2.3(c)). Epigenetic modifications, like methylation, can alter the the interactions between histone proteins and DNA, which influence diverse biological processes though changing the organization of DNA on this scale.

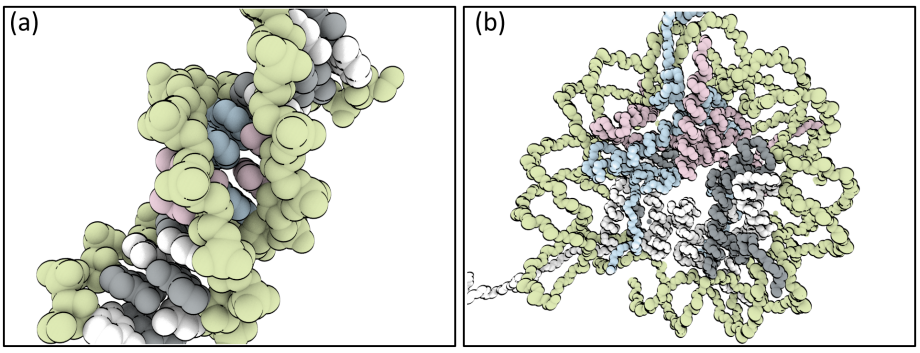


Figure 2.2: Illustrations of DNA and nucleosome. (a) Atomic structure of a section of B-form DNA (PDB code: 1BNA). The DNA backbones (green) form two parallel helical strands, which bind together via non-covalent interactions between the complementary nucleobase pairs, namely adenine(blue)-thymine(purple) and cytosine(white)-guanine(gray). Two grooves of different width, the major groove and the minor groove, are the two helical spaces between strands. (b) Crystal structure of the backbone atoms in a nucleosome core particle (PDB code: 1AOI). DNA is colored green, and the histones H2A, H2B, H3 and H4 are colored blue, purple, white and gray respectively.

Although naked DNA is condensed by histones, how the approximately 2

meters DNA (in humans) is packed into a nucleus of a few micrometers in diameter is not fully understood nowadays. *Chromatin*, the resulting complex fiber composed of DNA and histones, is folded repeatedly and is reorganized according to internal and external stimuli.

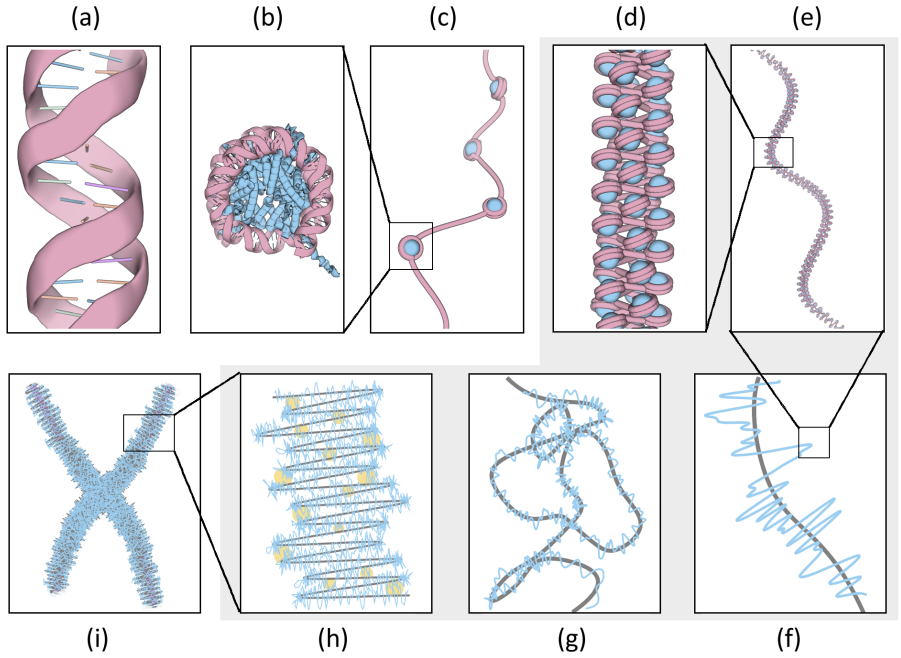


Figure 2.3: Hierarchies of DNA compaction in the nucleus of eukaryotic cells. (a) naked double-strand DNA. (b) nucleosome where DNA (red) wraps round the histone core (blue). (c) 10-nm "bead-on-a-string" chromatin fiber. (d-e) 30-nm chromatin fiber. (e-f) chromatin loops in interphase formed by attaching chromatin to scaffold proteins (gray). (h-i) chromosome in mitosis. Those subplots of gray background (d-h) are still under debates. Image adapted from [9].

In the next round of compaction, the linker histone H1 binds the nucleosomes at the entry and exit sites of the DNA and stacks nucleosomes into a fiber with a diameter of 30 nm (see Figure 2.3(d-e)). Some researchers are still suspicious about its existence. What's more, how the nucleosomes are arranged to form a 30-nm fiber is not clear. Several models, like the solenoid models and the

crossed-linker-models, have been developed to explain the nucleosome packing and stacking.

Above the 30 nm scale, based on the fluorescence in situ hybridization (FISH) [10] and HiC data [11, 1], chromatin fibers in interphase are suggested to form loops in size from several kilobases to a few megabases. An important question is what organizes the chromatin loop in cell nucleus. Possible candidates are the nucleus lamina, the nucleolus, transcription factories and certain architectural proteins, like CTCF (Section 2.2) or scaffold protein (see Figure 2.3(f-g)).

During mitosis the dispersed interphase chromatins undergoes a transition into rigid, more compacted objects *chromosomes*. It has not reached an agreement on the structure of the mitotic chromosomes. Many models have been proposed for the description of its organization, such as attaching the chromatin to a protein scaffold (so-called radial loop model in many textbooks [12], see Figure 2.3(h-i)) and a dynamic loop model with a restricted interaction range [13, 14].

On an even larger scale, it was found that chromatins do not occupy the space in nucleus at random. Depending on the compactness and the gene expression level, there are two types of chromatin, namely *Euchromatin* and *Heterochromatin*. Euchromatin is less condensed and contains genes expressed more frequently. while heterochromatin is more condensed and contains DNA transcribed infrequently. Electron microscopy images of nucleus with staining, like Figure 2.1(b), show that heterochromatin has a tendency to be located near the nucleus envelope. In Figure 2.4, as another example of the non-uniform distribution of chromatins, multiplex FISH images of sections of human nucleus clearly demonstrated that individual chromatins are organized as distinct *chromosome territories* even in the interphase [15, 16]. These results indicate that the high order spatial-temporal organization of chromatins is coupled to the

genome function and cell cycling.

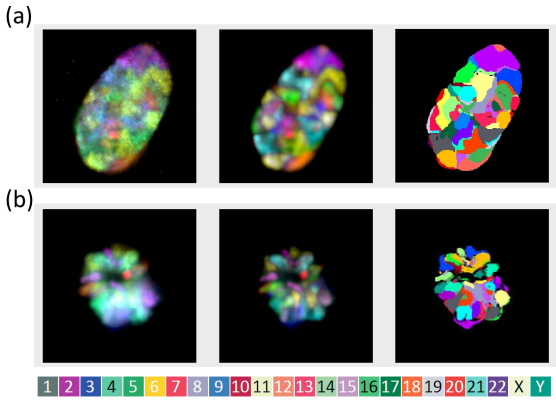


Figure 2.4: Mid-plane section of human male fibroblast nucleus (a) and prometaphase rosette (b) recorded with 24-color mupliplex FISH technique. Different chromosomes are labeled with different colors. Images from left to right are RGB images without deconvolution, after deconvolution and false color images after classification. Image adapted from [15].

One important family of approaches to determine the spatial approximation of genome regions is the *chromosome conformation capture* (3C) technique [18, 19]. As shown in Figure 2.5, 3C-based approaches have basic five steps: (1) addition of formaldehyde to crosslink DNA segments spatially close to each other. Hence interacting DNA sites are fixed, such as the association of an *enhancer* with a *promoter*, (2) cleavage of chromatin by restriction enzyme or sonicaiton. (3) ligation under a dilute condition such that ligation favors from DNA ends captured on the same complex over from random collisions between different complex, (4) reverse crosslinking at high temperature, (5) detection of ligation junction using different ways depending on the variant of the methods.

There are different 3C-based techniques which detect different physical interactions (see Figure 2.6). The basic 3C method tests the interaction between two known sites in the genome via quantitative polymerase chain reaction (qPCR). Circular chromosome conformation capture (4C) allows detecting unknown in-

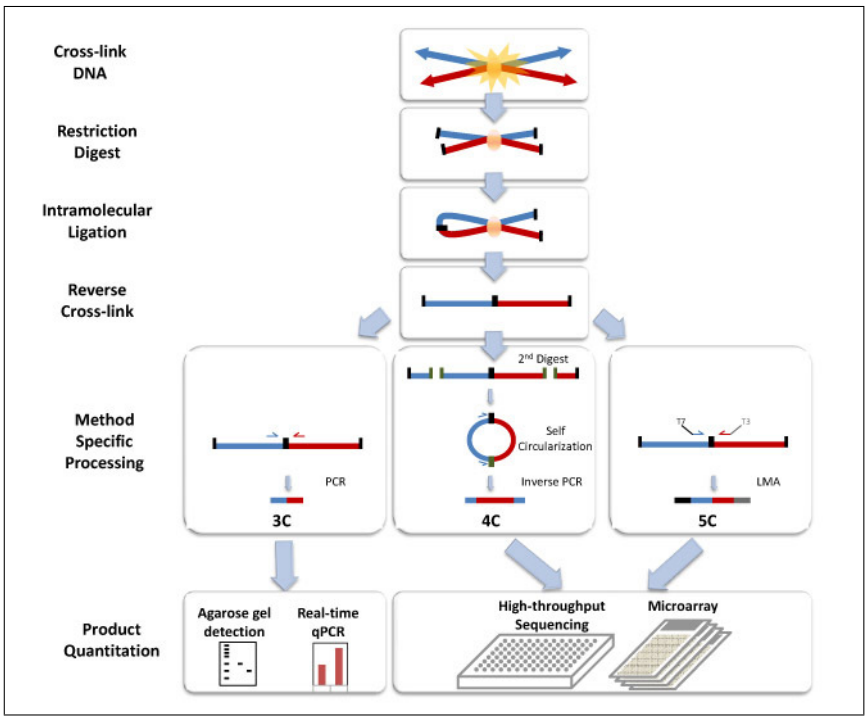


Figure 2.5: Basic protocol of chromosome conformation capture techniques. Image adapted from [17].

teractions of a known bait sequence. Carbon-copy chromosome conformation capture (5C) identifies all regions of interaction within a given genome domain, and Hi-C probes all occurring interactions in an unbiased fashion genome-wide. Further variants, like chromatin interaction analysis by paired-end tag sequence (ChIA-PET) and chromatin immunoprecipitation-loop (ChIP-Loop), determine genome interactions involving a specific protein of interest by incorporating an additional protein precipitation step.

With the advances of these approaches, more and more details about the high order chromatin organization have been revealed. Application of Hi-C to the human and mouse nucleus in embryonic stem cells (ESCs) and terminally differentiated cell types identified large, megabase-sized local chromatin interac-

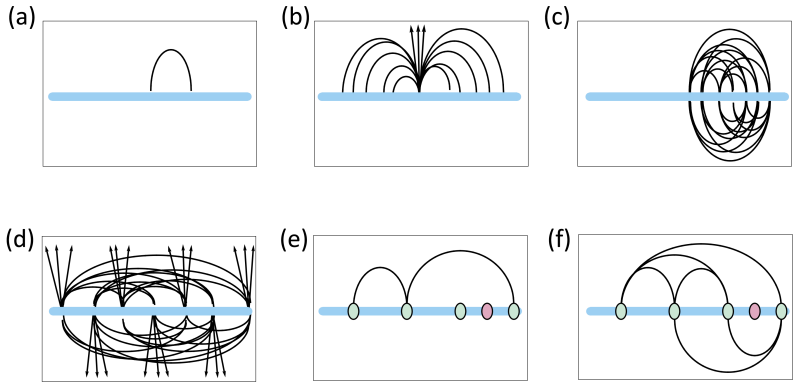


Figure 2.6: Physical interactions (black curves and arrows) detected by (a) 3C, (b) 4C, (c) 5C, (d) Hi-C, (e) ChIP-Loop and (f) ChIA-PET. Chromatin fiber is represented by the thick blue line. In subplots (e-f), the colored disks represent chromatin-associated proteins, and the specific protein of interest is colored green. Image adapted from [19].

tion domains, which are termed *topological domains* [1]. The domains are stable across different cell types and highly conserved cross species, indicating that these topological domains are an inherent property of mammalian genomes. Hi-C experiments with higher resolution showed that within these megabase-sized domains, there exist *topological subdomains* with a mean size of 520 kilobases [2]. As an example, Figure 2.7 shows a Hi-C interaction frequency heatmap of a segment on chromosome 2 in human genome. Genome positions are labeled on the underlying axis. One domain and one subdomain are outlined.

All these non-uniform non-random high order eukaryotic genome organization motifs (chromosome territories, topological domains, subdomains and perhaps sub-subdomains found by using even higher resolution in the future) suggest that chromatins might be architectured in a function-related way. Although *without* atomistic details, 3C-based techniques combined with other tools, such as ChIP-ChIP and ChIP-seq which will be introduced in the next section, provide us further possibilities to study *which* factors and *how* they organize the

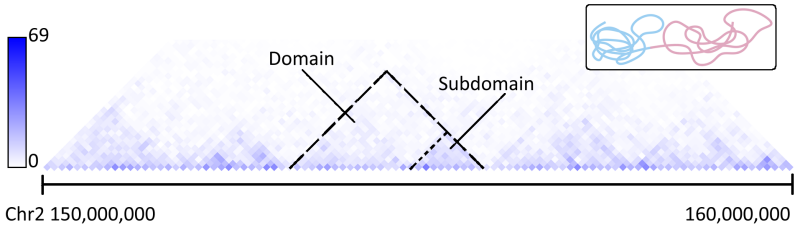


Figure 2.7: Observed intrachromosomal interaction frequency heatmap of a 10 Mb region on chromosome 2 in normal human cells. One topological domain and one subdomain are outlined. The top right insert sketches a chromatin fiber segmented into two (sub)domains depending on its self-contact. Data downloaded from [11] and image adapted from [2].

chromatin on a genome-wide scale, as well as at specific gene loci.

2.2 CTCF: from Genome Topology to Function

One of the most compelling candidates for organizing the genome in eukaryotes is the CCCTC-binding factor (CTCF). The full-length protein is composed of 727 amino acids. It contains a central DNA-binding domain, which is composed of 11 zinc fingers and is almost 100% amino acids sequence conserved among mouse, chicken and human. Based on its ability to employ different zinc fingers to bind to a wide range of *long* and *variant* nucleotide sequences, CTCF binding sites are ubiquitous in the human genome. It has been described as a "multivalent factor" [3, 4]. Many recent studies, using the *chromatin immunoprecipitation followed by DNA microarray* (ChIP-chip) or *followed by DNA sequencing* (ChIP-seq) techniques (see Figure 2.8), have been devoted into characterizing the genome-wide CTCF binding profile.

Ren and his colleagues performed a ChIP-chip analysis against CTCF and confirmed its binding to 13,804 regions in human fibroblast cells [21]. In ad-

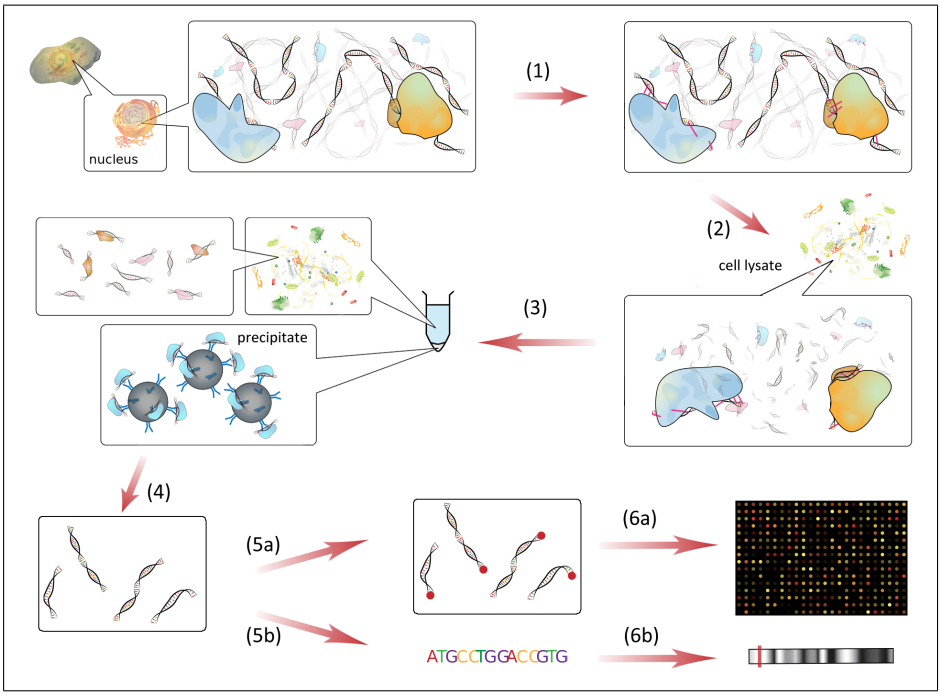


Figure 2.8: A basic workflow of a Chip-Chip or Chip-Seq experiment: (1) crosslink DNA-binding proteins to DNA, (2) lyse the cells and shear the DNA by sonication, (3) immunoprecipitate the crosslinked complexes with an antibody specific to the protein of interest, (4) reverse the crosslinking and isolate DNA strands. For ChIP-chip, the DNA fragments are labeled with a fluorescent tag (5a) and analyzed via DNA microarray (6a). For ChIP-seq, the immunoprecipitated sample is analyzed using high-throughput sequencing (5b), then mapped to the genome (6b). Image adapted from [20].

dition, most of its localization was found invariant across different cell types (compared with a hematopoietic progenitor cell line). By analyzing the conserved noncoding elements in the human genome, Lander et al. first discovered several groups of long nucleotide motifs which do not match any previously known motif, then they demonstrated by biochemical (Western blot) and computational methods that CTCF binds to the motifs in the largest group. A total of $\sim 15,000$ conserved binding sites of CTCF were found through this way [22].

Another computational study by Wang's group aimed to classify CTCF binding sites as cell type-specific (only found in one out of 38 cell lines), or conserved (found in all 38 cell lines). Approximately 66,800 CTCF binding sites were identified from each cell type. In cell type K562, 28% of the binding sites ($\sim 18,700$) were conserved [23]. A database of CTCF binding sites, CTCFBSDB, has also been constructed online, which now contains nearly 15 million experimentally determined CTCF binding sequences [24, 25].

CTCF was traditionally characterized as an *insulator* [26, 3, 27], which interferes with enhancer-promoter communication (*activator* or *repressor*), or buffers transgenes from chromosomal position effects (as a *chromatin barrier*) caused by heterochromatin spreading. In other words, it showed distinct functions at different loci depending on the biological context. Now more and more evidences [4, 28, 29, 5] strongly support that the mechanism underlying the diverse functions of CTCF is both a consequence and an effector of its contribution in organizing chromatin loops in the cell nucleus.

The most well known example comes from the analysis of the imprinted *H19*-insulin-like growth factor 2 (*IGF2*) locus [31, 4]. Figure 2.9(a) [2] shows the CTCF binding profile overlayed with 4C interaction profiles of five different viewpoints (highlighted in red) at the *H19-IGF2* locus on human chromosome 11. The *IGF2* promoter region (vp1) interacts strongly with an intergenic region, where CTCF binds, between *H19* gene and *IGF2* gene. This chromatin interaction is also confirmed by vp2. A region upstream of *H19* (vp3) also shows interactions with the intergenic region. In contrast, viewpoints in a CTCF-depleted region (vp4) and at the domain boundary (vp5) show only weak interactions. Based on several independent researches, a linear depiction and a simplified sketch of how CTCF mediates long-range chromatin contacts at the *H19-IGF2* locus [31, 30] are presented in Figure 2.9(b,d) and (c,e), for the ma-

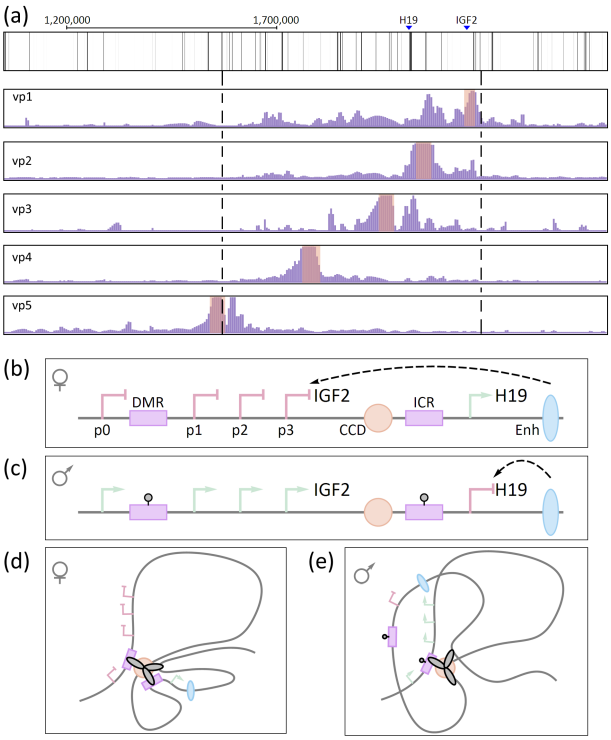


Figure 2.9: CTCF-mediated interactions at the imprinted *H19-IGF2* locus. (a) CTCF binding sites overlayed with 4C interaction profiles from five different viewpoints (highlighted in red). Two vertical dashed lines indicates the boundaries of a domain. (b,c) Linear depiction of the *H19-IGF2* locus. The maternally expressed noncoding *H19* gene is located downstream from the gene encoding Insulator-like growth factor 2 (*IGF2*) that is expressed exclusively from the paternal allele. The imprinting control region (ICR) between *H19* and *IGF2* contains CTCF binding sites and is essential for controlling entire locus. The differential methylated region (DMR) upstream of *IGF2* promoters (p1,2,3) and central conserved DNase I hypersensitive domain (CCD) act together to regulate allele-specific expression patterns with a shared set of downstream enhancers (Enh). DNA methylation of ICR and DMR in the paternal allele are represented by appending an additional small circle. (d,e) Schematic models illustrating allel-specific patterns of CTCF binding, DNA methylation and chromatin looping. Image adapted from [2, 4, 30].

ternal and paternal allele, respectively. On the maternal allele, the imprinting control region (ICR) is unmethylated, CTCF is bound, and the enhancer down-

stream of *H19* is prevented from accessing the *IGF2* promoter. On the paternal allele, the ICR is methylated, CTCF is unbound, which results in the spatial approximation between the enhancer and promoter and hence the expression of *IGF2*. Except for *H19-IGF2*, expression patterns at other loci involving CTCF, such as mouse β-globin, can also be better understood if CTCF is considered to mediate chromatin loops.

Statistic on the genome-wide comparison between the binding profile of CTCF and the distributions of other factors provides descriptions of CTCF binding on a larger scale.

CTCF normally is positioned in regions surrounded by well-positioned nucleosomes [23]. Approximately 50% of CTCF binding sites are located in intergenic regions, $\sim$ 15% are near promoters and $\sim$ 40% are intragenic. Although CTCF binding profile strongly correlates with gene density, similar to a canonical transcription factor such as TAF1, there are key differences between their distributions [21]. The majority of the TAF1 binding sites ($\sim$ 89%) are close to transcription starting sites, while the average distance from CTCF binding sites to promoters is further away (48,000 bp). It is also noticed that CTCF-depleted regions tend to include clusters of related gene families and coregulated genes, while CTCF-enriched regions often have multiple alternative promoters.

It has been shown that density of open chromatin tags, such as DNase I hypersensitive sites (chromatin region which has lost its condensed structure, and therefore it is sensitive to cleavage by the DNase I enzyme), is sharply elevated within the CTCF binding sites [23]. Analysis of the histone modification patterns in human genome [35] revealed that all the three states of H3K4 methylation are enriched in promoter regions of active genes, and highly correlates with CTCF binding sites. In addition, many of the CTCF binding are detected between histone methylation-defined active and silent chromatin

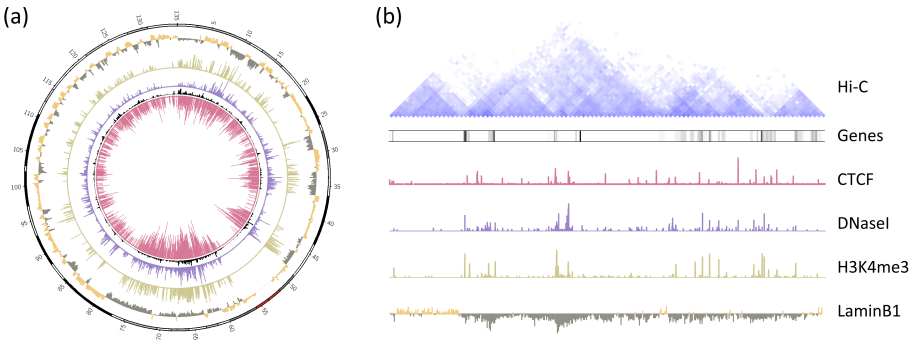


Figure 2.10: CTCF binding sites distribution. Profile of CTCF binding sites, gene density, Dnase I hypersensitive sites, histone modification H3K4me3 and lamin B occupancies (a) on human chromosome 11, (b) at Sox2 locus on mouse chromosome 3. Tracks in (b) are overlayed by a Hi-C interactions heatmap. (a) is generated using the Circos software package. Data fetched from [32, 33, 34] and image adapted from [36, 23].

domains, which is in consistent with its insulator role. By aligning the binding profile of CTCF and lamin B (fibrous protein in nuclear lamina), Wei and her colleagues [36] found that CTCF looping signals are depleted in lamin B-associated domains. Last but not least, Hi-C data in mammalian cells [1] showed that most topological domain boundaries are enriched for the binding of CTCF. As two examples, Figure 2.10 aligns the binding profile of CTCF, gene density, DNase I hypersensitive sites, histone modification H3K4me3 pattern and lamin B binding density on (a) human chromosome 11 and (b) Sox2 gene locus on mouse chromosome 3 with Hi-C interactions heatmap. More details are referred to [36, 23].

All these studies indicate that CTCF contributes to the construction of topological and functional domains of human genome. However, it should be pointed out that CTCF is not the sole player in genome organization [30, 37]. By examining the consequence of depletion of factors of interest on high order chromatin organization, Wendt et al. [2] observed a general reduce of chromatin inter-

actions but intact topological domains after *cohesin* depletion, and reduced intradomain interactions but increased interdomain interactions after CTCF depletion. They concluded that CTCF and cohesin contribute differently to chromatin organization. There lacks a comprehensive description of the functions of CTCF neither. For example, although CTCF binding site density is elevated, only 15% of its binding sites are located within topological domain boundaries.

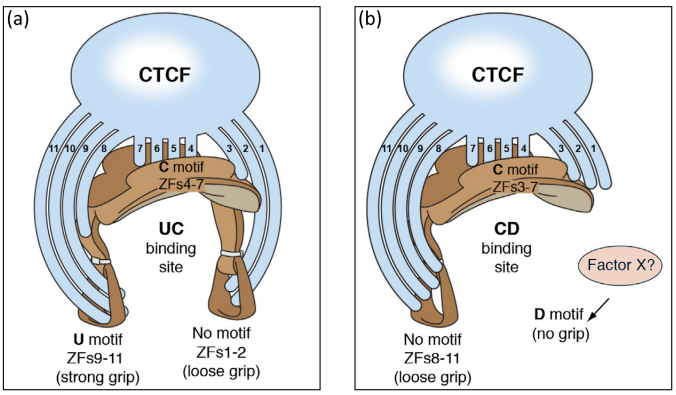


Figure 2.11: "Saddle" model of CTCF binding to DNA [38]. (a) CTCF binds tightly to a DNA site containing an upstream (U) and a core (C) motifs. While zinc fingers (ZFs) 4-7 interact with the core motif, the entire complex is stabilized by via ZFs 9-11 binding to the upstream motif. ZFs 1-2 contributes to the binding by associating with nonspecific sequence. (b) CTCF binds loosely to a DNA site containing a core and a downstream (D) motifs. ZFs 1-2 no longer associate with DNA or an unknown factor X outcompetes it for binding. Image adapted from [38].

Lots of efforts have been made to identify a consensus nucleotide sequence motif for CTCF binding, albeit no agreement has been reached yet. Early study by Ohlsson et al. [39] reported an 50-60 bp sequence. A large number of *in vivo* CTCF binding sites measured by ChIP-based techniques provides a larger opportunity to find a consensus motif. A 20 bp motif defined from ChIP-chip experiment was reported later [21]. By deleting individual zinc fingers and

mutating individual sites, Pedone et al. [40] determined a core 12 bp DNA motif to which CTCF binds with high affinity ($K_D \sim 10^{-10}$). What's more, it was shown that only 4 out of 11 zinc fingers are essential to *strong directional* binding, and the N- and C- terminal regions of the protein which flank the zinc finger domains are not required for DNA binding. More recently, based on the analysis of genome-wide binding profiles of CTCF zinc finger mutants, Casellas et al. [38] reported a $\sim$ 55 bp consensus sequence comprising a 10 bp upstream motif, a 20 bp core motif and a 10 bp downstream motif, which joined by two spacer sequences of $\sim$ 6 bp each. A "saddle" model of CTCF binding was proposed, which is presented in Figure 2.11.

As Phillips and Corces summarized at the end of a review article about CTCF in 2009 [4]:

> *Many important questions remain to be answered. Determination of the crystal structure of the zinc finger domain would lend significant understanding into how CTCF's conformation and the specific zinc fingers associated with DNA change upon binding to divergent sequences ... Organizing principles for loop formation should be established, in particular an unambiguous conclusion regarding whether chromatin interactions involve CTCF homodimerization or heterodimerization or if a single CTCF molecule can bring together multiple regulatory elements by serving as a substrate for proteins such as cohesin known to mediate chromatin contacts*,

current experimental data provide few insight into how CTCF associates with DNA on atomistic scale. The underlying reason is that CTCF belongs to *intrinsically disordered proteins* (IDPs), sometimes also called *intrinsically unstructured proteins*. IDPs contain domains that are unstructured in solution, but usually become structured on binding to their physiological targets. Oc-

currence of unstructured regions is quite common. Dunker and his colleagues [41] showed that more than 30% of eukaryotic proteins containing unstructured regions of size > 50 residues.

As shown in Figure 2.12, with a growing ratio of stable $3d$ content, structural characteristics change from highly extended, unstructured states, to compact but disordered molten globules, to proteins with multi folded domains which linked by flexible or disorder linker sequences and, finally, to mostly folded single domain proteins with only local disordered.

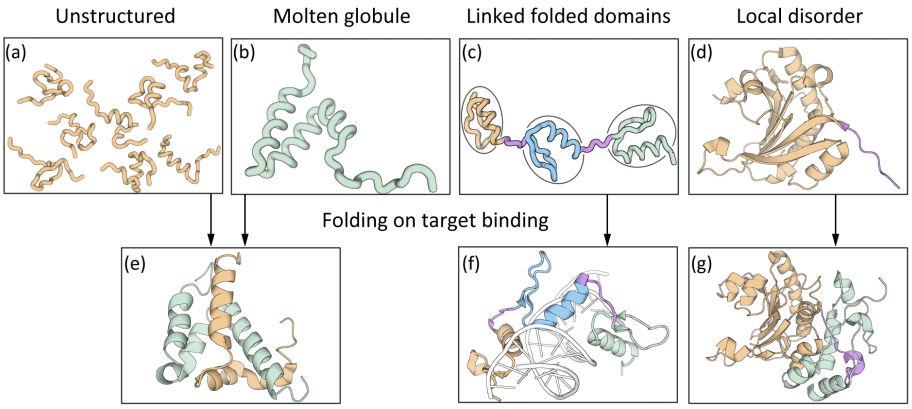


Figure 2.12: Examples of intrinsically disordered proteins. For the upper panels, from left to right, content of unstructured regions in the proteins increases. (a) An unstructured conformation ensemble of a region of nuclear receptor coactivator (ACTR). (b) A molten globule-like nuclear coactivator binding domain (NCBD) of CREB-binding protein (CBP) (PBD code: 2KKJ). (c) First three linked zinc finger domains of unbond transcription factor TFIIIA. (d) Free eukaryotic translation initiation factor 4E (eIF4E), which is mostly folded with unfolded N terminus (PDB code: 1EJH). Local disorder regions are labeled purple. The lower panels show well-ordered structures, as results from specific binding of proteins to their targets. (e) ACTR-NCBD complex (PDB code: 1KBH). (f) TFIIIA in complex with specific oligonucleotide (PDB code: 1TF3). (g) eIF4E-eIF4G complex (PDB code: 1RF8). Image adapted from [42].

Many eukaryotic proteins are modular and fall into the third subdivision, such as CTCF. They contain independently folded globular domains which are

separated by flexible linker regions. In the absence of their targets, modular proteins behaves as "beads on a flexible string", where the linker allows a relatively unhindered spatial search by the attached domains. Specific binding can induce a ordered structure of the linkers, and hence stabilize the structure of entire complex. A well known example is the binding of three Cys_2His_2 zinc finger proteins to target DNA loci, which will lead the linkers to fold, cap, and instruct consecutive zinc fingers to bind correctly in the major groove of DNA [43].

Because crystals of conformationally disordered molecules are difficult to form, crystallographic studies can not provide much information on unstructured states. It can only indicate their presence through the absence of electron density in local regions. Even if it succeeds to form a crystal, the crystallized structure may not be representative of the *conformation ensemble* of the IDP in solution. Instead, meaningful knowledge about the overall shape and size distribution (see Figure 2.13), long-range residue-residue contacts and backbone flexibility of IDPs is accessible through nuclear magnetic resonance (NMR) [44] and small-angle X-ray scattering (SAXS) experiments [45].

How to describe the unstructured states of IDPs and how to construct a conformation ensemble which represents the real dynamic conformations of IDPs are still under research. Based on the clear similarity, concepts in Polymer Physics, such as the radius of gyration (R_g) or end-to-end distance (R_e) (more details referred to Section 3.1), can help us to understand the conformational properties of IDPs. For example, Figure 2.13(a) shows the variety of the conformation of Tau protein [46] through the distribution of R_gs. In subfigure (b) [45], dependence of R_g on the chain length of IDPs is plotted in a log-log scale. The fitted straight solid line ($R_g \sim N^{\nu}$) assigns ν a value 0.522, which is close to that of a self-avoiding chain (0.588). Experimental data of IDPs or denatured

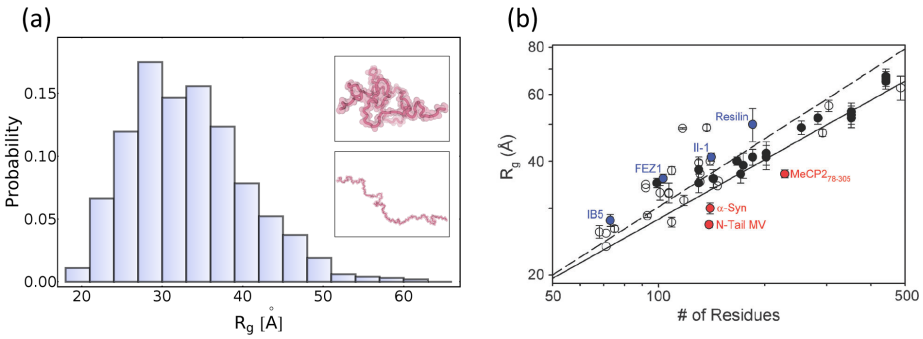


Figure 2.13: Radius of gyration (R_g) of intrinsically disordered proteins (IDPs). (a) Distribution of R_gs in a NMR-measured conformation ensemble of K18 domain of Tau protein [46]. Two inserts show the conformation of largest and smallest R_g, upper and below, respectively. (b) Experimentally measured R_gs versus the chain length of IDPs. Full dots represent various Tau protein constructs, and blue (red) dots label those more extended (compacted) IDPs with names. The straight solid line corresponds to the fitted Flory's relationship $R_g = (2.54 \pm 0.01) N^{0.522 \pm 0.01}$. Image (b) adapted from [45].

proteins from other sources, like single molecule fluorescence spectroscopy [47], have been also interpreted with these theoretical tools.

3 Methods

Two kinds of methods, *Monte Carlo method* and *molecular dynamics simulation method*, are most widely used to simulate *classical* biological macromolecular systems. Both methods have their pros and cons, and have been applied in this thesis. In general, it could be more efficient for computing the equlibrium properties of a system with Monte Carlo method than with molecular dynamics simulation method, while the latter method is a better choice if the non-equilibrium properties are of interest. In this chapter, we describe the basic principles of them, as well as a few examples.

3.1 Monte Carlo method

3.1.1 Importance sampling and Metropolis algorithm

In classic statistical mechanics, the probability of a system stays at $\boldsymbol{x}$ in phase space, where $\boldsymbol{x}$ stands for the set of variables describing the considered degrees of freedom, is given by the Boltzmann distribution in the canonical ensemble

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp(-\frac{H(\boldsymbol{x})}{k_B T}), \tag{3.1}$$

in which the partition function Z equals

$$Z = \int \exp(-\frac{H(\boldsymbol{x})}{k_B T})\,\mathrm{d}\boldsymbol{x}, \tag{3.2}$$

k_B is the Boltzmann factor, $H(\boldsymbol{x})$ is the Hamiltonian and T is the temperature of the system. Then the thermal average of any observable $A(\boldsymbol{x})$ is

$$< A(\boldsymbol{x}) >_T = \frac{1}{Z} \int A(\boldsymbol{x}) \exp(-\frac{H(\boldsymbol{x})}{k_B T})\,\mathrm{d}\boldsymbol{x}. \tag{3.3}$$

To compute the multi-dimensional integration in $< A(\boldsymbol{x}) >_T$ numerically, the most direct way in which we can do is to randomly select points $\{\boldsymbol{x}_i\}, i = 1, 2, ..., M$ from the phase space independently, and to approximate the integration by summations as

$$\overline{A(\boldsymbol{x})} = \frac{\sum_{i=1}^{M} \exp(-\frac{H(\boldsymbol{x}_i)}{k_B T}) A(\boldsymbol{x}_i)}{\sum_{i=1}^{M} \exp(-\frac{H(\boldsymbol{x}_i)}{k_B T})}. \tag{3.4}$$

However this so-called *simple sampling* is of little use in practice. Because according to equation 2.1, the Boltzmann weight on $\{\boldsymbol{x}_i\}$ will be negligibly small and terms associated with these points contribute little to the average. Instead, the idea behind an *importance sampling* is to sample more points in the phase space where the Boltzmann weight is large and fewer elsewhere. Now supposing the points $\{\boldsymbol{x}_i\}$ in the phase space are selected based on some probability $\omega(\boldsymbol{x}_i)$, the estimation of $< A(\boldsymbol{x}) >_T$ becomes

$$\overline{A(\boldsymbol{x})} = \frac{\sum_{i=1}^{M} \exp(-\frac{H(\boldsymbol{x}_i)}{k_B T})/\omega(\boldsymbol{x}_i) A(\boldsymbol{x}_i)}{\sum_{i=1}^{M} \exp(-\frac{H(\boldsymbol{x}_i)}{k_B T})/\omega(\boldsymbol{x}_i)}. \tag{3.5}$$

Obviously, the most natural choice for $\omega(\boldsymbol{x}_i)$ would be $\omega(\boldsymbol{x}_i) \propto \exp(-H(\boldsymbol{x}_i)/k_B T)$.

Then Equation (2.5) reduces to an arithmetic average

$$\overline{A(\boldsymbol{x})} = \frac{1}{M}\sum_{i=1}^{M} A(\boldsymbol{x}_i). \tag{3.6}$$

This process was first proposed by Metropolis et al. in 1953 [48] using a Markov process where each state $\boldsymbol{x}_{i+1}$ is constructed from its previous state $\boldsymbol{x}_i$ via a proper transition probability $\omega(\boldsymbol{x}_i \to \boldsymbol{x}_{i+1})$. With the transition probability

$$\frac{\omega(\boldsymbol{x}_i \to \boldsymbol{x}_{i+1})}{\omega(\boldsymbol{x}_{i+1} \to \boldsymbol{x}_i)} = \exp(-\frac{H(\boldsymbol{x}_{i+1}) - H(\boldsymbol{x}_i)}{k_B T}) = \exp(-\frac{\delta H}{k_B T}), \tag{3.7}$$

it is well known that the probability distribution of generated states tends to the Boltzmann distribution as the number of generated states approaches infinity. One example of the explicit form of Metropolis algorithm in implementation is

$$P(\boldsymbol{x}_i \to \boldsymbol{x}_{i+1}) = \begin{cases} 1 & \delta H \leq 0, \\ \exp(-\frac{\delta H}{k_B T}) & \delta H > 0. \end{cases} \tag{3.8}$$

where $P(\boldsymbol{x}_i \to \boldsymbol{x}_{i+1})$ is the probability to accept a trial move, which try to change the system from state $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$.

Given a sequence of states $\{\boldsymbol{x}_i\}$ generated by a Monte Carlo simulation, a key concept in analyzing the results is the autocorrelation time τ_{auto} for any observable $A(\boldsymbol{x})$. Let's define the normalized autocorrelation function $\rho(t)$ as

$$\rho(t) = \frac{< A(\boldsymbol{x}_i), A(\boldsymbol{x}_{i+t}) > - < A(\boldsymbol{x}_i) >^2}{< A(\boldsymbol{x}_i), A(\boldsymbol{x}_i) > - < A(\boldsymbol{x}_i) >^2}. \tag{3.9}$$

$\rho(t)$ measures the correlation of $A(\boldsymbol{x})$ between system states, which are separately sampled by t Monte Carlo steps. As t increases from 0, $\rho(t)$ decreases from 1 to 0 exponentially or in most case slower than exponentially. An integrated

autocorrelation time τ_{auto} is then defined as

$$\tau_{auto} = \frac{1}{2} \sum_{t=-\infty}^{+\infty} \rho(t) \approx \sum_{t=0}^{M} \rho(t) - \frac{1}{2}. \tag{3.10}$$

The approximation in last formula is given by an "automatic winding" algorithm [49] to estimate τ_{auto}, where M is chosen to be the smallest integer such that $M \geq c \times \tau_{auto}(M)$. If $\rho(t)$ roughly decays exponentially, it would be suffice to take $c \approx 4$. Otherwise, it is prudent to take $6 \leq c \leq 10$.

From an initial state $\boldsymbol{x}_0$, the Metropolis algorithm will first drive the system to its thermal equilibrium. To eliminate this initialization bias, $10 \sim 20 \times \tau_{auto}$ states sampled in the beginning of the simulation will be discarded. After the equilibration, states will be saved in every $10 \times \tau_{auto}$ Monte Carlo steps, which are considered to be uncorrelated and will be averaged out to calculate $A(\boldsymbol{x})$ using equation 2.6. Other factors, like the boundary condition and the finite size of the simulated system, may also have significant effects that need to be considered to interpret the Monte Carlo simulation results properly.

Many important systems in statistical mechanics have been studied by Monte Carlo method. Here are two simplest examples. Figure 3.1 shows the initialization stage of a Monte Carlo simulation for a standard zero-field ferromagnetic two-state Ising model [50] ($H = -J \sum_{i,j} s_i s_j, s_{i,j} = \pm 1$) on a triangular lattice, with $J = 0.2$ in (a) and $J = 0.3$ in (b). Configuration of the system, with different spin states in different colors, at Monte Carlo step $t = \{25, 50, 75, 100\}$ are plotted in (c)-(f) with $J = 0.2$, and in (g)-(j) with $J = 0.3$ respectively. It is clear from (a) and (b) how the energy of the system fluctuates and reaches a equilibrium sooner or later. Comparing the system configuration in equilibrium with $J = 0.2$ (f) to the configuration with $J = 0.3$ (j), spontaneous magnetization happens when J increases (or the temperature of the system decreases beyond the Curie temperature) as expected.

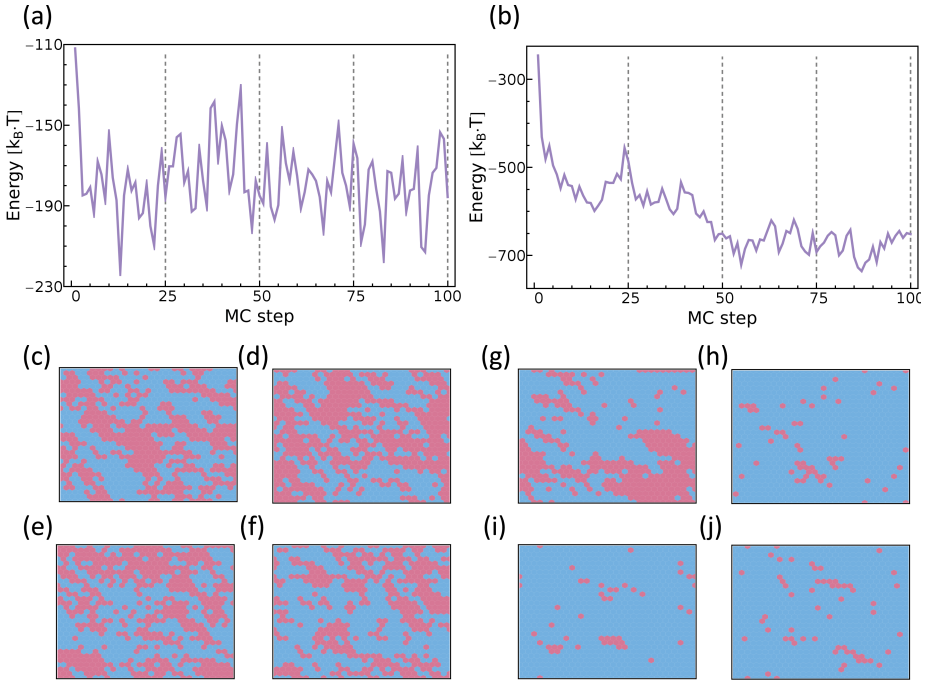


Figure 3.1: Hamiltonian and configurations of a ferromagnetic two-states Ising model simulated using Monte Carlo method on a 30×30 triangular lattice. Hamiltonian of the system versus the simulation step in the initialization stage with $J = 0.2$ (a) and with $J = 0.3$ (b). Configurations at Monte Carlo step $t = \{25, 50, 75, 100\}$ with $J = 0.2$ in (c)-(f), and with $J = 0.3$ in (g)-(j) respectively. Different spin states are labeled via different colors.

The second example is a two-dimensional Lennard-Jones fluid model in canonical ensemble [51], again simulated on a triangular lattice. Given the distance between the nearest neighbor sites to be σ, the Hamiltonian of the system $H = \sum_{i,j} V_{LJ}(r_{i,j})$, where the Lennard-Jones potential V_{LJ} has a form

$$V_{LJ}(r_{i,j}) = \begin{cases} 4\epsilon((\frac{\sigma}{r_{i,j}})^{12} - (\frac{\sigma}{r_{i,j}})^{6}) & r_{i,j} < 2.5\sigma, \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

Progress of the Hamiltonian and configuration of the system in the initialization

stage of a Monte Carlo simulation are presented in Figure 3.2 with temperature $T = 0.1\,\epsilon/k_B$ in (a)(c)-(f), and with $T = 0.2\,\epsilon/k_B$ in (b)(g)-(j), respectively. The simulations start from a "liquid" phase. As it shows, molecules stay aggregated when the temperature is low. But evaporation begins as soon as the temperature climbs up.

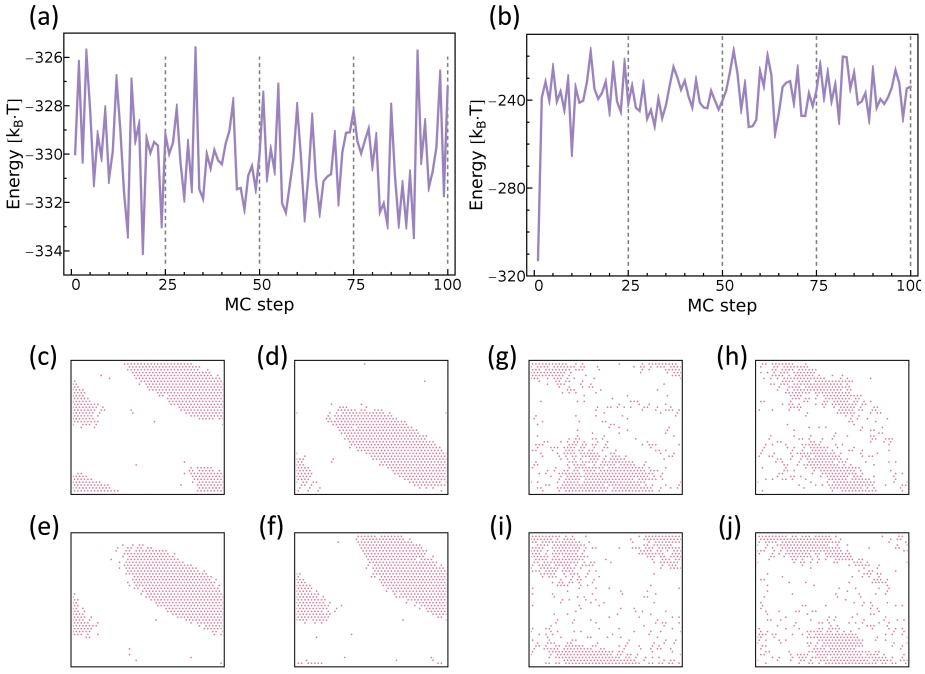


Figure 3.2: Hamiltonian and configurations of a two-dimensional Lennard-Jones fluid simulated via Monte Carlo method on a 45×45 triangular lattice, with system temperature $T = 0.1\,\epsilon/k_B$ in (a)(c)-(f), and with system temperature $T = 0.2\,\epsilon/k_B$ in (b)(g)-(j). Configurations are plotted at Monte Carlo step $t = \{25, 50, 75, 100\}$, where dots represent molecules on lattice.

Not only in Statistical Physics, Monte Carlo method has its wide usage in many other fields, such as Pharmacy [52] and Finance [53]. In the next subsection, first some physical quantities and models in Polymers Physics will be introduced, then we will show how to simulate polymers using Monte Carlo

method.

3.1.2 Applications in Polymer Physics

A *polymer* is a macromolecule which is composed of many repeated subunits *monomers*. Depending on its structure, polymers can be classified into linear polymers and branched polymers. Depending on the properties of the subunits, polymers can be categorized into homopolymers (all monomers of the same type) and copolymers (monomers of different types). Many biological macromolecules are polymers, or can be easily modeled as polymers on a coarse-grained scale.

Given a polymer consisted of $N+1$ monomers of positions $\{\boldsymbol{r}_0, \boldsymbol{r}_1, ..., \boldsymbol{r}_N\}$, there are a few quantities frequently used to describe the conformation of the chain.

- The end-to-end distance R_e is defined as

$$R_e = |\sum_{i=1}^{N} \boldsymbol{r}_i - \boldsymbol{r}_{i-1}|, \tag{3.12}$$

  which denotes the distance from the first monomer to the last monomer.

- The radius of gyration R_g, which measures the dimension of the polymer chain, is given by

$$R_g = (\frac{1}{N+1}\sum_{i=0}^{N} |\boldsymbol{r}_i - \boldsymbol{r}_{com}|^2)^{1/2}, \tag{3.13}$$

  where $\boldsymbol{r}_{com}$ is the center of mass of the polymer, e.g., for homopolymers $\boldsymbol{r}_{com} = \frac{1}{N+1}\sum_{i=0}^{N} \boldsymbol{r}_i$.

- The persistence length ξ_p is calculated from the decay of the bond angle correlation

$$\exp(-l/\xi_p) =< \boldsymbol{u}(l_0) \cdot \boldsymbol{u}(l_0 + l) >_{l_0}, \tag{3.14}$$

where l is the contour length along the polymer chain, $\boldsymbol{u}(l)$ is a unit vector parallel to the chain segment at contour length l, and the average $<>_{l_0}$ is carried out over all possible l_0 along the chain. The persistence length ξ_p is a measurement the stiffness of the polymer. The more rigid the chain is, the larger the ξ_p is.

Corresponding to different types of polymers, different physical models have been proposed based on different assumptions.

- The most simple model is the so-called *freely-joined chain* (or *ideal chain*) model. The bond length is a constant, i.e., $|\boldsymbol{b}_i| = |\boldsymbol{r}_i - \boldsymbol{r}_{i-1}| = b, \forall i \in \{1, 2, .., N\}$. In addition, the direction of the bonds are completely independent of each other. We get

$$< R_e^2 >= \sum_{i=1}^{N}\sum_{j=1}^{N} < \boldsymbol{b}_i \cdot \boldsymbol{b}_j >= \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_{i,j} b^2 = Nb^2. \tag{3.15}$$

  This can also be written in a scaling relation as

$$< R_e^2 > \sim N^{2\nu} \tag{3.16}$$

  with $\nu = 0.5$. Equation (2.15) is the same as the mean squared displacement of a N-step random walk with a step length b. Therefore, a freely-joined chain is also sometimes called a *random-walk chain*.

- Usually the length of a chemical bond in macromolecules is fluctuating, instead of a fixed value. The *Gaussian chain* model describes the flexibility of the bond length using a Gaussian distribution,

$$p(\boldsymbol{b}) = \frac{3}{2\pi b^2}^{3/2} \exp(-\frac{3|\boldsymbol{b}|^2}{2 < b^2 >}). \tag{3.17}$$

  The resulting system is equivalent to a polymer with its monomers con-

nected by harmonic springs, where the potential energy in the springs is $\frac{\kappa}{2}\sum_1^N |\boldsymbol{b}_i|^2$ with $\kappa = \frac{3}{<b^2>}k_BT$. The distribution for the end-to-end vector $\boldsymbol{R}_e$ is

$$p(\boldsymbol{R}_e) = (\frac{3}{2\pi N < b^2 >})^{3/2} \exp(-\frac{3R_e^2}{2N < b^2 >}). \tag{3.18}$$

It can be easily obtained that

$$< R_e^2 >= N < b^2 > . \tag{3.19}$$

- Another probable important role, which has not been considered yet, is the stiffness of a polymer. In the continuous limit ($N \to \infty, b \to 0, Nb \to L$) of the *Worm-like chain* model, the bending energy of the polymer is calculated via

$$H = \frac{B}{2}\int_0^L (\frac{\partial u(l)}{\partial l})^2 \, \mathrm{d}l \tag{3.20}$$

with the bending modulus B proportional to the persistence length, $B = k_BT\xi_p$. The mean squared end-to-end distance could be derived as

$$< R_e^2 >= 2L\xi_p(1 - \frac{\xi_p}{L}(1 - \exp(-\frac{L}{\xi_p}))) = L^2 f_D(\frac{L}{\xi_p}), \tag{3.21}$$

where the Debye function $f_D(x) = \frac{2}{x^2}(x - 1 + \exp(-x))$. For very short chain $L \ll \xi_p$, we can substitute $\exp(-\frac{L}{\xi_p})$ with $1 - \frac{L}{\xi_p} + \frac{L}{\xi_p}^2$, and get $< R_e^2 >= L^2$. This means that the chain does not bend at all. For very long chain $L \gg \xi_p$, we get $< R_e^2 >= 2L\xi_p \sim N$, which is the same as the scaling relation in freely-joined chain model. Together this shows that the effect of the stiffness of a polymer depends on the polymer length (or the coarse-grained scale).

- Last but not least, the excluded volume interaction is necessary to be considered in most applications. Each monomer occupies certain space

exclusively, and the polymer is also called *self-avoiding chain*. Flory first gave a theoretical solution to self-avoiding chain in 1949 [54]. The total free energy F of a polymer with end-to-end vector $\boldsymbol{R}_e$ is made up of two parts. An entropy part F_1 equals

$$F_1(\boldsymbol{R}_e) = -TS = -Tk_B \log(p(\boldsymbol{R}_e)) = k_B T \frac{3}{2Nb^2} R_e^2 + const., \quad (3.22)$$

where Equation (2.18) is applied in the last step. The second part F_2 is the energy of excluded volume interactions. When we consider only pairwise interactions,

$$F_2(\boldsymbol{R}_e) \simeq \beta k_B T c^2 R_e^3 = \beta k_B T \frac{N^2}{R_e^3}, \quad (3.23)$$

where $c \simeq \frac{N}{R_e^3}$ is the local monomer concentration and β is the strength for the excluded volume interactions. Minimizing $F = F_1 + F_2$ with respect to R_e yields

$$R_e^2 \sim N^{2\nu} = N^{2 \cdot \frac{3}{5}}. \quad (3.24)$$

Numerical simulations estimated $\nu = 0.588$, which is quite close to the Flory's result.

Besides these general models, many more specific polymer models have been designed to simulate biological macromolecules which are essential for all known form of life. Regarding adenine, thymine, cytosine and guanine as four types of monomers, a single-strand DNA molecule can for example be considered as a polymer on a coarse-grained scale. To understand the spatial organization and interactions of chromatin fibers with accessible computing power nowadays, the chromatin fiber of over 100 Mbp need to be modeled via a more coarse-grained polymer. As an example, Figure 3.3(a) shows the mean squared spatial distance

as a function of the genomic distance given by the *random loop model* [55]. The random loop model was developed from the Gaussian chain model. Except for the harmonic springs connecting successive monomers, additional springs connect monomer i and j with a probability p. Therefore, the Hamiltonian of the polymer has the form

$$H = \frac{\kappa_{backbone}}{2} \sum_{i-1}^{N} |\boldsymbol{r}_i - \boldsymbol{r}_{i-1}|^2 + \frac{\kappa_{i,j}}{2} \sum_{j>i+1}^{N} |\boldsymbol{r}_i - \boldsymbol{r}_j|^2, \tag{3.25}$$

where

$$\kappa_{i,j} = \begin{cases} \kappa_{loop} & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases} \tag{3.26}$$

In Figure 3.3(a) this homogeneous model produces a plateau in an intermediate range of genome distance in the $< R_n^2 >$ profile. This was observed in fluorescence in situ hybridization (FISH) experiment in human interphase cells [10], but can not explained by any general model mentioned above (Equation (2.16)(2.19)(2.21)(2.24)).

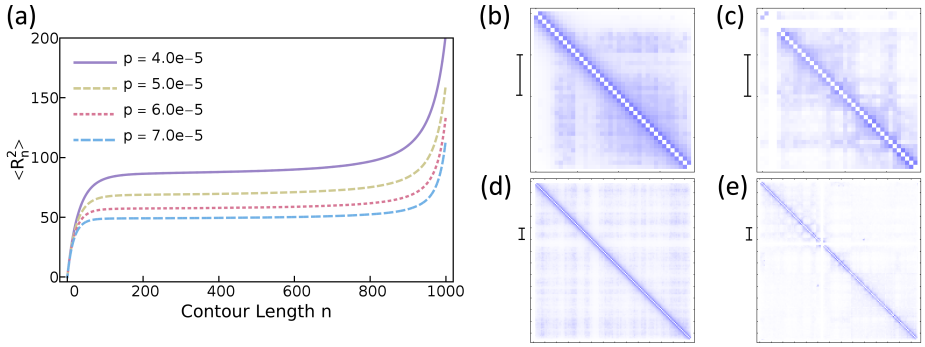


Figure 3.3: Random loop model. (a) Mean squared spatial distance versus genomic distance with different looping probabilities (Equation (2.26)) [55]. (b) Intra-chromatin interactions of human chromosome 21 (b) and 11 (d), compared to the experimental data (c) and (e) respectively. The vertical bars in (b-e) represent 10 Mbp in length.

Inhomogeneity could be easily introduced into this model. By determining

$$\kappa_{i,j} = \begin{cases} \frac{\kappa_{loop}}{2}(c_i^{ctcf} c_j^{cohesin} + c_j^{ctcf} c_i^{cohesin}) & \text{with probability } p, \\ 0 & \text{with probability } 1-p, \end{cases} \tag{3.27}$$

where c_i^A is the binding site density of protein A at genome loci i (of bins 1 Mbp), the intra-chromatin interaction frequencies can be plot as a heat map for human chromosome 21 (b) and 11 (d), with comparison to the experimental data (c) and (e) respectively. Based on this model, by explicitly taking excluded volume interaction into consideration, *dynamic loop* model [10] (with restricted interaction range [13, 14]) could be applied to describe the chromatins in interphase (in mitosis).

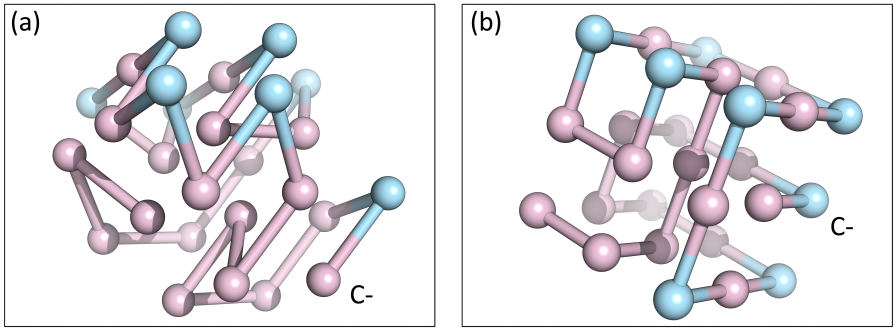


Figure 3.4: One optimal structure of protein of sequence HHHHHPHHPHPHPHPHPHPHHHHHHPH predicted by HP model in a face centered cubic lattice (a) and in a simple cubic lattice (b) [56, 57]. Hydrophobic residues are colored red, and polar ones are colored blue.

Proteins, which is mostly composed of 22 kinds of amino acids, can also be thought of an inhomogeneous polymer. One example is *hydrophobic polar (HP) model*, which is developed by Lau and Dill [58] to study the protein folding problem (see Figure 3.4). Each amino acid is represented by one monomer, either hydrophobic(H) or polar(P). The Hamiltonian of the protein is given by

the H-H contacts of non-successive monomers. With $s_i = \{1, 0\}$ for monomer i is H or P, it can be formulated as

$$H = -J \sum_{|i-j|>1}^{N} s_i s_j \tag{3.28}$$

for all monomer pairs $\{i, j\}$ which contacts with each other. Figure 3.4 shows one optimal structure of an amino acid sequence HHHHHPHHPHPHPHPHP-PHHHHHHPH predicted by HP model in a face centered cubic lattice (a) and in a cubic lattice (b) [56, 57]. Hydrophobic and polar residues are colored red and blue, respectively. It is clear that hydrophobic residues form a hydrophobic core which is more or less shielded by the polar residues on the protein-water interface. There are lots of possible modifications on HP model. Many other polymer models of protein, such as Gō-type model [59, 60, 61], are not discussed here.

So far we have discussed about the principle of Monte Carlo algorithm and various models of polymers or biological macromolecules. Before one starts simulating polymers using Monte Carlo method, one question still need to be solved is how we try to change the conformation of a polymer in each step. In other words, what is our Monte Carlo *trial move*? The answer to this question depends on many properties of the system of interest, such as dilute or dense, homogeneous or inhomogeneous. Here two examples are illustrated.

The first one is the so-called *bond fluctuation* model, which was originally designed by Carmesin and Kremer [62, 63]. In a $3d$ simple cubic lattice of unit spacing (see Figure 3.5(a)), each monomer occupies a cubic grid cell so that eight vertices of the cell are blocked for occupation by other monomers. This guarantees the excluded volume constraint and leads to a minimum bond length $b_{min} = 2$. The Monte Carlo trial move is set to ± 1 in either $\{x, y, z\}$ direction. To further preserve the topology of the polymer, i.e., no bonds intersection, a

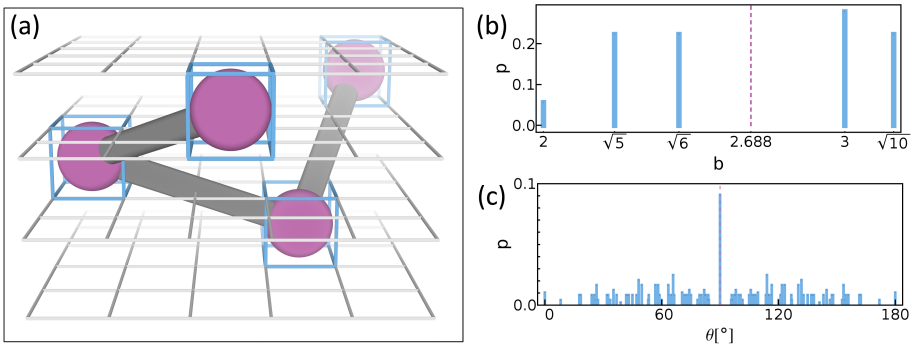


Figure 3.5: Bond fluctuation model. (a) Every monomer occupies a grid cell (the blue framed cubic), so that eight vertices of the cell is blocked from occupation by other monomers. *Priori* probability distribution of the allowed bond length b (b) and the allowed bond angle θ (c). The vertical dashed line in (b)(c) represents the mean value.

set of allowed bond vectors $\{\boldsymbol{b}\}$ could be derived from

$$\{\boldsymbol{b}\} = \boldsymbol{P}(2,0,0) \cup \boldsymbol{P}(2,1,0) \cup \boldsymbol{P}(2,1,1) \cup \boldsymbol{P}(2,2,1) \cup \boldsymbol{P}(3,0,0) \cup \boldsymbol{P}(3,1,0), \quad (3.29)$$

where $\boldsymbol{P}(\delta x, \delta y, \delta z)$ represents the set of all permutations and sign combinations of $\pm\delta x, \pm\delta y, \pm\delta z$. The set $\{\boldsymbol{b}\}$ contains 108 bond vectors, and results in a set of *fluctuating* bond length $\{b\} = \{2, \sqrt{5}, \sqrt{6}, 3, \sqrt{10}\}$. The *priori* probability distribution for bond length b is plotted in Figure 3.5(b), with its mean value $< b >= 2.688$ indicated by a vertical dashed line. The *priori* probability distribution for bond angle θ formed by successive two bonds is shown in Figure 3.5(c). Although there exist only two angles values $0, \pi$ in the underlying lattice, θ have many possible choices in the range $[0, \pi]$. The distribution is symmetric around the mean bond angle $< \theta >= \pi/2$. Bond fluctuation model has been used a lot in simulating polymers in lattice. It is also easy to make extensions based on it, such as the dynamic loop model and work in Chapter 6.

The second example is to simulate a semiflexible chain adsorbed on a sphere in continuous space (off-lattice) [64]. Two kinds of Monte Carlo moves are

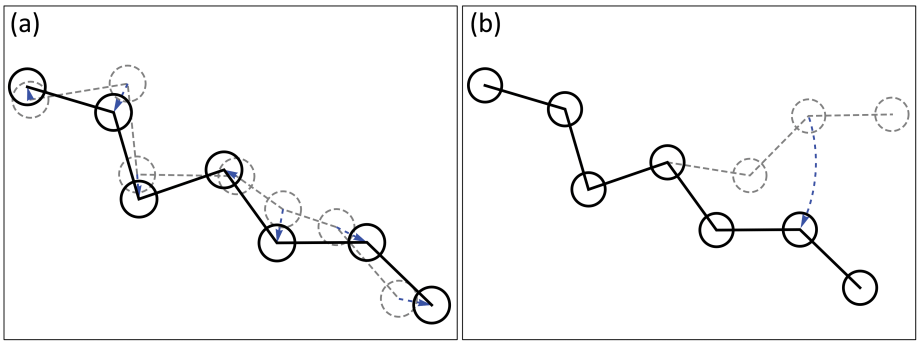


Figure 3.6: Two kinds of Monte Carlo trial moves for polymer simulated in continuous space [64]. (a) Translate each monomer in a random direction by a random distance, which changes the polymer conformation locally. (b) Pivot either tail of the polymer by a random angle, which changes the polymer conformation globally. The old, new conformation is plotted with dashed, solid line respectively.

adopted (see Figure 3.6). One *local* trial move try to translate a monomer in $3d$ space (a) by $\boldsymbol{\Delta r} = s(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, where s, θ, ϕ are randomly chosen from $[0, s_{max}]$, $[0, \pi]$ and $[0, 2\pi]$ respectively. This is similar to the trial move used in the bond fluctuation model. Trajectories generated using local moves might capture some dynamic properties of the system. In contrast another trial move is *global*, which pivots either tail of the polymer by a randomly chosen angle. Sometimes global move greatly reduces the correlation between successive polymer conformations, so that the computation demand decreases. However it could lead to a small acceptance ratio of trial moves in some cases, e.g., a very compact system. Therefore it is quite often to take advantages of both kinds of moves, and to apply them together.

Figure 3.7 shows the initialization stage of a simulation for a polymer of large stiffness. In equilibration, the polymer is in fully contact with the curved surface and wraps around the whole sphere (d). The trajectory (a-d) does not reveal the *real* adsorption process of the polymer. Ensemble statistics will be done on snapshots taken only after this initialization stage.

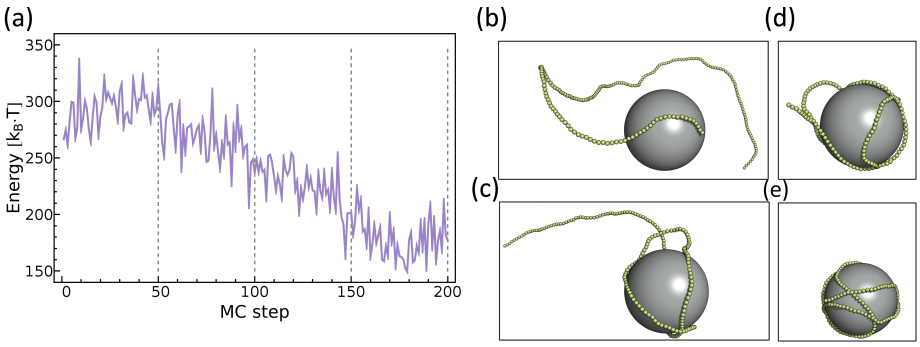


Figure 3.7: Hamiltonian (a) and configurations (b-e) of a semiflexible chain adsorbed on a sphere simulated via Monte Carlo method in continuous space. Snapshots are taken at Monte Carlo step $t = \{50, 100, 150, 200\}$ in order.

3.2 Molecular Dynamics Simulation

Molecular dynamics simulation method is a more preferable choice than Monte Carlo method when dynamic properties, such as transport coefficient, are concerning.

3.2.1 Integrators

For Monte Carlo method, a designed trial move propagates the system while the velocities of particles in the system do not explicitly play any role. For molecular dynamics simulation, Hamiltonian of both coordinate and velocity $H(\boldsymbol{r}, \boldsymbol{p})$ controls the evolution of the system containing N particles, according to the Newton's law

$$\dot{\boldsymbol{r}}_i = \boldsymbol{p}_i / m_i, \tag{3.30}$$

$$\dot{\boldsymbol{p}}_i = \boldsymbol{f}_i, \tag{3.31}$$

for all $i \in \{1, 2, ..., N\}$. There are several ways , so-called *integrator*s, to implement these two equations on computer. Two most widely used algorithms are introduced here. The first one is *leap-frog* algorithm, which is the default integrator in the simulation software GROMACS [65]. It updates $\{\boldsymbol{r}_i, \boldsymbol{p}_i\}$ via

$$\boldsymbol{p}_i(t + \frac{1}{2}\delta t) = \boldsymbol{p}_i(t - \frac{1}{2}\delta t) + \delta t \boldsymbol{f}_i(t) \tag{3.32}$$

$$\boldsymbol{r}_i(t + \delta t) = \boldsymbol{r}_i(t) + \delta t \boldsymbol{p}_i(t + \frac{1}{2}\delta t)/m_i. \tag{3.33}$$

The other integrator is called *velocity verlet* algorithm, which is the default integrator in the simulation package ESPResSo [66, 67] and can be written as

$$\boldsymbol{p}_i(t + \frac{1}{2}\delta t) = \boldsymbol{p}_i(t) + \frac{1}{2}\delta t \boldsymbol{f}_i(t) \tag{3.34}$$

$$\boldsymbol{r}_i(t + \delta t) = \boldsymbol{r}_i(t) + \delta t \boldsymbol{p}_i(t + \frac{1}{2}\delta t)/m_i \tag{3.35}$$

$$\boldsymbol{p}_i(t + \delta t) = \boldsymbol{p}_i(t + \frac{1}{2}\delta t) + \frac{1}{2}\delta t \boldsymbol{f}_i(t + \delta t). \tag{3.36}$$

Two algorithm are literally equivalent, if the initial condition $\{\boldsymbol{r}_i(0), \boldsymbol{p}_i(0)\}$, couplings and constraints are not considered. Both of them are time reversible and of third order of δt for $\boldsymbol{r}_i$, i.e.,

$$\boldsymbol{r}_i(t + \delta t) = 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t - \delta t) + \delta t^2 \boldsymbol{f}_i(t)/m_i + \boldsymbol{O}(\delta t^4). \tag{3.37}$$

In practice, given a single initial point, they will yield different trajectories. Figure 3.8 shows the difference between the step schemes of these two integrators.

3.2.2 Thermostats

Since the Hamiltonian is conserved in the Newtonian scheme, disregarding the computation error, a trajectory obtained by basic leap-frog or velocity verlet algorithm maps to a *microcanonical ensemble* (NVE). However this is not the

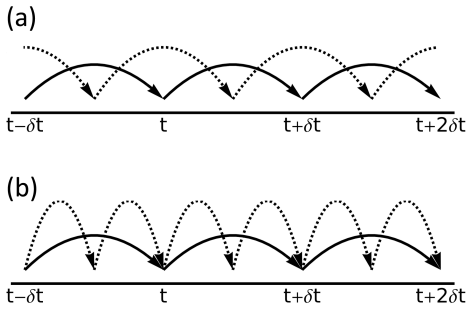


Figure 3.8: Integration scheme for (a) leap-frog algorithm and (b) velocity verlet algorithm. The solid line represents the propagation of $\boldsymbol{r}_i(t)$ and the dashed line represents the propagation of $\boldsymbol{p}_i(t)$.

condition under which most experiments are performed. The *canonical ensemble* (NVT), *isothermal-isobaric ensemble* (NPT) and *grand-canonical ensemble* (μVT) are more suited to imitate real experiment. We focus on the canonical ensemble and briefly discuss how to modify the integration scheme to satisfy the constraint on temperature in molecular dynamics simulations (so-called *thermostat*). In principle, a thermostat works by coupling the system of our interest, which has an instantaneous temperature $\mathcal{T}(t) \propto \sum_i m_i \dot{\boldsymbol{r}}_i^2$, to a heat bath of a different temperature T_0. Heat transfers back and forth between the heat bath and the system so that $\mathcal{T}$ has certain designed properties. Depending on the coupling scheme, there are different thermostates [68, 69, 70], such as velocity rescaling coupling, Andersen coupling and Nosé-Hoover coupling. *Berendsen thermostat* and *Langevin thermostat* will be introduced in the following, which are applied in GROMACS and ESPResSo, respectively.

The Berendsen thermostat also belongs to the velocity rescaling coupling methods. The corresponding equation of motion can be written as

$$\dot{\boldsymbol{p}}_i = \boldsymbol{f}_i - \frac{1}{2\tau}\left(\frac{T_0}{\mathcal{T}(t)} - 1\right)\boldsymbol{p}_i. \tag{3.38}$$

According to this scheme, one can show that the instantaneous temperature $\mathcal{T}(t)$ changes following

$$\dot{\mathcal{T}}(t) = (T_0 - \mathcal{T}(t))/\tau. \tag{3.39}$$

Thus the coupling strength parameter τ in Equation (2.38) describes how fast $\mathcal{T}(t)$ relaxes to T_0. If this parameter is too large $\tau \to \infty$, the heat flows into and out of the system so slowly that the thermostat is actually disabled. Then Equation (2.38) will reduce to Equation (2.31), which leads to a microcanonical ensemble sampling. On the other hand, a too small τ will results in an unrealistic small temperature fluctuation. In atomistic simulations, it is typically set $\tau \approx$ 0.1 ps. On should notice that the Berendsen thermostat does *not* produce a canonical ensemble in general. One exception of this thermostat with $\tau = \delta t$, which generates a canonical distribution of configurations (but not momenta).

The Langevin thermostat, in contrast, produces a trajectory which converges to a canonical distribution. It relies on the Langevin equation of motion

$$\dot{\boldsymbol{p}}_i = \boldsymbol{f}_i - \gamma_i \boldsymbol{p}_i + \boldsymbol{w}_i(t), \tag{3.40}$$

where $\boldsymbol{w}_i$ is a stochastic force and γ_i is the atomic friction coefficient. To obtain a canonical ensemble of system temperature T_0, it has been shown that $\boldsymbol{w}_i(t)$ must be uncorrelated with $\boldsymbol{p}_i(t')$ and $\boldsymbol{f}_i(t')$ for $t' < t$, and have following properties

$$< \boldsymbol{w}_i(t) > = 0, \tag{3.41}$$

$$< \boldsymbol{w}_i(t)\boldsymbol{w}_j(t') > = 2m_i\gamma_i k_B T_0 \delta_{ij}\delta(t' - t). \tag{3.42}$$

If $\gamma_i = \gamma \, \forall i$, on an intermediate timescale (short compared to the experimental timescale, but long compared to the time separating atomic collisions), the effect

of the thermostat can be formulated as

$$\dot{\mathcal{T}}(t) = 2\gamma(T_0 - \mathcal{T}(t)). \tag{3.43}$$

Comparing Equation (2.43) with (2.39), it is clear that γ controls the rate of energy exchange between the heat bath and the system (like τ), and its value should be set within a proper range.

Besides the integration scheme and the temperature control, many other techniques, such as implementation of additional constraints (e.g., constraints on bond length or bond angle), the boundary condition and calculation of long-range forces etc., are all important to set up an appropriate and efficient molecular dynamics simulation. However even a simulation is carefully performed, we still have problems on how to interpret its result. Does the generated trajectory accurately predict the "*real*" trajectory of macromolecules in our cell? The answer is *no*. Then is the trajectory close to the "*real*" trajectory? The answer is *no* one has proved it so far, but it is generally trusted *on belief* in spite of the so-called *Lyapunov instability* beneath all molecular dynamics simulation (see Figure 3.9) [69].

In order to reduce computational demand, quite often biological macromolecules are simulated without explicit surrounding solvent molecules. In an implicit solvent simulation, motion of the macromolecules of interest are described by the *Langevin dynamics* (see Equation (3.40)). What's more, in the overdamped limit, i.e.,

$$0 = \boldsymbol{f}_i - \gamma_i \boldsymbol{p}_i + \boldsymbol{w}_i \tag{3.44}$$

or

$$\boldsymbol{p}_i = \boldsymbol{f}_i/\gamma + \boldsymbol{w}_i/\gamma, \tag{3.45}$$

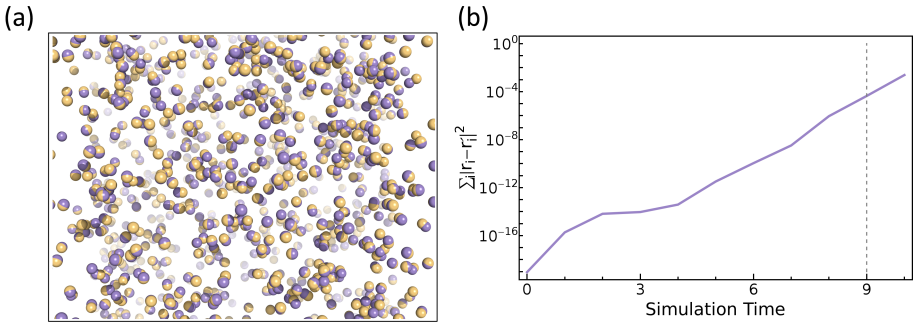


Figure 3.9: Lyapunov instability. Two molecular dynamics simulations for 10^3 particles, interacting with each other via Lennard-Jones potential of depth 1.0 $/k_BT$ and equilibrium distance $2^{1/6}$, are performed in a simulation box of dimensions 40^3. They start from an identical equilibrated state, except that the x component of the velocities of two particles in the second run are changed by $\pm 10^{-10}$. (a) The onset of the snapshots of discernible difference between two trajectories (in different colors). (b) The difference $\sum_i |\boldsymbol{r}_i - \boldsymbol{r'}_i|^2$ grows exponentially with simulation time. Image adapted from [69].

the Langevin motion becomes Brownian motion (see Figure 3.10).

By defining $\zeta = \gamma m$, and using the Einstein relation $D = k_BT/\zeta$, the equation of motion for the *Brownian dynamics* then can be rewritten as

$$\dot{\boldsymbol{r}} = \frac{D}{k_BT}\boldsymbol{f} + \boldsymbol{w}', \tag{3.46}$$

where D is the diffusion coefficient and $\boldsymbol{w}'$ is stochastic forces of variance $\sqrt{2D}$. Compared with molecular dynamics, Brownian dynamics can access longer timescales and simulate system of larger spatial scales.

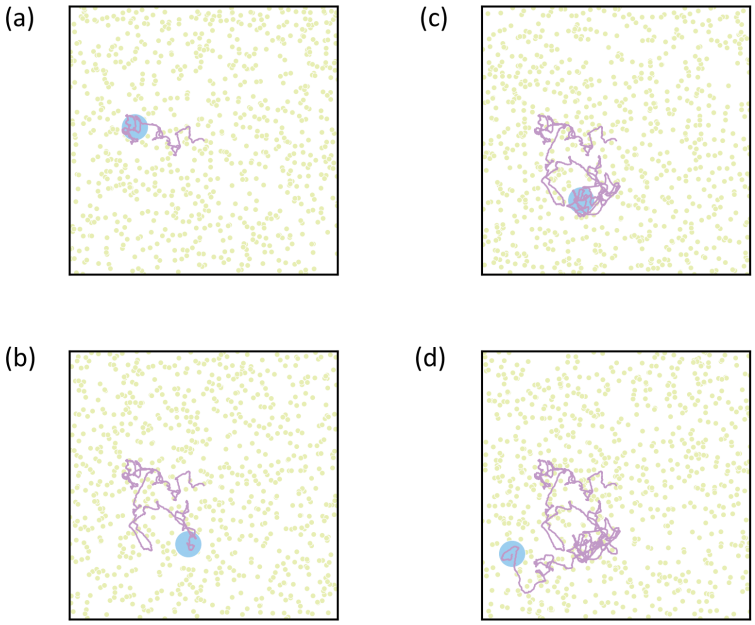


Figure 3.10: Brownian motion. A simulated trajectory of a pollen grain (the large blue dot) suspended in water. Velocity of the pollen grain is changed by elastic collisions with surrounding water molecules (small green dots).

4 Unbound multi-Cys_2His_2 zinc finger protein

Zinc finger proteins and the 3D organization of chromosomes, Christoph J Feinauer, Andreas Hofmann, Sebastian Goldt, Lei Liu, Gabriell Mate and Dieter W Heermann, Advances in Protein Chemistry and Structural Biology. Published. Lei Liu's contribution is model for CTCF.

A multiscale approach to simulating the conformational properties of unbound multi-Cys_2His_2 zinc finger proteins, Lei Liu, Rebecca C Wade and Dieter W Heermann, In preparation.

Chapter Summary. The conformational properties of unbound multi-Cys_2His_2 (mC_2H_2) zinc finger proteins are studied using a multiscale approach. Three methods on different length scales are utilized. First, atomistic molecular dynamic simulations confirmed that the zinc finger is more rigid than the most typical linker. Second, the end-to-end distance distributions of mC_2H_2 zinc finger proteins are computed using a more efficient pivoting algorithm, which only takes excluded volume interaction into consideration. The end-to-end distance distribution gradually changes its profile, from left-skewed to right-skewed, as the number of the zinc fingers increases. We explained this with a worm-like chain model. For proteins of a few zinc fingers, an effective bending constraint favors an extended conformation. Only for proteins containing more than nine zinc fingers, a somehow compact conformation is preferred. Third, a mesoscale model is modified to study both the local and global conformational properties of mC_2H_2 zinc finger proteins. Simulations of the protein CCCTC-binding factor (CTCF), which includes ten C_2H_2 and one C_2HC zinc fingers, on the molecular level are presented.

4.1 Introduction

Multi-Cys_2His_2 (mC_2H_2) zinc finger proteins are composed of tandem repeats of a Cys_2His_2 (C_2H_2) zinc finger motif connected by short flexible linkers. They form a class of transcription factors which control a wide range of cellular processes [71, 72]. Each C_2H_2 finger has a general sequence motif X_2-C-$X_{2\text{-}4}$-C-X_{12}-H-$X_{3\text{-}5}$-H, and forms a $\beta - \beta - \alpha$ secondary structure motif with a central coordinating Zn^{2+} ion. There are few researches on the conformational property of unbound mC_2H_2 zinc finger protein, which might be helpful for understanding its binding mechanism to DNA, RNA and other proteins.

Comparison of the NMR data of the three linkers in the Wilms' tumor zinc fingers transcription factor in complex with DNA and in solution indicated that the whole protein would be segmentally disordered, while the individual zinc fingers are structured [73]. Following this idea, multi-zinc finger proteins are usually classified as intrinsically disordered proteins, that "fold while binding", and form a globally disordered chain when unbound [42, 74]. On the one hand, NMR relaxation measurements for the first three zinc fingers of the Xenopus transcription factor TFIIIA showed that on time scales shorter than 10 ns, the motions of individual zinc fingers are highly correlated, and that the average end-to-end distance of the polypeptide chain is longer than that in its crystallographic conformation bound with DNA [75]. On the other hand, an attraction between the zinc finger in protein GATA-1 and the zinc finger in protein SP1 or in freind of GATA-1 (FOG) was observed in both isothermal calorimetric titration and CD spectra [76, 77]. Assuming interactions between zinc fingers are always attractive and strong, zinc fingers will aggregate and the unbound the polypeptide chain will collapse, which are contradictory to the NMR observation.

Among various mC_2H_2 zinc finger proteins, the CCCTC-binding factor, or CTCF, is of special interest because of its potential function in genome organization. CTCF contains tandem 10 C_2H_2 and 1 C_2HC zinc fingers. As a ubiquitous transcription factor in eukaryotes, more than 13,000 CTCF binding sites have been identified experimentally in human genome [35, 78, 22]. Since first isolated from chicken in 1990 [26], CTCF has been reported to play several roles in gene regulation in different contexts, such as promoter repression, activation, enhancer blocking, X-chromosome inactivation and genomic imprinting. Recently, in applications of chromosome conformation capture techniques, CTCF binding sites genome-wide correlate with both intra- and inter-chromosome interactions. It has become the most compelling candidate for the genome architecturer, and has been dubbed as "The Master Weaver of the Genome" [4]. Several CTCF-mediated chromosome loop models have been proposed [4, 10, 79]. However, due to the flexibility of the mC_2H_2 zinc finger protein, it is not easy to determine its conformational ensemble in experiments, and an understanding of unbound CTCF dynamics is still lacking.

In this chapter, we want to investigate the conformational space of the unbound mC_2H_2 zinc finger protein. Does it collapse into a globular conformation due to the interactions between zinc fingers, curl in a coil due to the intrinsically disordered linkers, or extend like a straight rod? It is difficult to study the complete CTCF protein in solution using atomistic detailed simulation because of its large molecular weight (over 82 kDa) and expected long relaxation time. A multiscale approach extending from atomic to mesoscale was developed for this purpose.

4.2 Materials and Methods

Three approaches on different length scales are utilized in this study: 1) for proteins composed of single zinc finger or of three zinc fingers, atomistic molecular dynamics (MD) simulations with explicit water and ions are applied; 2) for proteins containing more zinc fingers, i.e. for CTCF, a pivot algorithm is devised, which only takes excluded volume interaction into consideration; 3) finally, a mesoscale polypeptide model with implicit solvent is constructed for both single- and multi-C_2H_2 proteins. The reference codes for the different systems studied in this work are composed of three parts: the PDB code, the index of the first zinc finger and the index of the last zinc finger. These codes and the corresponding names of studied polypeptides are listed in Table 4.1. Besides these zinc finger proteins, the most typical linker (the so called conserved linker) is labeled "cLinker", which is a short peptide of sequence TGEKP. For the sake of brevity, we use ZFi to stand for the i-th zinc finger in the following.

Table 4.1: Reference codes and corresponding protein names.

Reference code[a]	Protein name	Indices of zinc fingers[b]	Number of residues
1aay_1	Zif268	1	32
1tf3_1_3	TFIIIA	1∼3	91
ctcf_4_8	CTCF	4∼8	150
ctcf_3_9	CTCF	3∼9	206
ctcf_2_10	CTCF	2∼10	267
ctcf_1_11	CTCF	1∼11	314

[a] The reference code is composed of three parts: the PDB code, the index of the first zinc finger, and the index of the last zinc finger. For example, 1tf3_1_3 represents a polypeptide containing the N-terminal three zinc fingers in the protein structure of PDB code 1tf3.

[b] Indices of zinc fingers are the indices in the protein structure deposited in PDB.

4.2.1 Atomic Simulations

As a protein of 727 amino acid residues (AA), CTCF contains 10 C_2H_2 and 1 C_2HC zinc finger, with linkers of 5∼8 AA between them. The aligned sequence motifs with conserved Cys and His residues, which coordinate the Zn^{2+} ion, are shown in Figure 4.1 (a). The sequence disorder propensities predicted by DisEMBL, DISOPRED2 and GLOBPLOT2 [80, 81, 82] of the entire peptide chain are shown in Figure 4.1 (b), with the 11 zinc fingers labeled in gray. It shows that the tandem zinc finger central segment is flanked by two intrinsically unstructured tails, of about 265 and 148 AA each, at the N and C termini. Only the central multi-zinc finger domain is studied here.

The initial conformation of single zinc finger for MD simulation was taken from a crystal structure of protein Zif268 (PDB code: 1AAY), which probably is the most frequently studied mC_2H_2 zinc finger protein. The NMR resolved structure of ZF1∼ZF3 in protein TFIIIA (PDB code: 1TF3), whose unbound conformational properties have been investigated by another NMR research, was chosen as the initial conformation of a polypeptide with three zinc fingers. For CTCF, NMR resolved atomic structures of ZF6∼ZF7 and ZF10∼ZF11, each are contained in PDB file 2CT1 and 1X6H. Other zinc fingers were built using a homology modeling approach with the MODELLER 9.10 program [83]. The zinc finger structure templates were chosen from non-redundant PDB sequences of $< 95\%$ sequence identity, that facilitates to derive unbiased statistics on their structure and evolution. Single or multiple templates used for the modeling had at least 35% sequence identity to the target CTCF zinc finger. Compared with the templates, the generated homology models have backbone root mean square deviations (RMSD) ≤ 1 Å, and have Z-scores ≥ 3.5 (see Table 4.2). A typical conformation is rendered in Figure 4.2.

Standard all atom MD simulations were applied to 1aay_1 and 1tf3_1_3.

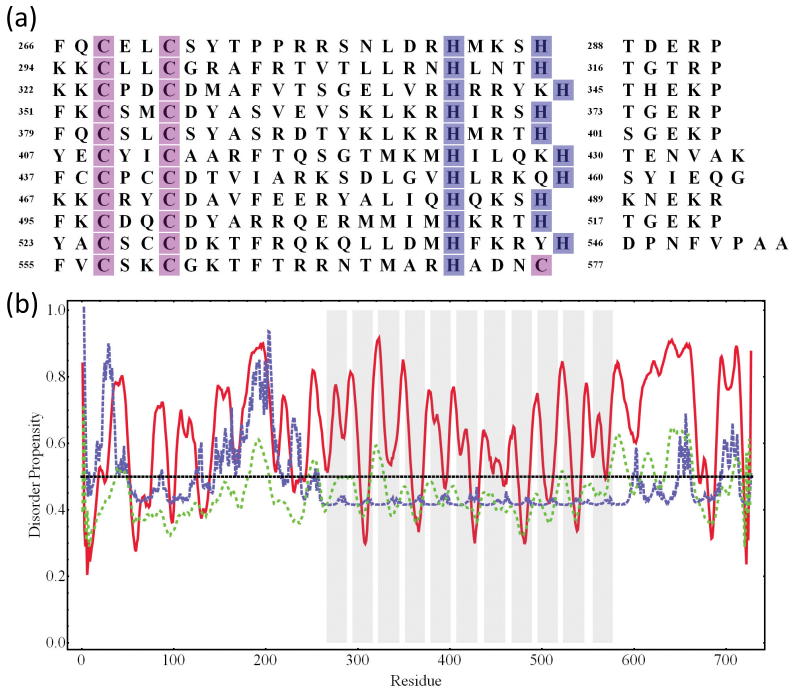


Figure 4.1: Sequence properties of the central zinc finger containing domain of CTCF. (a) Zinc finger and linker sequences of CTCF, with Cys and His residues colored purple and blue, which coordinates the Zn^{2+} ion. (b) Sequence disorder propensities predicted by DisEMBL (solid red), DISOPRED2 (dashed blue), and GLOBPLOT2 (dotted green), with zinc fingers labeled gray.

All calculations were carried out using the GROMACS software [65] with the Amber force field Amber99SB [84]. The Zn^{2+} ion type, which plays a key role in folding and stabilizing the $\beta - \beta - alpha$ structure motif in each zinc finger, was simulated using the Cationic Dummy Atom (CaDA) method [85]. One tetrahedron-shaped zinc divalent cation was represented by four peripheral cationic dummy atoms, which interact with other atoms electrostatically but not sterically and impose the requisite orientational requirement for the zinc four-ligand coordination (Figure 4.3 (a)).

The system was first energetically minimized to remove unfavorable contacts.

Table 4.2: Parameters of homology modeling for CTCF zinc fingers.

Zinc finger	Template PDB code[a]	Tempalte/Target		
		Sequence identity[b] (%)	RMSD[c] (Å)	Z-score[d]
ZF1	1SRK	40	0.55	3.7
	1X5W	48	0.75	3.5
ZF2	2EN4	54	0.94	3.7
	2KMK	50	0.36	3.9
ZF3	2EME	56	0.58	3.7
	2KMK	46	1.00	3.7
ZF4	2DMD	62	0.50	3.7
	2ELQ	52	0.68	3.9
ZF5	2DMD	50	0.83	3.7
	2ELQ	60	0.32	3.9
ZF8	2ELS	35	0.87	3.7
	2EM3	42	1.01	3.7
ZF9	2DMD	48	0.43	3.7

[a] Corresponding PDB code of the structure template.

[b] Percentage of sequence identity between template and model.

[c] Backbone RMSD between template and model.

[d] Proteins with similar fold typically have RMSD $\leq$ 1.0 Å, and Z-score $\geq$ 3.5.

Next it was solvated in NaCl solution of defined ionic strength, followed by another energy minimization. Then, to equilibrate the zinc fingers and surrounding water molecules and ions, 30 ps MD at 200 K was carried out with constraints on all bond lengths. This was implemented using the LINCS algorithm with a harmonic constraint force constant of 1.0×10^3 kcal $\cdot$ mol^{-1} $\cdot$ nm^{-2}. Afterward, unconstrained equilibration of 12.5 ps with constant volume and constant target temperature, and 100ps with constant pressure and constant temperature were performed. Finally, 10 ns for 1aay_1, 30 ns for 1tf3_1_3 production runs were carried out with an integration step of 1fs. The temperature and pressure of the system were controlled via the Berendsen algorithm with coupling time constants of 0.1 and 1.0 ps, respectively. Coordinates from generated trajectory were written every 0.5 ps.

With different ionic strengths $\{0.00, 0.01, 0.10\}$ M and different tempera-

Figure 4.2: A homology model of the central zinc fingers domain of CTCF. Consecutive zinc fingers are encapsulated in ellipsoids of alternating colors.

tures $\{300, 330\}$ K, ten independent runs were perfored for 1aay_1 under each environmental condition. For 1tf3_1_3, ten runs were carried out with an ionic strength of 0.1 M and temperature of 300 K.

4.2.2 An atomistic pivoting algorithm

For proteins with more zinc fingers, assuming that each zinc finger does not change its secondary structure and that the internal motion of each zinc finger is not strongly coupled to the collective motion of the whole peptide chain, a pivoting algorithm [86] was devised to take only excluded volume interactions, i.e., atom clashes, into consideration.

As shown in Figure 4.4, we first randomly pick pivoting points along the backbone ($N - C_\alpha$ or $C_\alpha - C'$ bonds) of residues in the linkers of a multi-zinc finger protein of conformation i. Then random rotations are made around these points. The resulting conformation is accepted as a new conformation $i + 1$, if no excluded volume violation is detected. The pivoting process is repeated until

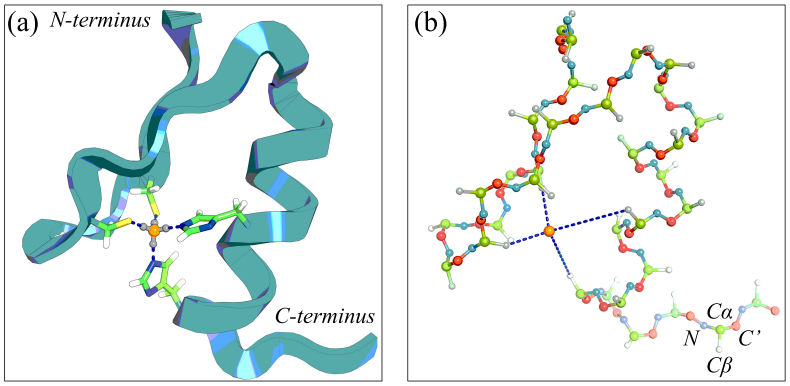


Figure 4.3: Atomic and mesoscale models of 1aay_1. (a) The CaDA representation, where the orange Zn^{2+} ion is surrounded by four gray dummy atoms. (b) Mesoscale model, where each AA is made up of three backbone beads (N, C_α, C') and one side chain bead (C_β). The Zn^{2+} ion is explicitly modeled as one additional bead, which interacts with C_β of its coordinating residues.

a minimum number of conformations are sampled.

This algorithm is much more efficient for sampling regions in conformational space which are unlikely to be visited through naive atomistic MD simulation. What's more, subject to the above assumptions, it provides a first-order approximation to the conformational ensemble of the unbound multi-zinc finger protein under physiological conditions (see results in Section 4.3.2).

4.2.3 Mesoscale Simulations

To further reduce the computational demand while keeping the AA sequence specificity, we adapted a mesoscale model, *peptideB*, developed by Bereau and Deserno, which was designed for protein folding and aggregation [87].

As shown in Figure 4.3 (b), each amino acid in peptideB is modeled by three (Glycine) or four (non-Glycine) beads. These beads represent the amide group N, the central carbon C_α, the carbonyl group C' and the side group C_β. Both bonded and nonbonded interactions have been systematically parameterized (see

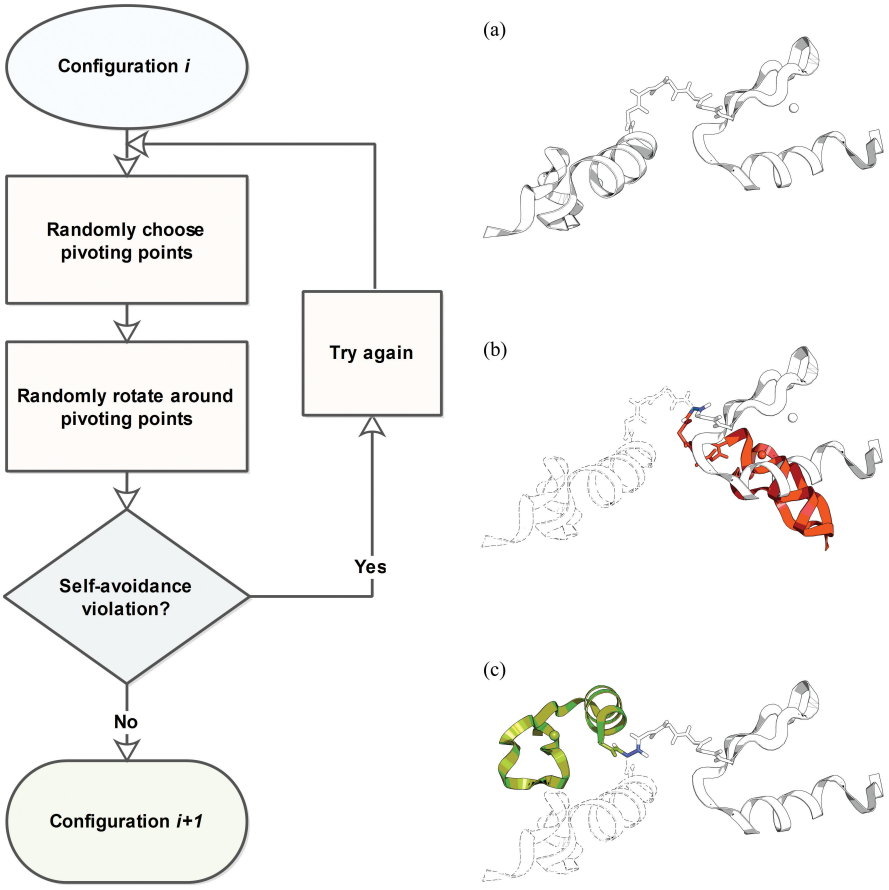


Figure 4.4: Illustration of the atomistic pivoting algorithm. It shows a trial to pivot a residue backbone bond, the blue bond in (b,c), which is randomly chosen from the linkers in conformation i (a). If there are atom clashes (b), reject it and try again. Otherwise it is accepted as a new conformation $i + 1$ (c).

Appendix B). In contrast to other coarse grained protein models, such as the Gō-type model and MARTINI force field [88, 89, 90], which forbid changes of the secondary or tertiary structure of the simulated peptide, the peptideB model does not bias the structure toward any particular secondary motif and hence is suitable for the flexible multi-zinc finger proteins. However, some modifications

are necessary:

(i) Because of the critical role of Zn^{2+} in stabilizing the individual zinc finger motifs, the Zn^{2+} ion is explicitly modeled by an extra bead. Additional G$\bar{o}$-type constraints between the Zn^{2+} and its coordinating residue beads are included as

$$V_{Zn-bond}(r) = \frac{1}{2}\kappa_{Zb}(r - r_0)^2, \tag{4.1}$$

$$V_{Zn-angle}(r) = \frac{1}{2}\kappa_{Za}(\theta - \theta_0)^2. \tag{4.2}$$

The equilibrium distance r_0 and angle θ_0 are set equal to the values obtained from the atomic coordinates of the reference conformation. The spring constants κ_{Zb} and κ_{Za} are tuned to give the best nativeness order parameter Q of single zinc finger, which is defined by $< exp[-(r_{i,j}^{ref} - r_{i,j}^{sim})^2/9] >_{i,j}$. The distance $r_{i,j}$ is measured between a pair of C_α beads, each in residue i and j, in the reference conformation and during a simulation. The average goes over all bead pairs $\{i, j\}$ [91]. The advantage of Q over RMSD is that no structure alignment is involved.

(ii) The hydrophobic interaction between side chain beads is reduced so that the calculated standard free binding energy for a CCHC zinc finger (FOG) binding to a CCCC zinc finger (GTAT-1) get closer to the experimental value (see Appedix C).

With an energy scale $\varepsilon = k_B T_r = 1.38\times10^{-23}\text{J}\cdot\text{K}^{-1}\times300\text{K} \approx 0.6\text{kcal}\cdot\text{mol}^{-1}$ and a time scale $\tau = 0.1$ ps, all mesoscale simulations were carried out using the ESPResSo 3.1.0 package [66, 67]. To avoid bead clashes in the initial conformation, Lennard-Jones forces were capped for 200 cycles of 50 steps, with maximum force strength increasing gradually from 0 to normal force strength. Then, simulations longer than ten nanoseconds (e.g., 15 ns for 1tf3_1_3) with constant simulation box volume and constant temperature $0.5 \sim 0.7\ T_r$ were

performed. The integration step was 0.01 τ, roughly corresponding to 1 fs. The whole system was coupled to a Langevin thermostat with friction coefficient 1.0. Multiple runs from different initial conformations, resulting from the atomistic pivoting algorithm, were carried out for each studied protein to ensure adequate sampling. Then thermodynamic calculations were performed by the weighted histogram analysis method (WHAM) [92, 93], which combines energy histograms from canonical simulations at different temperatures (see also Appendix C). Compared to a single histogram method carried out at one temperature, it is known that WHAM constructs a more precise free energy landscape.

4.3 Results and Discussion

4.3.1 The rigidity of single zinc finger

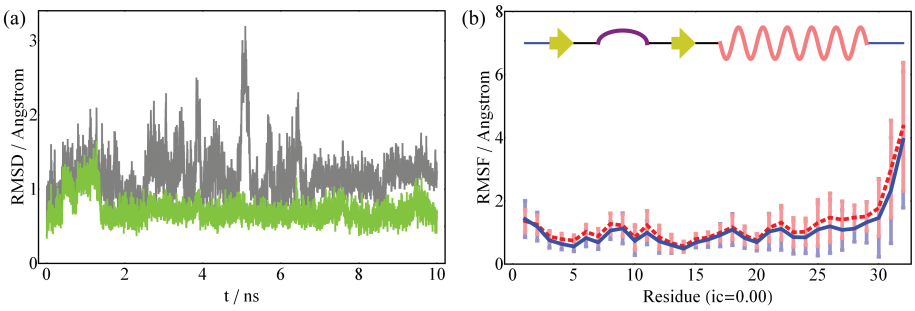


Figure 4.5: RMSD and RMSF of atomic MD simulations of 1aay_1. (a) RMSD against simulation time of the whole protein (gray) and only the zinc finger (green). (b) RMSF of each residue at 330 K (dashed red line) and 300 K (solid blue line), with the secondary structure shown above. The error bar is calculated over independent runs.

Compared to the linker, the rigidity of the zinc finger was verified by all-atom MD simulation. A typical RMSD of C_α atoms in 1aay_1 versus the simulation time, where the protein trajectory was first aligned with the crystallographic

structure, is shown in Figure 4.5 (a). The deviation of the whole peptide is 2∼4 fold greater than that for only one zinc finger. The difference between AAs with specific secondary structure and AAs without stable secondary structure can be analyzed in more detail from the root mean square fluctuation (RMSF) calculated for each residue, as shown in Figure 4.5 (b). As expected, AAs in α-helix or β-strands show low fluctuations, while AAs in the linker at the C-terminus show much higher fluctuations. In addition, we performed a principle component analysis (PCA) on the protein trajectory [94]. The high frequency internal motion modes and the low frequency collective motion modes have distinct amplitude patterns for AAs of zinc finger and for AAs of linker, which indicates the coupling of the collective motion of whole peptide chain and the internal motion of zinc finger is weak.

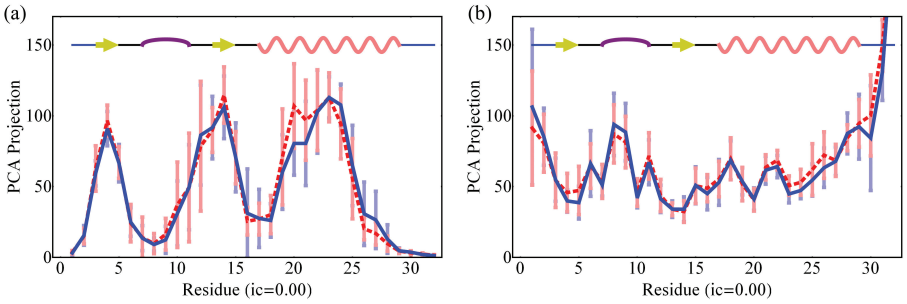


Figure 4.6: Principle component analysis of 1aay_1 trajectories. (a) Fastest five modes projection by residue at 330 K (dashed red) and 300 K (solid blue). The Zn^{2+} coordinating AAs {Cys5, Cys10, His23, His27} and the hydrophobic core above Zn^{2+} ion {Ala4, Phe14, Leu20}, which show high frequency fluctuations, play a critical role in stabilizing the $\beta - \beta - \alpha$ fold. (b) Slowest five modes projection by residue.

The effects of ionic strength in the surrounding medium and the environmental temperature on the stability of zinc finger were investigated by checking the dependence of radius of gyration (R_g) of zinc fingers (Figure 4.7 (a)) and the number of hydrogen bonds between the protein and solvent (Figure 4.7

(b)) on these factors. Both factors have minor impacts. The small increase of about 0.05 Å of R_g induced by the higher temperature mainly contributes to the elongation along the major axis of zinc finger ellipsoid. The number of surface hydrogen bond varies within 5%. These results reveal that i) the zinc finger is structurally more stable than the linker under the studied conditions, ii) its fast internal motions are decoupled from the slow motions of the linkers. These observations pave the foundation of our atomistic pivoting algorithm.

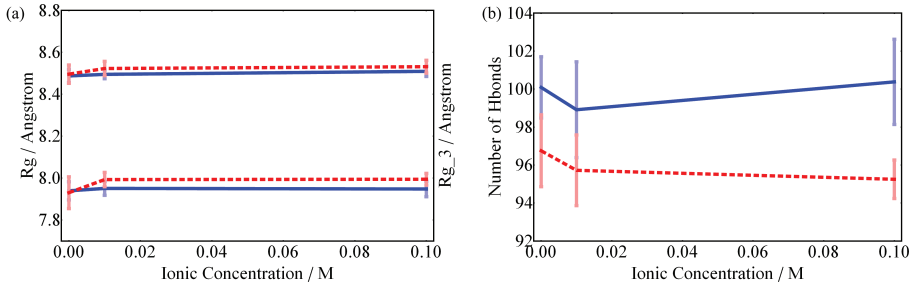


Figure 4.7: Properties of 1aay_1 simulated at variant ionic strengths and at 330 K (dashed red) or 300 K (solid blue). (a) Radius of gyration R_g and the component of R_g along the major axis of the zinc finger, R_g_3. (b) The number of protein-solvent hydrogen bonds.

4.3.2 Conformations of multi-zinc finger proteins

The time dependent end-to-end distance R_e of 1tf3_1_3, calculated from an all-atom MD trajectory, is shown in Figure 4.8 (a). In agreement with the NMR observation, the entire polypeptide chain extends from the initial conformation, i.e., the DNA-bound conformation.

However, a simulation of 30 ns is not long enough to obtain good statistics over the R_e distribution of this protein. As Figure 4.8 (b) shows, (ϕ, ψ) of Gly39 in the first linker (also a conserved linker) of 1tf3_1_3 was still centered in the α_L region, which formed a C-capping motif with Ser35 [95, 43]. This α-helix

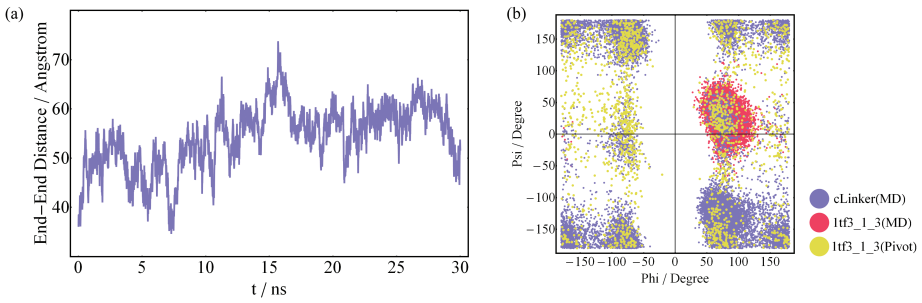


Figure 4.8: All-atom MD simulations of 1tf3_1_3. (a) End-to-end distance R_e versus simulation time. (b) Ramachandran plot of Gly39 in the first linker (also a conserved linker) of 1tf3_1_3, simulated using all-atom MD (red) and atomistic pivoting (yellow). They are compared with the Ramachandran plot of Gly2 in a cLinker simulated using all-atom MD (blue).

capping has been suggested to be a determinant of the binding affinity of zinc finger to DNA, and was identified as a conformational characteristic relevant to its bound state. On the contrary, backbone dihedrals of Gly2 in an atomistic MD simulation of cLinker itself showed a much wider, standard Ramachandran distribution for Glycine under the same simulation conditions. Therefore, the MD simulation for 1tf3_1_3 was too short to visit the entire phase space.

To overcome this insufficient sampling problem, an atomistic pivoting algorithm was devised (Section 4.2.2). It is shown in Figure 4.8 (b) that when this algorithm was applied to 1tf3_1_3, Gly39 could take all possible local conformations. The R_e distributions of multi-zinc finger proteins calculated by the pivoting algorithm are shown in Figure 4.9. For ctcf_1_11 and ctcf_2_10 (a,b), the distributions are skewed to the right. For other shorter polypeptides (c∼e), the distributions are skewed to the left. We use a worm-like chain model to explain the dependence of the R_e distribution shape on the number of residues of polypeptides [47]. Considering each zinc finger as a monomer, they are connected and form a polymer chain. The excluded volume of the zinc fingers

prevents the polypeptide bending too much, which *effectively* exerts a bending constraint on the chain. The strength of this constraint can be described by a persistence length ξ_p, where the orientational correlation between any two segments on the chain, with contour separation l, roughly decays as $\exp(-l/\xi_p)$. Then the shape of R_e distribution is dependent on the ratio of the contour length of the chain L to ξ_p. When L is small ($< 10\xi_p$), the excluded volume encourages an extended conformation. Otherwise, the chain is more likely to take a somewhat bended conformation.

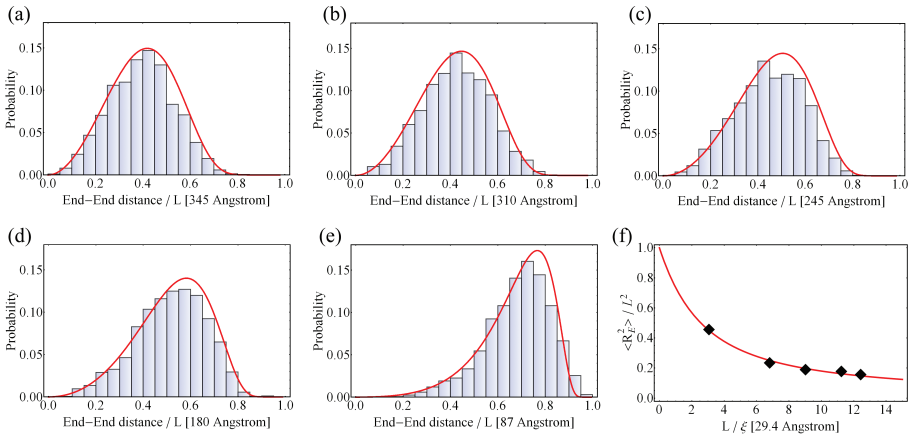


Figure 4.9: End-to-end distance (R_e) distributions of multi-zinc finger proteins. Scaled R_e distribution of (a) ctcf_1_11 (b) ctcf_2_10 (c) ctcf_3_9 (d) ctcf_4_8 (e) 1tf3_1_3 simulated by the atomistic pivoting algorithm, compared with that of a worm-like chain (red line). (f) Mean squared end-to-end distance $< R_e^2 >$ over the contour length L, fitted with the Debye function (red line).

For each mC_2H_2 polypeptide, here we defined L as the largest R_e that it can take, while still keeping the $\beta - \beta - \alpha$ motif of all zinc fingers. By fitting the relation betwen the mean squared end-to-end distance $< R_e^2 >$ and L

$$< R_e^2 > /L^2 = f_D(\frac{L}{\xi_p}), \tag{4.3}$$

where f_D is the Debye function $f_D(x) = 2(x-1+e^{-x})/x^2$, we obtained $\xi_p \approx 29.4$ Å (Figure 4.9 (f)). The R_e distribution of a worm-like chain was given by [96]

$$p(r) = \frac{\kappa}{N} \sum_{k=1}^{\infty} \pi^2 k^2 (-1)^{k+1} e^{-\kappa \pi^2 k^2 (1-r)}, \tag{4.4}$$

with $\kappa = \xi_p / L$ and a normalization factor N. These distributions are plotted using red lines in Figure 4.9 (a∼e), with k truncated at 48.

The *effective* bending constraint produces orientational correlation between adjacent zinc fingers, which may influence the transcription factor-DNA binding, e.g. part of the free energy barrier that needs to be overcome from unbound to bound state, is contributed by the protein itself (see Section 4.3.3). While mutation of the linker sequence affects the protein-DNA binding dynamics and conformations, the sequence and length of the linker will also impact the conformational properties of unbound multi-zinc finger proteins by changing ξ_p. For example, replacing Glycine by Proline will increase ξ_p and extending the linker will decrease ξ_p, resulting in a more straightened or more compact polypeptide, respectively.

4.3.3 The mesoscale model

Qualitative agreement between the RMSF of 1aay_1 from all-atom MD simulations and that from mesoscale simulations is shown in Figure 4.10 (a). The fluctuation of AAs decreases in the first β-strand and the α-helix, and it increases dramatically at both terminals. Note that the mesoscale potential does not stabilize the second β-strand as well as the atomic potential does, which is probably due to the coarse graining. To complement the analysis, we calculated and plotted the nativeness order parameter Q of 1aay_1 over time in Figure 4.10 (b). An average Q of ∼0.93 demonstrates that the rigidity of the single zinc finger is captured by our mesoscale model, where $Q \geq 0.6$ has been chosen as a

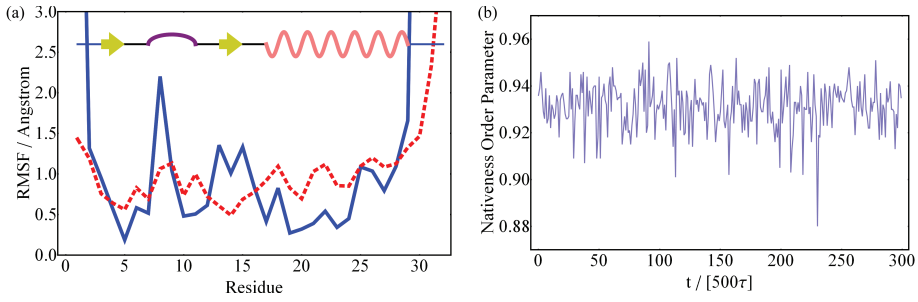


Figure 4.10: Stability of single zinc finger in the mesoscale model. (a) Comparison of RMSF of 1aay_1 simulated with all-atom MD (dashed red line) and with mesoscale MD (solid blue line). (b) Nativeness order parameter versus time in a typical mesoscale simulation.

threshold for peptide successful folding [87].

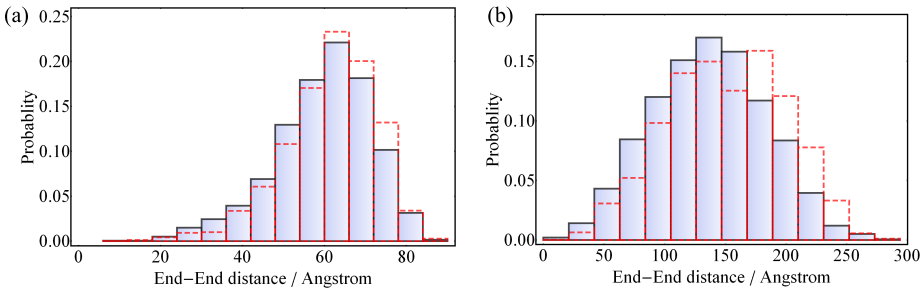


Figure 4.11: R_e distribution of (a) 1tf3_1_3 and (b) ctcf_1_11, simulated with the atomic pivoting algorithm (solid blue), and with mesoscale MD (dashed red).

Compared to the results from the atomistic pivoting algorithm, R_e distributions of 1tf3_1_3 and ctcf_1_11, calculated with mesoscale MD, are plotted in Figure 4.11 (a) and (b) respectively. It shows that while the excluded volume interaction dominates the global packing of the polypeptide chain, the mesoscale model favors more extended conformations, as a result of other interactions, namely the hydrogen bonds and dipole-dipole interaction.

In addition, we calculated the free energy difference landscape of 1tf3_1_3

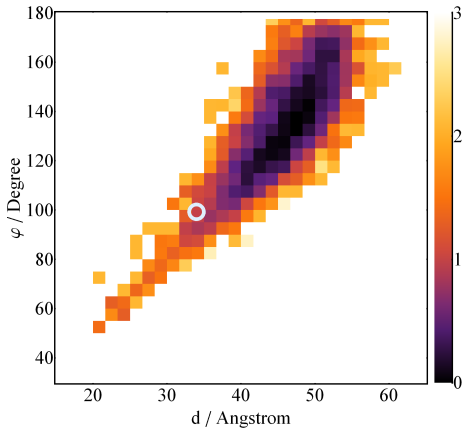


Figure 4.12: Free energy landscape of 1tf3_1_3. d is the distance between the center of geometry of ZF1 and that of ZF3. φ is the angle between the centers of geometry of three zinc fingers. The color represents the free energy difference relative to the lowest value, in units of $k_B T_r$. The DNA-bound state is the center of the white circle.

using WHAM. It is projected onto a plane in Figure 4.12 with two axes corresponding to $d = |\boldsymbol{r}^3_{cog} - \boldsymbol{r}^1_{cog}|$ and $\varphi = \angle \boldsymbol{r}^1_{cog} \boldsymbol{r}^2_{cog} \boldsymbol{r}^3_{cog}$, where $\boldsymbol{r}^i_{cog}$ is the center of geometry of the i-th zinc finger. It shows several characteristics of the conformations of 1tf3_1_3: i) too small d (< 15 Å) and too small φ ($< 40°$) values are forbidden due to the excluded volume of the zinc finger; ii) small d with large φ and large d with small φ are forbidden due to the length constraint of linkers; iii) it corresponds well with the R_e distribution shape, i.e., weighted on the larger d value side; iv) compared to the DNA-bound state (34 Å, 99°), most favorable unbound conformations of d 40 $\sim$ 50 Å and φ 120 $\sim$ 150° are more elongated. The free energy penalty from most populated unbound states to the bound state is around $k_B T_r$.

The mesoscale MD is more efficient than the all-atom MD simulation. To sample sufficient conformations to calculate a R_e distribution like in Figure 4.11 (a), it takes 104 CPU hours using the atomistic pivoting algorithm, 324 CPU hours using mesoscale MD, whereas 32000 CPU hours of all-atom MD

simulation is still not long enough.

4.4 Conclusion

In this chapter, we studied the conformational properties of unbound mC_2H_2 zinc finger proteins using multiscale approaches. First, a homology model of the tandem zinc finger domain of transcription factor CTCF was constructed. All-atom MD simulations showed that single zinc finger is a stable structural unit, independent of the studied environmental conditions. In agreement with the NMR observation of the N-terminal three zinc fingers of TFIIIA, the polypeptide becomes more extended from a DNA-bound state to an unbound state. Next, an atomistic pivoting algorithm, which considers only the excluded volume interaction, was developed to investigate the global conformational characteristics. It showed that as the number of zinc fingers increases, the end-to-end distance distribution gradually changes its shape, from skewed to the left to skewed to the right. This was explained using a worm-like chain model. The *effective* bending constraint should apply not only to multi-zinc finger proteins, but also to other multi-domain proteins connected by short flexible linkers. Finally, a mesoscale peptide model was modified for mC_2H_2 proteins, which is efficient while providing similar conformational properties as those given by atomistic models.

Due to the limitation on computational resources, our all-atom MD simulation was not long enough to calculate a R_e distribution of 1tf3_1_3, which could be further compared to that calculated with the mesoscale model. Based on the modified mesoscale model, how mC_2H_2 protein binds to double stranded DNA or how it searches for its DNA target loci will be studied in the next chapter.

5 Multi-Cys_2His_2 zinc finger protein in complex with DNA

The interaction of DNA with multi-Cys_2His_2 zinc finger proteins, Lei Liu and Dieter W Heermann, Journal of Physics: Condensed Matter. Accepted.

Chapter Summary. The multi-Cys_2His_2 (mC_2H_2) zinc finger protein, like CTCF, plays a central role in the three-dimensional organization of chromatin and gene regulation. The interaction between DNA and mC_2H_2 zinc finger proteins becomes crucial to better understand how CTCF dynamically shapes the chromatin structure. Here we study a coarse-grained model of the mC_2H_2 zinc finger proteins in complexes with DNA, in particularly, study how a mC_2H_2 zinc finger protein binds to and searches for its target DNA loci. On the basis of coarse-grained molecular dynamics simulations we present several interesting kinetic conformational properties of the proteins, such as the rotation coupled sliding, the asymmetrical roles of different zinc fingers and the partial binding partial dangling mode. In addition, two kinds of studied mC_2H_2 zinc finger proteins, of CG-rich and AT-rich binding motif each, were able to recognize their target sites and slid away from their non-target sites, which shows a proper sequence specificity in our model and the derived force field for mC_2H_2-DNA interaction. A further application to CTCF shows that the protein binds to a specific DNA duplex only with its central zinc fingers. The zinc finger domains of CTCF asymmetrically bend the DNA, but do not form a DNA loop alone in our simulations.

5.1 Introduction

In human cells, the chromosomes three-dimensional structure is dynamically shaped by certain proteins and the nuclear lamina *in vivo*. More and more recent chromosome conformation capture results suggest that the protein CCCTC-binding factor (CTCF), as well as cohesin, might be responsible for the spatial organization of chromatin in mammalian nuclei [97, 98, 2]. They mediate the long-range chromatin looping genome wide and their binding sites are enriched at the boundaries of both topological domains and "subdomains". Since first isolated from chicken in 1990, CTCF has been reported to play many different roles in gene regulation in different contexts [99, 100, 101]. Several possible DNA looping mechanisms of CTCF, where it regulates gene expression via reshaping the chromatin structure, have been proposed [101, 4, 102]. Although CTCF binds to a lot of DNA loci, different CTCF consensus binding motifs have been also reported by different groups [40, 21, 38]. However, little is known about how CTCF binds to these different DNA loci and how it searches for its target loci.

The central binding domain of CTCF is composed of ten Cys_2His_2 (C_2H_2) zinc fingers and one C_2HC zinc finger, hence it is classified as a C_2H_2 zinc finger protein. Since the transcription factor TFIIIA was first identified in *Xenopus laevis*, the C_2H_2 zinc finger protein has attracted wide range of interest for several decades [72, 103, 104]. Each C_2H_2 finger contains one central Zn^{2+} ion coordinated by two cysteines and two histidines, with a sequence motif X_2-C-$X_{2\text{-}4}$-C-X_{12}-H-$X_{3\text{-}5}$-H, and folds into a stable β-β-α secondary structure domain. Multi-C_2H_2 zinc finger proteins (mC_2H_2), usually tandem repeats of C_2H_2 zinc fingers connected by highly conserved short peptide linkers, are ubiquitous in eukaryotic cells and affect a broad range of biological functions.

The most fascinating character of C_2H_2 zinc finger appears to be its modularity for sequence-specific binding to DNA [103, 105]. As a representative, Egr-1 (also known as Zif268 [106, 107]) which contains three zinc fingers and has a great influence in the brain and cardiovascular system, binds specifically to a 9 base-pair (bp) CG-rich sequence, while each zinc finger contacts a 3∼4 bp subsite along the major groove of DNA. This "canonical" zinc finger-DNA binding mode provided an expectation of a DNA recognition code and hence a framework for the design of novel zinc finger combinations recognizing desired DNA target sites. Considerable effort has been devoted into this field and it turned out that a single versatile DNA recognition code for mC_2H_2 does not exist. First, there are other types of zinc fingers which uniquely bind to DNA sites of AT-rich sequences (e.g., $TATA_{ZF}$ [108]). Second, the binding affinity of a mC_2H_2 is neither a simple summation nor multiplication of the binding affinities of its component zinc fingers. It is also determined by the cooperativity between zinc fingers, as well as the linkers [103, 43]. Different zinc fingers in mC_2H_2 might play different roles in a complex with DNA, such as the N-terminal six zinc fingers of protein TFIIIA [109]. Prediction of the binding motif and conformation of the generic mC_2H_2 zinc finger protein with DNA remains an unsolved challenge [110, 111].

The kinetics of mC_2H_2 zinc finger protein, which is usually intrinsically disordered when unbound [73, 42, 112], searching for its target DNA loci is also a quite interesting problem. The asymmetrical roles of zinc fingers of Egr-1 in DNA-scanning process was revealed by a recent NMR study [89]. The first zinc finger (ZF1) of Egr-1 undergoes more intensive domain motions than the second and third one (ZF2 and ZF3 respectively) when Egr-1 slides along DNA, and ZF1 mainly dissociates while ZF2 and ZF3 bind to the DNA in the nonspecific DNA complex. This is consistent with the "search and fold" mechanism (also

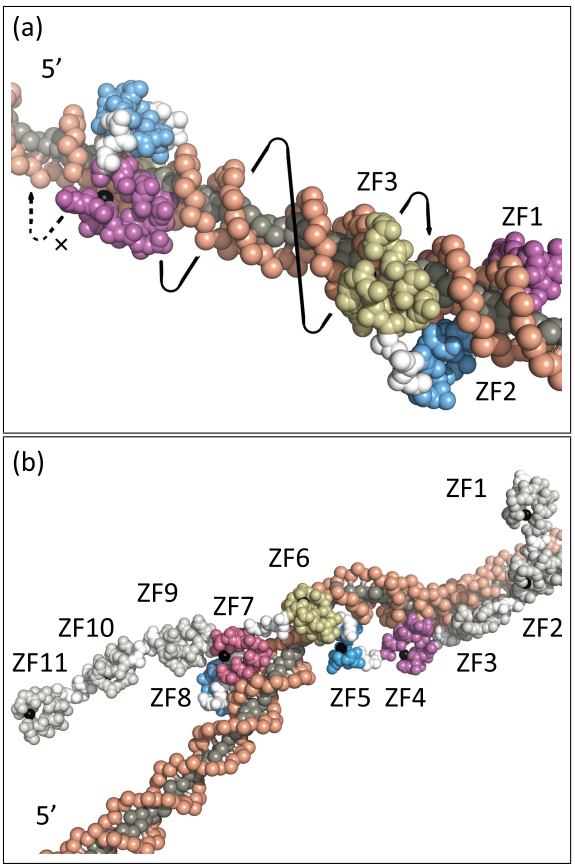


Figure 5.1: Multi-Cys_2His_2 zinc finger proteins in complexes with DNA. (a) Rotation coupled sliding motion of protein Egr-1 when it slides along DNA. Egr-1 binding domain contains three zinc fingers (labeled as ZF1, ZF2 and ZF3 from N to C terminus), which are joined by two flexible white linkers. (b) A snapshot of the simulated eleven zinc finger domains of CTCF bound to DNA. The central five zinc fingers ZF4∼ZF8, which contact DNA, are highlighted in different colors.

called two-mode model) which emphasizes the coupling between protein binding and partial protein folding [42, 74]. Coarse-grained (CG) molecular dynamics simulation, which only considered the electrostatic interaction between protein and DNA, succeeded in orientating Egr-1 and showing a higher domain mobility of ZF1 [89]. However, the effect of mutation on the cognate DNA sequence is

out of the scope of this sequence nonspecific model.

Aimed at understanding how CTCF changes the chromatin structure in more detail, in this paper we studied the conformational properties of mC_2H_2 zinc finger proteins while it binds to or scans double stranded DNA, with sequence specific protein-DNA interactions. In section 5.2, the mapping schemes from atomistic to CG scale for both amino acids and nucleotides are described. In section 5.3, we will briefly explain how the protein-DNA CG force-field was derived, the simulation details and all the conformational properties of interest. Then simulation results of the force-field parameterization are presented for one training case, two testing cases and a first application to the zinc finger domains of CTCF. Finally, we give a short discussion and summary of our work.

5.2 Coarse-grained model

Due to the intrinsically disorder of mC_2H_2, CG peptide models which forbid changes of the secondary or tertiary structure of the protein during simulation are not suitable for our purpose. We have adjusted the *peptideB* model, which was designed by Bereau [87] for studying protein folding and aggregation, as the CG description of mC_2H_2 zinc finger proteins (figure 5.2(a)). Each amino acid is modeled by three (Glycine) or four (non-Glycine) beads, that represent the amide group N, central carbon C_α, carbonyl group C' and side chain group C_β. The first three beads are backbone beads and the fourth bead determines the sequence specificity of peptides. Considering the importance of the Zn^{2+} ion in maintaining the secondary structure of single zinc finger, the Zn^{2+} ion is explicitly represented by an extra bead. G$\bar{o}$-type constraints between the Zn^{2+}

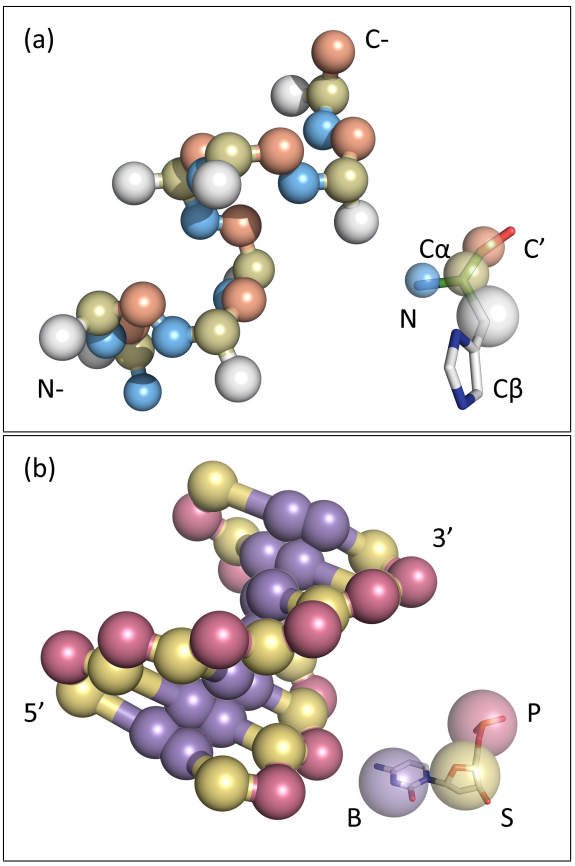


Figure 5.2: Coarse-grained (CG) models for proteins (a) and DNA (b). Each amino acid in proteins are represented by four beads (amide group N, central carbon C_α, carbonyl group C' and side chain group C_β). Each nucleotide in DNA are represented by three beads (phosphate group P, sugar group S and base group B).

and the C_β of its coordinating residues are added as

$$V_{Zn-bond}(r) = \frac{1}{2}\kappa_{Zb}(r - r_0)^2 \tag{5.1}$$

$$V_{Zn-angle}(r) = \frac{1}{2}\kappa_{Za}(\theta - \theta_0)^2, \tag{5.2}$$

where the equilibrium distance r_0 of Zn^{2+}-C_β and equilibrium angle θ_0 of C_β-Zn^{2+}-C_β are obtained from reference structures. By simulations of a single

ZF, the interaction strength κ_{Zb} and κ_{Za} have been tuned to give the best nativeness Q ($\sim$ 0.93), which has the same expression as equation 7 ($Q =< exp[-(r_{i,j}^{ref} - r_{i,j}^{sim})^2/9] >_{i,j}$), but averaged over any bead pairs $\{i, j\}$ in the ZF.

To simulate B-form double stranded DNA, we use the 3SPN.1 model [113, 114] which has been successfully applied to study the DNA sequence preference of nucleosomes (figure 5.2(b)). Each nucleotide is modeled by three beads, representing the backbone phosphate group P, the backbone sugar group S, and the base group B$\in$\{A,T,C,G\}.

Both *peptideB* and 3SPN.1 are implicit solvent models, while the effect of salt concentration in solvent is considered via the Debye-Hueckel approximation. For simplicity, we set the temperature to T=300 K and the salt concentration in sodium chloride solvent to [Na^+]=150 mM in this work.

5.3 Methods

5.3.1 Parametrization

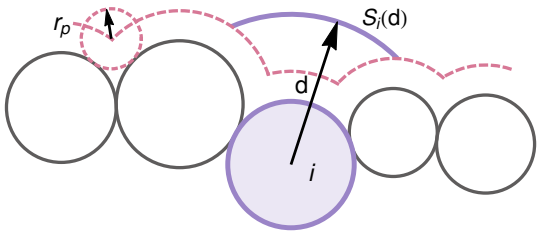


Figure 5.3: The accessible surface area $S_i(d)$ of radius d for a bead of type i. r_p is the radius of a probe, and the dash line represents the probe's accessible surface.

Based on our CG model for mC_2H_2-DNA complexes, the protein-DNA interactions are mainly composed of the nonbonded interactions between the twenty different types of side chain beads C_β of the protein and the six different types

of beads of DNA. We derived the 20×6 interactions of each pair of type i in the protein and type j in the DNA as follows. First, all the atomistic mC_2H_2-DNA complexes deposited in the RCSB Protein Data Bank [115], which was filtered to remove those of high amino acid sequence similarity, were mapped to our CG representation (see also appendix). Then an initial guess of the "statistical potentials" were inferred from these CG complexes using the formula [116, 117]

$$G_{i,j}(d) = -k_B T < ln(N_{obs}(i,j,d)/N_{exp}(i,j,d)) >, \tag{5.3}$$

where N_{obs} and N_{exp} are the number of observed and expected occurences of a bead pair $\{i,j\}$ of a distance $d \pm \delta$, and the term inside the angle brackets is averaged over all complexes. Given the coordinates of beads in CG complexes, it is straightforward to count N_{obs}. We calculated N_{exp} as

$$N_{exp}(i,j,d) = \chi_i(d)\chi_j(d)N_{tot}(d) \tag{5.4}$$

$$= \frac{S_i(d)}{\sum_i S_i(d)} \frac{S_j(d)}{\sum_j S_j(d)} N_{tot}(d). \tag{5.5}$$

$N_{tot}(d)$ is the total number of pairs of mC_2H_2-DNA beads of a distance $d \pm \delta$, and $\chi_i(d)$ is the probability of any DNA bead appearing at a distance $d \pm \delta$ from a protein bead of type i. $S_i(d)$ is the contribution from beads of type i to the total accessible surface area of radius d, which is defined as the area accessible to the center of a probe of radius r_p while the surface remains at a distance d from the bead i (figure 5.3). Hence, $\chi_i(d)$ and $\chi_j(d)$ can be calculated from the mC_2H_2 structure and the DNA structure in a CG complex respectively. The parameter d was varied from 5 to 18 Å with a step size of 0.5 Å, i.e., $\delta = 0.5$ Å. $N_{obs}(i,j,d)$ will be too small to give a meaningful and smooth profile of $G_{i,j}(d)$, if d is smaller than 5 Å or larger than 18 Å. The same consideration applies for the choice of δ. The radius of a probe r_p, is set to be equal to 3.44 Å, which

is the mean value of all bead radii. According to the shape of the subtracted potentials, for i representing C_β beads $G_{i,j}$ was then fitted to

$$U_{i,j}^{cb-DNA}(d) = \epsilon_{i,j}[(\frac{a(\sigma_{i,j}+18)}{d+18})^8 - (\frac{a(\sigma_{i,j}+18)}{d+18})^6] \tag{5.6}$$

with $a = (3/4)^{1/2}$, $\sigma_{i,j} \leq d \leq d_{cutoff}$. When $d < \sigma_{i,j}$,

$$U_{i,j}^{cb-DNA}(d) = \epsilon_{i,j}[(\frac{a\sigma_{i,j}}{d})^8 - (\frac{a\sigma_{i,j}}{d})^6]. \tag{5.7}$$

For other types of i (N, C_α, C'), a Weeks-Chandler-Andersen (WCA) pure repulsive potential was used as

$$U_{i,j}^{bb-DNA}(d) = 4\epsilon_{bb-DNA}[(\frac{\sigma_{i,j}}{d})^{12} - (\frac{\sigma_{i,j}}{d})^6 + 1/4] \tag{5.8}$$

when $d \geq 2^{1/6}\sigma_{i,j}$. Otherwise, $U_{i,j}^{bb-DNA}(d) = 0$. Note that the superscript *cb* means C_β beads, and *bb* stands for backbone beads in the above equations. The interaction between Zn^{2+} bead and DNA was ignored due to small value of N_{obs}.

$U_{i,j}^{cb-DNA}$ was further divided into two parts, the DNA-sequence dependent interaction $U_{i,j}^{cb-\{A,T,C,G\}}$ and the DNA-sequence independent interaction $U_{i,j}^{cb-\{S,P\}}$, which includes 20×4 and 20×2 pairs of $\{\epsilon_{i,j}, \sigma_{i,j}\}$ respectively. Since the fluctuation of the binding energy landscape along the DNA affects the stability of the protein-DNA complex at the target site [118, 119], we multiplied $U_{i,j}^{cb-\{A,T,C,G\}}$ by a prefactor f^{SP}. In addition, because the mean value of binding energy determines the dissociation rate of a protein from DNA, we added a shift m^{SP} to $U_{i,j}^{cb-\{A,T,C,G\}}$. $U_{i,j}^{cb-\{S,P\}}$ and $U_{i,j}^{bb-DNA}$ was manipulated by $\{f^{NS}, m^{NS}\}$ and $\{f^{WCA}, m^{WCA}\}$ likewise. Taking the interactions between C_α beads and DNA base groups as an example, the final potential that we used

during the simulation has a form,

$$V_{i,j}^{cb-\{A,T,C,G\}}(d) = f^{SP} \times U_{i,j}^{cb-\{A,T,C,G\}}(d) + m^{SP}. \tag{5.9}$$

The shift parameter does not change the depth of a potential well.

The parameters $\{\epsilon_{i,j}, \sigma_{i,j}\}$ are obtained by fitting the potential function $U_{i,j}(d)$ to the statistical potential $G_{i,j}(d)$. Rather than further tuning $\{\epsilon_{i,j}, \sigma_{i,j}\}$ we introduced six scaling or shifting parameters $\{f^{SP}, m^{SP}, f^{NS}, m^{NS}, f^{WCA}, m^{WCA}\}$. For these we found the optimum values by force matching, which facilitate the binding specificity of our training case. More precisely, their values are tuned to stabilize Egr-1 in the complex with DNA for the training case, but to destabilize the complex with nonspecific site. All the parameter values used in the molecular dynamics simulation are given in Table 5.1 and Table 5.2.

Table 5.1: PDB complexes used to derive mC_2H_2-DNA potential.

3UK3	4F6M	2WBS	2WBU	2KMK	2PRT
2I13	1LLM	1P47	1F2I	1G2F	1TF6
1YUI	1AAY	1MEY	1UBD	2DRP	2GLI

5.3.2 Characterization and simulation details

To quantitatively characterize a CG mC_2H_2-DNA complex structure, we defined and calculated the following features [120, 121].

(a) A protein-DNA contact interface is composed of all the mC_2H_2-DNA bead pairs within a threshold distance of 7 Å. Then the interface nativeness Q is defined as

$$Q =< exp[-(r_{i,j}^{ref} - r_{i,j}^{sim})^2/9] >_{i,j}, \tag{5.10}$$

where $r_{i,j}$ is the distance between bead pair $\{i, j\}$ in the reference or simulated conformation, and the average is taken over all the bead pairs $\{i, j\}$ which belong

Table 5.2: $\{\epsilon_{i,j}, \sigma_{i,j}\}$ in mC_2H_2-DNA potential.

mC_2H_2	DNA (kcal/mol, Å)					
	P	S	A	T	C	G
N	0.09, 4.9	0.09, 4.9	0.09, 4.9	0.09, 4.9	0.09, 4.9	0.09, 4.9
C_α	0.09, 5.3	0.09, 5.3	0.09, 5.3	0.09, 5.3	0.09, 5.3	0.09, 5.3
C'	0.09, 5.2	0.09, 5.2	0.09, 5.2	0.09, 5.2	0.09, 5.2	0.09, 5.2
Ala	0.12, 4.8	0.12, 4.2	1.51, 6.4	2.41, 6.0	2.44, 6.1	2.47, 8.0
Pro	0.11, 4.5	0.09, 5.2	0.65, 6.1	1.87, 7.6	1.92, 4.3	1.28, 7.0
Glu	0.17, 4.8	0.18, 5.5	1.25, 8.1	1.56, 4.5	3.26, 6.5	3.54, 8.0
Gln	0.13, 4.8	0.11, 6.1	3.68, 6.5	2.90, 6.4	2.28, 6.3	2.49, 7.4
Asp	0.11, 6.5	0.13, 5.1	2.31, 5.9	2.77, 6.6	5.95, 5.7	7.28, 7.1
Asn	0.10, 5.6	0.10, 6.1	3.38, 5.1	4.40, 7.8	2.96, 7.8	1.81, 7.4
Ser	0.14, 3.9	0.16, 5.6	2.76, 4.8	3.31, 5.7	3.81, 6.4	2.31, 5.5
His	0.28, 3.9	0.22, 6.5	2.95, 6.5	3.51, 4.6	4.03, 7.3	4.50, 7.5
Lys	0.14, 3.9	0.12, 4.8	1.27, 7.8	3.73, 7.2	4.14, 7.2	2.19, 6.5
Arg	0.14, 5.2	0.10, 5.3	3.77, 7.8	5.00, 7.3	7.03, 7.7	6.49, 8.2
Thr	0.18, 3.9	0.16, 5.2	2.06, 6.4	3.29, 6.1	3.40, 6.6	2.39, 7.6
Val	0.12, 5.0	0.11, 5.0	1.79, 6.7	3.09,10.3	2.27, 8.3	1.44, 9.3
Ile	0.19, 4.3	0.15, 5.6	1.31, 9.2	2.71, 8.0	2.28, 6.8	4.39, 9.8
Leu	0.19, 6.0	0.11, 5.9	1.29, 8.0	2.89, 8.0	2.04, 6.5	0.28,10.5
Met	0.17, 7.7	0.12, 7.7	0.51, 4.8	2.25,10.3	0.82, 8.0	1.46,10.1
Phe	0.28, 5.6	0.18, 6.1	1.25, 8.0	3.60, 8.8	2.52, 6.6	0.76, 8.5
Tyr	0.11, 5.2	0.11, 4.5	0.30, 6.4	1.48, 6.8	1.32, 5.6	0.63, 4.4
Cys	0.22, 6.1	0.19, 5.6	1.04, 8.2	2.30, 7.2	2.22, 8.2	0.86, 8.7
Trp	0.11, 5.6	0.08, 6.9	0.57, 9.3	2.74, 7.8	2.51, 7.8	1.01, 8.7

to the reference contact interface. The recognition region of a protein is made up of all the protein beads on the contact interface.

(b) For any protein bead of a snapshot taken from a simulation, first in each DNA strand we determine the phosphate bead which is closest to this protein bead, then we calculate the distance between these two phosphate beads. If the separation is larger than 15 Å, this protein bead is defined to locate in the major groove of DNA. Otherwise, it is in the minor groove. A small change of the criterion 15 Å, e.g., by 1 Å, will not change the conclusions obtained from the simulation results. The percentage of the recognition region of a protein in the major groove is obtained by applying this judgment to all the protein beads in its recognition region.

(c) Since DNA is dynamic in the simulation, local axis $\{\boldsymbol{l}_x, \boldsymbol{l}_y, \boldsymbol{l}_z\}$ are defined bp by bp along the DNA. For the i-th bp,

$$\boldsymbol{l}_x^i \parallel \boldsymbol{r}_i^S - \boldsymbol{r}_i^{cen} \tag{5.11}$$

$$\boldsymbol{l}_y^i \parallel \boldsymbol{r}_{i+1}^{cen} - \boldsymbol{r}_i^{cen} \tag{5.12}$$

$$\boldsymbol{l}_z^i \parallel \boldsymbol{l}_x^i \times \boldsymbol{l}_y^i \tag{5.13}$$

where $\boldsymbol{r}_i^S$ is the position of the i-th sugar bead on a single DNA strand, and $\boldsymbol{r}_i^{cen}$ is the position of the center of the i-th pair of sugar beads. The orientation of a protein domain relative to DNA is determined as the angle between the center of geometry of its recognition region $\boldsymbol{r}_{reg}^{cog}$, the center of geometry of the domain $\boldsymbol{r}_{dom}^{cog}$, and the projection point of $\boldsymbol{r}_{dom}^{cog}$ to $\boldsymbol{l}_y^i$, where i minimize $|\boldsymbol{r}_{dom}^{cog} - \boldsymbol{r}_i^{cen}|$. Although there still exist some controversies, we still hope to find a quantity to characterizes the bending of double stranded DNA observed in our simulations. With the 3SPN model, de Pablo has studied the dependence of the DNA persistence length on the environment's salt concentration [113, 114], and Takada has studied the effect of P53 binding on DNA bending [121]. A similar DNA bending score is defined in our work as $< \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_{i+10} >$, where $\hat{\boldsymbol{u}}_i$ is a unit vector parallel to the vector $\boldsymbol{r}_i^{Cen} - \boldsymbol{r}_{i-10}^{Cen}$, and $\boldsymbol{r}_i^{Cen}$ is the position of the center of the i-th pair of sugar beads. With this definition, for a DNA segment of n bp, the bending score is calculated from the tenth bp to $(n-10)$-th bp. It decreases as the DNA bends more severely.

We studied four mC_2H_2 zinc finger proteins in complex with DNA of different sequences, which are summarized in table 5.3. Egr-1 are simulated with a DNA duplex of 28 bp, of either specific (SP) sequence which contains the target site or nonspecific (NS) sequence without it. The sequence of SP28 and NS28 were taken from the recent experimental research of Iwahara [89]. The initial conformation of Egr-1:SP28 (or NS28) complex for the molecular dynam-

Table 5.3: Simulated mC_2H_2-DNA complexes.

mC_2H_2	DNA			
	Label	Sequence[a,b]		
Egr-1	SP28	GTACCGATT	GCGTGGGCG	GAACCTTCAG
	NS28	GTACCGATT	GCAGATTCC	GAACCTTCAG
$TATA_{ZF}$	SP27	GCCCCGGAC	GCTATAAAA	GGAGGGGCC
	NS27	GCCCCGGAC	CACCATCCG	GGAGGGGCC
TFIIIA	SP61	TGATCTCAG	AAGCGATAC	AGGGTCGGG
		CCTGGTTAG	TACCTGGAT	GGGAGACCG
		CCTGGGA		
CTCF	SP160	CGGCTTATG	TGATCTCTC	GATCGAATT
		AGTTTACTT	TGCCTGCAC	CCCCAGCAG
		CGCTGCAGT	ACCGCGCTT	GGCCGCGAG
		GTGGCGCCA	TTGCTCCAC	GATTGACGC
		GCGCCCCCC	GCGTTTAAC	GTATAAGGG
		ACGCCTAGC	CGGCTTTCA	ACAGGCA

[a] From 5' to 3'.

[b] Target sites in SP or non-target sites in NS have been underlined.

ics simulation was built by superimposing the sugar group beads of G10∼G18 in a standard B-form structure of SP28 (or NS28) to the sugar group beads of the target site in the reference structure (PDB code: 1AAY [107]), while the protein remained the same as in 1AAY. A similar preparation was made for $TATA_{ZF}$ and TFIIIA, with reference structures of PDB codes 1G2F and 1TF6, respectively [108, 109].

As for CTCF, the atomistic structures of the sixth and seventh, tenth and eleventh zinc fingers were contained in the PDB structure 2CT1 and 1X6H respectively. Other zinc fingers were prepared using a homology modeling approach with the MODELLER 9.10 program [83]. Then the constructed atomistic structure of the central eleven zinc finger domains of CTCF was mapped to the CG scale and simulated without DNA. Next, we randomly chose an unbound CTCF conformation, translated and rotated it so that the seventh zinc finger was embedded into the major groove of the DNA sites C84∼G86. We use this

transformed CTCF, together with the 160 bp specific DNA duplex [21], as the initial conformation of CTCF:SP160 complex.

Our molecular dynamics simulations were carried out using the ESPResSo 3.1.0 package [66, 67]. In case that beads clash with each other in the initial conformation, a complex structure was energy minimized, and "warmed" up by gradually increasing the allowable maximum force strength from zero to the normal value. Given the CG time scale $t \sim 0.1$ ps, a subsequent simulation at constant volume and constant temperature (NVT) was performed using a Langevin thermostat with friction constant of t^{-1}, an integration step of $0.01t$, a sample step of $\tau \sim 500t$ and total simulation time of $1.5 \times 10^5 t$. Then we analyzed the conformational properties of this complex based on the trajectories of tens of simulations performed with different random seeds.

5.4 Results

5.4.1 Egr-1

Rotation coupled sliding.

Whether a protein slides along the DNA helical pitch (rotation coupled sliding) or not is governed by the type and details of the dominating interaction between the protein and DNA, and may affect the rate of the protein to find its target site. Given that Egr-1 is embedded in the major groove of DNA in the crystal structure, we first examine the percentage of the recognition region of Egr-1 located in the major groove (figure 5.4(a)). For both specific and nonspecific DNA duplexes, more than 80% of the recognition region of Egr-1 stay in the major groove during the simulations. In addition, we calculated the trajectory averaged distance from the C_α bead to its nearest DNA bead for each amino acid residue of the protein in figure 5.4(b). For each zinc finger (bounded via

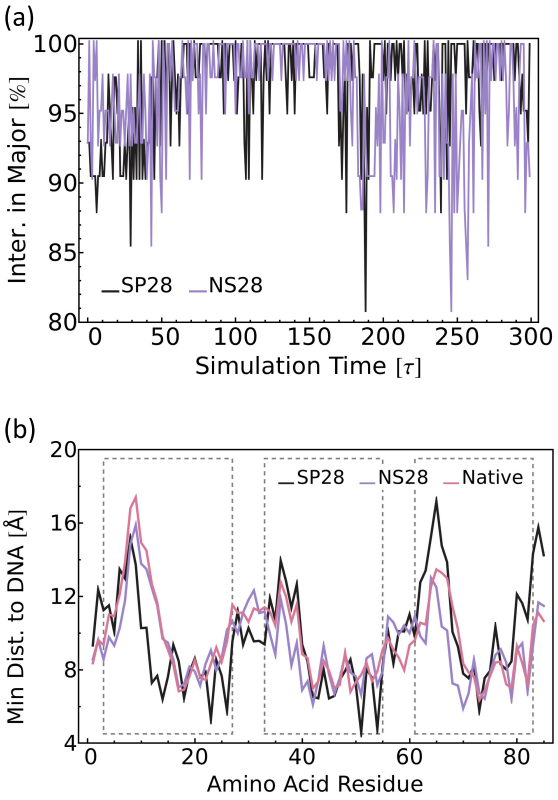


Figure 5.4: Conformational characteristics of Egr-1 sliding along specific (SP28) and nonspecific (NS28) DNA. (a) Percentage of the recognition region of Egr-1 located in the major groove of DNA during simulation. (b) Trajectory averaged minimal distance from the C_α bead to DNA for each amino acid residue, compared with the distance in the crystallographic Egr-1:DNA complex (Native). Residues in three zinc fingers are bounded via three dashed boxes.

a dashed box), the distance grows to a high peak, then decreases and oscillates on a low plateau. A similar distance profile appears in the crystal structure of Egr-1:DNA complex, because the C_2H_2 zinc finger uses its C-terminal α-helix as its DNA-binding interface. Together this suggests that the protein slides along DNA with the same contact interface as in the crystal complex, no matter the DNA site is specific or nonspecific.

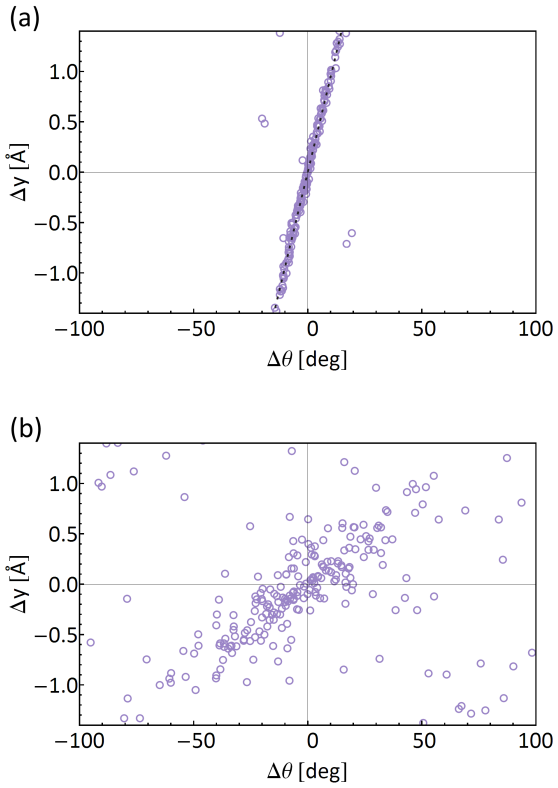


Figure 5.5: Distribution of the changes in the rotation angle θ ($\Delta\theta$) and in the diffusion displacement y (Δy) for (a) the second zinc finger of Egr-1 and (b) Egr-1.

To further clarify the rotation-diffusion coupled motion, given the definition the rotation angle θ based on the local axes $\{\boldsymbol{l}_x, \boldsymbol{l}_z\}$ and the diffusion displacement y based on $\boldsymbol{l}_y$ (see section 5.3.2(c)), we plot the distribution of the changes in rotation value $\Delta\theta$ and changes in diffusion displacement Δy between successive samplings in figure 5.5, for the second zinc finger of Egr-1 (a) and Egr-1 (b). The dashed line in figure 5.5(a) corresponds to a diffusion along the helical groove of an ideal B-form DNA, 34 Å displacement with 360° rotation, which matches quite well to the simulated data. As for the whole protein, θ and y are still coupled with a Pearson correlation coefficient of 0.874, but the slope of $\Delta\theta$

over Δy has changed. The latter point can be interpreted as a result of the fact that different zinc fingers have different characteristics of motion, which will be discussed in the following.

Asymmetrical roles of zinc fingers.

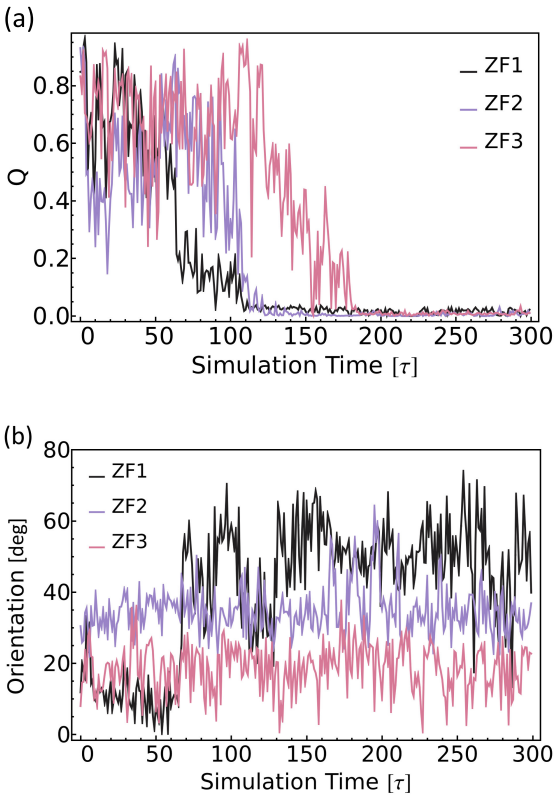


Figure 5.6: Conformational properties of different zinc fingers during the simulation of Egr-1 in complex with nonspecific DNA. (a) The interface nativeness Q. The smaller the Q is, the farther the zinc finger deviates from its initial gesture relative to DNA. (b) The orientation angle relative to DNA. A large orientation angle indicates a dissociation of the zinc finger from DNA.

To show the asymmetrical roles of zinc fingers in Egr-1, namely ZF1 almost

dissociates when Egr-1 diffuses on nonspecific DNA, we calculated and compared (a) the interface nativeness Q and (b) the orientation relative to DNA versus the simulation time for different zinc fingers. In figure 5.6(a) Qs of three zinc fingers drop to zero one by one. This shows that the first zinc finger initiated the sliding, followed by the second then the third zinc finger. In figure 5.6(b), while the orientation angles of the second and third zinc fingers remain at low values, the orientation angle of the first one is comparatively large, which suggests the dissociation of the first zinc finger. A direct influence of the more intensive motion of ZF1 is that during our simulations of the Egr-1:NS28 complex, the protein always slides towards the 3' end of the DNA duplex (see also figure 5.1(a)), which has also been experimentally verified.

DNA sequence dependency.

We examined the sequence specificity of the "statistical potentials" via a comparison of the interface nativeness of the mC_2H_2 zinc finger protein in complexes with specific and nonspecific DNA. In figure 5.7(a), it is clear that Egr-1 resides on its target site and forms a stable complex with the specific DNA duplex (SP28), but it slides away from its non-target site when it is simulated with the nonspecific DNA duplex (NS28). The probability of Egr-1 sliding away for SP28 and NS28, calculated from 100 independent simulations, are 0.17 and 0.74 respectively (see table 5.4).

5.4.2 $\mathbf{TATA_{ZF}}$

To check whether the potential of protein-DNA interaction is biased to the "canonical" mC_2H_2-DNA complex or not, we further studied $TATA_{ZF}$, which binds specifically to 9 bp AT-rich DNA motif, also in complexes with specific DNA (SP27) and nonspecific DNA (NS27). The interface nativeness during a

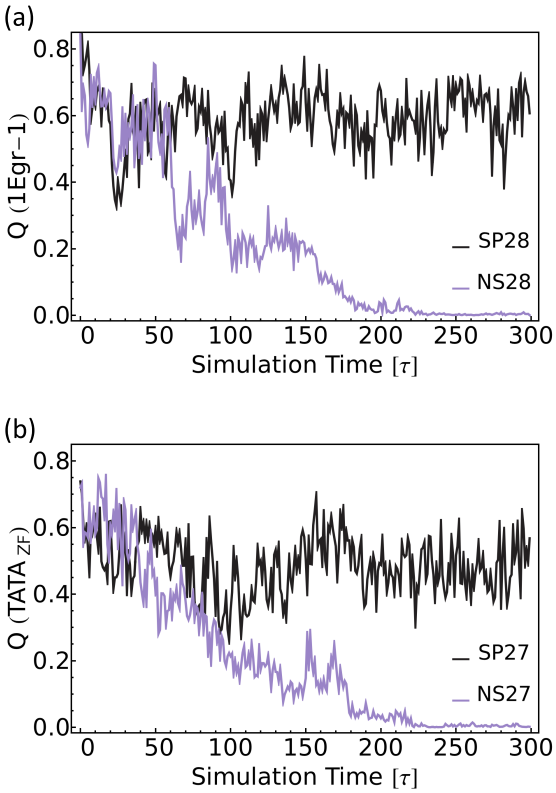


Figure 5.7: Interface nativeness Q, which is the average of Qs of three component zinc fingers versus simulation time for (a) Egr-1 and (b) $\mathrm{TATA_{ZF}}$ in complex with specific or nonspecific DNA duplex.

typical simulation is shown in figure 5.7(b). Similar to the results for Egr-1, $\mathrm{TATA_{ZF}}$ slides away on non-target site while stays on target site. The probability of this protein sliding away with SP27 and NS27 are 0.09 and 0.67. These values suggest that our CG potential captures the essential sequence specificity for $\mathrm{mC_2H_2}$-DNA recognition.

Table 5.4: DNA sequence specificity.

mC_2H_2	DNA	Probability[a]
Egr-1	SP28	0.17
	NS28	0.74
$TATA_{ZF}$	SP27	0.09
	NS27	0.67

[a] Probability of mC_2H_2 to slide away in the simulations.

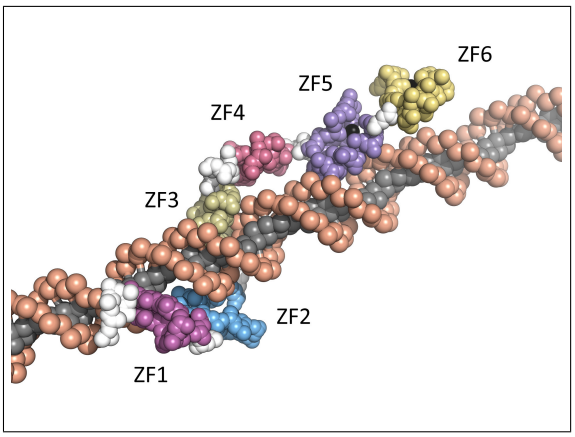


Figure 5.8: Different roles of the N-terminal first six zinc fingers of protein TFIIIA. Six zinc fingers are labeled and colored with different colors, joined by short flexible white linkers. ZF1∼ZF3 bind to DNA while ZF4 and ZF6 dissociate.

5.4.3 TFIIIA

Another test was performed via applying our CG model to the N-terminal six zinc fingers of the protein TFIIIA, which is known from the crystal structure that it tightly binds to DNA with the first three zinc fingers ZF1∼ZF3 and the fifth one ZF5, while other zinc fingers ZF4 and ZF6 dissociate. As shown in Figure 8, ZF1∼ZF3 wraps around the major groove of DNA, like the ZFs in Egr-1. In contrast, ZF4∼ZF6 align roughly parallel to the DNA axis, and form an extended structure.

The trajectory averaged minimal distance from the C_α bead to its closest

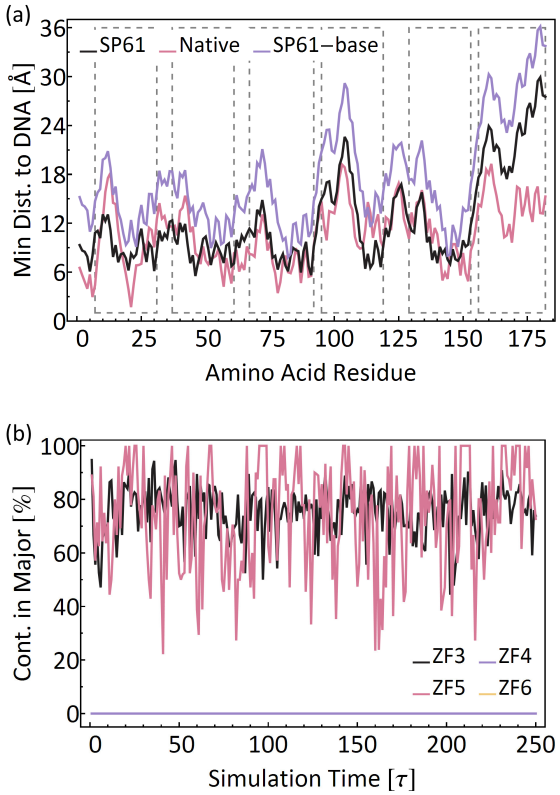


Figure 5.9: Conformation properties of TFIIIA in complex with DNA. (a) Trajectory averaged minimal distance from the C_α bead in each residue to DNA (SP61), and to DNA bases (SP61-base), compared with the distance in the crystal structure (Native). Residues in six zinc fingers are bounded via six dashed boxes. (b) Percentage of DNA contacting beads in ZF3∼ZF6 of TFIIIA located in the major groove of DNA versus the simulation time. Note that lines for ZF4 and ZF6 completely coincide with each other.

DNA (base) beads for the residues in TFIIIA is plotted in figure 9 (a), with comparison to the distance in the crystal complex. The percentage of DNA contacting beads in ZF3∼ZF6 of TFIIIA located in the major groove of DNA, as a function of the simulation time, is shown in figure 9 (b). Here the DNA contacting beads are defined as the protein beads whose minimal distance to

DNA beads is less than 7 Å.

In general the distance is smaller for the first three zinc fingers than that for the fourth and sixth one, which agrees with the partial binding partial dangling mode. If 7 Å is chosen as the threshold to determine a protein bead in contact with DNA or not, it seems that ZF4 also binds to the DNA during the simulation. However, additional analysis of the minimal distance of C_α to the DNA *bases* shows that ZF4 is far away from the DNA bases. Hence in our simulations, ZF4 is closer to the DNA backbone than in the crystal structure, which might result from the long-range coulomb interactions are not explicitly included in our force field. What's more, although the α-helix region of ZF4 is near to the DNA, it is always located in the minor groove of the DNA.

Consistent with the native complex conformation, Figure 9 (a) also shows that ZF5 contacts DNA bases. It has a wider range of motion than the first three zinc fingers, e.g., ZF3, and it resides in the major groove of DNA most of the time (see Figure 9 (b)).

5.4.4 CTCF

The trajectories averaged minimal distance from each amino acid residue to DNA of the eleven zinc finger domains of CTCF in complex with DNA (SP160) are presented in figure 5.10(a). It is quite clear that only the central five zinc fingers ZF4∼ZF8 contact DNA (see also figure 5.1(b)). while ZF5∼ZF7 are consecutively embedded in the major groove, ZF4 and ZF8 also reside in the major groove but cross the minor groove with the linkers between ZF4 and ZF5, ZF7 and ZF8, respectively. Concerning the unbound zinc fingers, one may notice the difference of the profiles monotonicity between the N- and C-terminal parts in figure 5.10(a), which indicates that the N-terminal ZF1∼ZF3 are more rigid than the C-terminal ZF9∼ZF11. The DNA bending scores of nucleotide

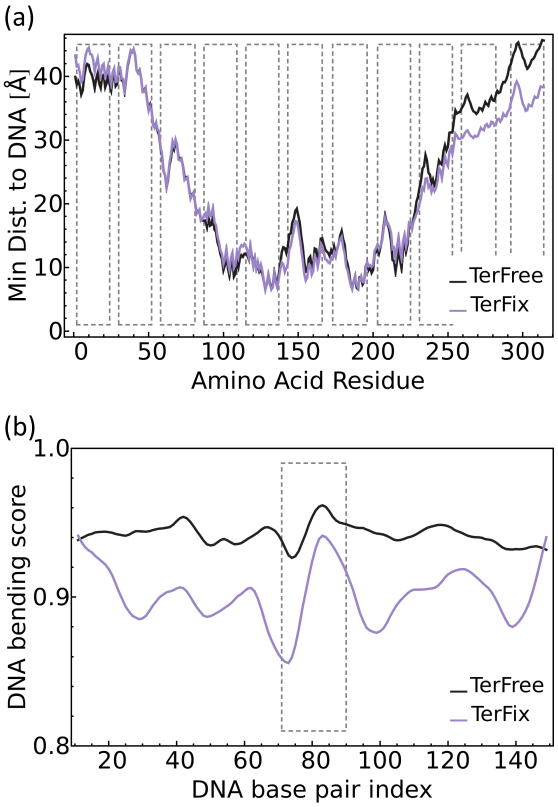


Figure 5.10: Trajectory averaged conformational properties of CTCF:SP160 complexes with or without fixed DNA terminals (TerFix or TerFree). (a) Minimal distance from the C_α bead to DNA for the zinc finger domains of CTCF, where the eleven zinc fingers are bounded by dashed boxes. (b) DNA bending score of each nucleotide base pair, where the dashed box indicates the position of the consensus 20 bp CTCF binding motif.

bps are plotted in figure 5.10(b), where the consensus 20 bp CTCF binding sites are bounded with a dashed box. As suggested by the gel mobility shift analysis experiments [122, 123], the zinc fingers of CTCF bend double stranded DNA while it binds to. The asymmetry profile in the binding sites confirmed a directional binding of CTCF, as well as different roles of different zinc fingers in CTCF.

5.5 Conclusions

To investigate how the transcription factor CTCF, as a typical mC_2H_2 zinc finger protein, organizes the spatial structure of chromatin which probably results in gene regulation, we presented a CG molecular dynamics study on mC_2H_2 zinc finger proteins in complexes with DNA. Using sequence specific "statistical potentials", which are derived from the mC_2H_2-DNA complex structures in the PDB database and tuned by cautious scaling and shifting, we confirmed several experimental key features in our simulation results. The three zinc finger protein Egr-1 slides in the major groove of DNA duplex, and contacts the DNA with the same interface as in the specific crystal structure. The first zinc finger of Egr-1 was found to undergo more intense domain motion and to dissociate from nonspecific DNA when sliding on. What's more, no matter the target DNA motif is CG-rich or AT-rich, mC_2H_2 proteins recognized their target sites and slide away from non-target sites. A further testing performed on six zinc finger domains of a mC_2H_2 protein, also presented a partial binding partial dangling mode found in the experiment.

An application to the eleven zinc finger domains of CTCF with a specific DNA duplex shows that the protein binds only with its central five zinc fingers ZF4∼ZF8. These zinc fingers are embedded in the DNA's major groove, but perhaps not continuously. ZF4 and ZF8 cross the minor groove with the C- and N-terminal linker, respectively. it is also shown that CTCF asymmetrically bends the DNA duplex. It is not clear from experiments whether a CTCF or the zinc finger domains of CTCF can organize a DNA loop alone [122, 123]. We don't find a DNA loop formed by CTCF binding in our simulation, but it is still possible that, besides the central zinc finger domains, the unstructured N- or C-terminal is necessary for CTCF to organize a DNA loop. Other possibilities,

like two CTCFs, which contacts a DNA site respectively, aggregate together and form a bridged DNA loop, will be studied in the future.

The mC_2H_2-DNA interaction potentials, we developed here, have some inborn limitations. First, some important factors, like the ionic strength in the solvent, are missing in the force field. As in the Mullinax and Noid's study [124, 125], structures deposited in PDB database can be regarded as an extended canonical ensemble composed of a collection of canonical ensembles for distinct systems (e.g., different environment's pH and salt concentration), each at a finite temperature T. Then those variables can be explicitly considered in the force field and further parameterized. Second, the statistical potentials do not contain any long-ranged interaction, such as Coulomb interaction which dominates nonspecific protein-DNA association. As a possible reason, the β-helix in ZF4 of TFIIIA is closer to the DNA backbone during our simulations than in the crystal structure. This may also influence the motion of a protein searching for its DNA target locus. Hopping and correlated transfer might be more easier with longer ranged interactions [120]. Long-ranged electrostatic interactions, as a function of salt concentration, can be *ad hoc* included into a coarse-grained force field using a Debye-Hueckel potential, as Bereau did for the *peptideB* model [87]. This will be one of next objectives to optimize the interactions.

6 General one Chain adsorbed onto another

Phase transition and winding properties of a flexible polymer adsorbed to a rigid periodic copolymer, Lei Liu, David Schubert, Min Chu and Dieter W Heermann, Physical Review E. Under review. Lei performed most simulations and wrote the manuscript. David and Min contributed other materials to this work.

Chapter Summary. Motivated by the non-covalent binding of polypeptides to DNA, the adsorption of a flexible polymer to a rigid periodic copolymer is studied in $2d$ and $3d$. The fraction of adsorbed monomers, the specific heat, and the Binder cummulant are analyzed and compared with analytical results for an ideal chain. As the interaction strength ϵ increases a second-order phase transition occurs from a non-adsorbed state to an adsorbed state, in $2d$ and a higher-order transition in $3d$. The transition point is estimated as $\epsilon_0 \sim 2.2$ for $d = 2$ and $\epsilon_0 \sim 2.1$ for $d = 3$, where ϵ is given units of $k_B T$. The dependence of the number of adsorbed monomers N_{ads} on the chain length L of the flexible polymer shows a power law scaling relation $N_{ads} \sim L^{\phi}$ with $\phi \sim 0.46, 0.42$, for $d = 2, 3$, respectively. We also find a optimal $\epsilon \sim 2.8$ for the winding of the flexible polymer around the rigid one in $3d$. Compared to the adsorbed monomers, the successive non-adsorbed monomers contribute more to the winding. When the interaction is strong $\epsilon > 3.5$, the winding value or the number of winding turns of the flexible polymer becomes linearly dependent on the chain length.

6.1 Introduction

Contrary to the traditional view that a functional protein usually possesses a stable three-dimensional structure, more and more functional intrinsically disordered protein domains of significant size are reported [42, 126]. They interact with DNA, RNA and other protein domains, and play several important roles in cells such as trascriptional regulation, translation and cellular signal transduction [127, 128]. Many intrinsically disordered proteins undergo a transition from a random-coil-like unbound state to a more ordered bound state of stable secondary or tertiary structure, i.e. a so-called 'folding while binding' process [129, 130]. For example, the binding of the multi-Cys_2His_2 (mC_2H_2) zinc finger protein, which behaves like a worm-like chain [112], to its target DNA sites results in an orientational restraint of successive zinc fingers and facilitates the whole protein to wind around the DNA along its helical major groove [43]. It is well known that the giant lose of entropy of a protein from unbound to bound state should be compensated with the protein-DNA binding enthalpy gain [131, 132].

Another macromolecular system of current interest is the polymer-carbon nanotube hybrid, which consists of a carbon nanotube (CNT) coated with a self-assembled monolayer of flexible, or semi-flexible polymer chains [133, 134]. Several experiments confirmed that wrapping is a general phenomenon occurring between polymers and CNTs, and some polymers are reported to wrap CNTs in a distinct, helical-type conformation, like poly(sacchrides) [135], poly(dialkylsilanes) [136] and single-stranded DNA [137, 138]. This non-covalent polymer wrapping can effect the properties of the CNTs, such as the solubility, dispersity, strength, toughness, and conductance, and hence enhances its functionality in numerous proposed applications [139, 140, 141].

There are studies on both intrinsically disordered protein-DNA and polymer-CNT, with either Monte Carlo or molecular dynamics methods on a coarse-grained or atomistic scale [142, 143, 144]. For example, Levy's group uncovered the asymmetric role of zinc fingers in DNA-scanning process of the inducible transcription factor Egr-1 based on a Go-type model [89, 145]. Tallury et al. found that polymers with stiff and semiflexible backbones tend to wrap around the CNTs with more distinct conformations than those with flexible backbones via atomistic molecular dynamics simulations [146, 147]. However, all these studies focused on one or a few specific molecular systems, and it was not fully understood how the adsorptive interaction between the polypeptides and DNA, or the polymer and CNT, influences the binding and the winding.

To answer these questions, a generic polymer-polymer coarse-grained model was developed. The intrinsically disordered protein is modeled as a flexible polymer chain, and the DNA is modeled as a rigid periodic copolymer. Conformation properties of the polymer-polymer complex were investigated with different adsorptive interaction and different chain lengths. The phase transition from a non-adsorbed state to an adsorbed state, and the characteristics of the flexible polymer wrapping around the rigid one are analyzed. In section 6.2, we discuss the theoretical work on the adsorption of an ideal chain. In section 6.3, the model and the simulation method are briefly introduced. Then the results for the phase transition and winding are discussed in two different parts. Finally, we present a short summary of our main conclusions.

6.2 Theory

Research interest on similar problems dates back to the 1960's. Rubin studied the adsorption of an ideal chain on a long rigid-rod molecule by the transfer-matrix method [148]. There the adsorbing rodlike molecule is represented by the

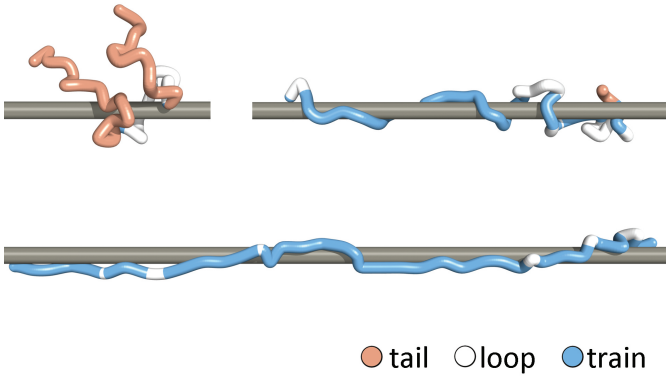


Figure 6.1: Typical conformations of a flexible polymer of chain length $L = 40$ adsorbed to a rigid polymer, with the adsorptive interaction strength $\epsilon = 0.4, 2.7, 4.7$ (top-left, top-right and bottom panel, respectively). The flexible polymers are divided into three kinds of segments, tail(brown), train(blue) and loop(white).

lattice sites on the z axis of a cubic lattice. The adsorptive interaction strength is ϵ, and the adsorption energy per monomer is $-\epsilon$ in units of k_BT. Given that the first monomer is grafted on the z axis and the length of the flexible chain approaching infinity, the average fraction of adsorbed monomers f_{ads} is found to be equal to zero below a transition point $\epsilon < \epsilon_0$. What's more, the kth derivative of f_{ads} at ϵ_0 equals zero for any $k \geq 1$ suggesting an infinite-order phase transition.

For the adsorption of an ideal chain to an impenetrable straight line in $2d$, one can directly apply the solution of the adsorption of an ideal chain to an impenetrable flat surface [149]. Consider a lattice model of the chain-surface system in which the adsorbing surface corresponds to the x-y plane and the chain is represented by a random walk in half of the space $z > 0$. Each lattice site is surrounded by Z nearest-neighbor sites, while Z_0 of them of the same z value are called to locate in the same layer. At any moment, the walker can only move to one of the current nearest-neighbor sites in the next step. For a

random walker who starts on the adsorbing surface, the probability that at the Nth step the random walker is located in the kth lattice layer is $P_k(N)$. The key recurrence equation for adsorption on a plane is

$$\begin{aligned} P_k(N) = & \frac{1}{2}aP_{k+1}(N-1) + (1-a)P_k(N-1) \\ & + \frac{1}{2}aP_{k-1}(N-1) \end{aligned} \tag{6.1}$$

for $k \geq 1$, where $a = (Z - Z_0)/Z$. This describes that if at the Nth step the random walker is in the kth layer, he must be in the $k-1$, k or $(k+1)$th layer at the $(N-1)$th step. It is reported [149] that this system undergoes a second-order phase transition as the chain length $L \to \infty$. Again the chain is non-adsorbed ($f_{ads} = 0$) below ϵ_0, but the specific heat $C = \langle (E - \langle E \rangle)^2 \rangle / L k_B T^2$ jumps discontinuously from zero to a finite peak at ϵ_0.

Concerning the problem of the adsoprtion of an ideal chain on an axis in $2d$, $P_k(N)$ can be also regarded as the probability that a random walker, in half x-y plane ($y > 0$), is located in the kth lattice layer ($y = k$) at the Nth step, when he starts from the attracting x axis. Then in a $2d$ simple square lattice, one simply sets $a = (4-2)/4 = 0.5$ and it should have the same phase transition behavior as above.

A more general scaling analysis for the adsorption of flexible chain onto any object S [150] shows that $f_{ads} = 0$ for $\epsilon < \epsilon_0$ and $f_{ads} > 0$ for $\epsilon > \epsilon_0$ when the chain length $L \to \infty$. Close to ϵ_0, the number of adsorbed monomer N_{ads} follows the relation [151]

$$N_{ads} = L^{\phi} F((\epsilon - \epsilon_0) L^{\nu}) \,, \tag{6.2}$$

where $F(x)$ is a scaling function. The relation between the crossover exponent ϕ and the critical exponent ν has been studied. Regarding the adsorption of

a polymer on a surface in $3d$, a lattice simulation by Eisenriegler, Kremer and Binder reports that $\phi \simeq 0.59 \simeq \nu_{3d}$ [152], which is also obtained by Blumen et al. [153]. Using a different algorithm, Hegger and Grassberger find $\phi \sim 0.5$ [154], and this results is supported by other simulations [155]. If the surface is penetrable and neutral (with $\epsilon_0 = 0$), ϕ is related to ν via $\phi = 1 - \nu$ [156]. Bhattacharya et al. found that the value of ϕ depends essentially on the degree of interaction between different loops in a polymer, and varies in the range of $0.34 \leq \phi \leq 0.59$ [157].

6.3 Model and Simulation

In our model, both in $2d$ and $3d$, the rigid molecule (e.g., DNA) is represented by an infinitely long copolymer with periodically distributed adsorption sites on it (see Fig.6.2). The flexible molecule (polypeptides) is modeled as a flexible polymer of length L. We implement this using the bond fluctuation model [62] for the cubic lattice, where the bond lengths of the flexible polymer can varies from 2 to $\sqrt{10}$. The rigid polymer lies on the x axis. The distance between the adsorbing sites is 3. Because of the excluded volume, one adsorption site can not be occupied by two monomers simultaneously, and the distance from the monomer of the flexible polymer to the rigid one is $s \geq 2$. One monomer of the flexible polymer is considered to locate on the surface of the rigid molecule if $s = 2$ in $2d$, or $2 \leq s \leq \sqrt{8}$ in $3d$. But it is adsorbed *only* when it resides on the surface of an adsorbing site, i.e., it has the same x coordinate value as an adsorbing site.

The simulations were performed using the standard Metropolis algorithm [50], where an adsorbed monomer can only leave the adsorption site with a probability $\exp(-\epsilon/k_B T)$. It is guaranteed that at least one monomer of the flexible polymer, not adsorbed necessarily, is on the surface of the rigid copoly-

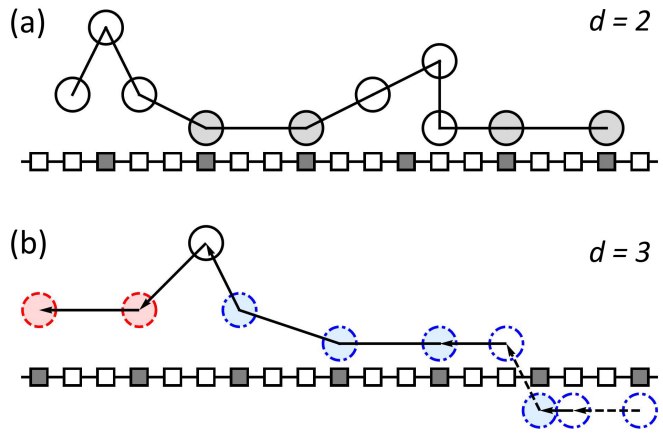


Figure 6.2: An illustration of our model in $2d$ (a) and in $3d$ (b). The z-axis is perpendicular to and points out of the plane. Monomers of the rigid polymer are represented by squares, where the adsorbing sites are colored gray. Monomers of the flexible polymer are represented by circles of solid black edges ($z = 0$), dashed red edges ($z > 0$) and dash-doted blue edges ($z < 0$). Circles representing adsorbed monomers are filled. In (b), bonds from monomer i to $i+1$ are draw using solid lines (parallel to the plane), solid arrows (going out of the plane) and dashed arrows (going into the plane).

mer. For different $L \in \{10, 20, 40, 80, 160, 200\}$ and $\epsilon \in [0.0, 5.0]$ with a step 0.1, we first calculated the sampling interval Δt from the autocorrelation time of the radius of gyration R_g of the flexible polymer (e.g., $\Delta t \sim 10^7$ MC steps for $L = 200$, $\epsilon = 3.0$), Then after equilibration, 10^4 independent conformations with interval Δt were sampled for each pair of parameters $\{L, \epsilon\}$ to calculate the ensemble averaged properties of interest.

Since the first monomer of the flexible chain is always fixed and adsorbed in the theoretical work, we also calculated the adsorption with the first monomer grafted.

6.4 Results

6.4.1 Phase Transition

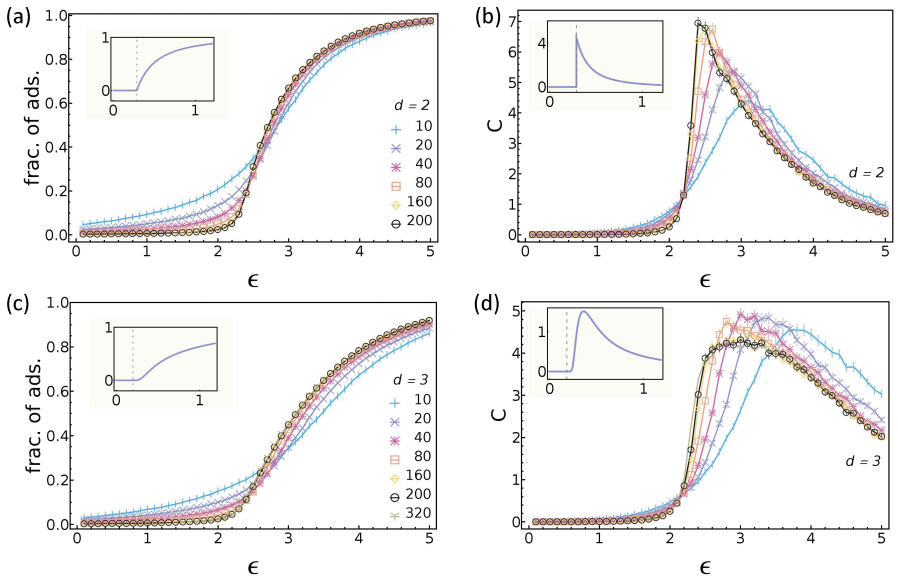


Figure 6.3: Fraction of adsorbed monomers (a, c) and specific heat (b, d) for different chain lengths $L \in \{10, 20, 40, 80, 160, 200, 320\}$ and different adsorptive interaction strength ϵ in dimension $d = 2, 3$. Subplot (b, d) has the same legend as (a, c), respectively. The corresponding theoretical results for an ideal chain are shown in the inserts, while the vertical dashed line indicate the location of the transition point ϵ_0.

The dependence of the fraction of adsorbed monomers f_{ads} and the specific heat C on the adsorptive interaction strength ϵ for various chain lengths are presented in Fig.6.3. In both dimensions, due to the finite size effect, the transition gets sharper when L increases. For longer polymers, the f_{ads} is almost zero for small ϵ. It is apparent that there is a steeper rise within the transition region in $2d$ than in $3d$. Also a higher fraction of the flexible polymer is adsorbed in $2d$ than in $3d$ when the interaction is strong (e.g., $\epsilon = 5.0$). Concerning the specific

heat, for $L = 200$, C roughly jumps vertically to a higher peak in $2d$, while it climbs up to a lower maximum with a flatter slope in $3d$. All of these features appear in the theoretical results for an ideal chain too. Therefore we expect a second-order phase transition for $d = 2$, and a higher (larger than two) order transition for $d = 3$. In addition, as the flexible polymer becomes longer and longer, the peak height of the specific heat increases monotonically in $2d$, and it starts increasing followed by a decline in $3d$. However, it converges in either case. Taking the number of adsorbed monomers N_{ads} as an order parameter, the dependence of susceptibility $\chi =< N_{ads}^2 > - < N_{ads} >^2$ on ϵ as $L \to \infty$ (data not shown here) also support the conclusion drawn from the specific heat.

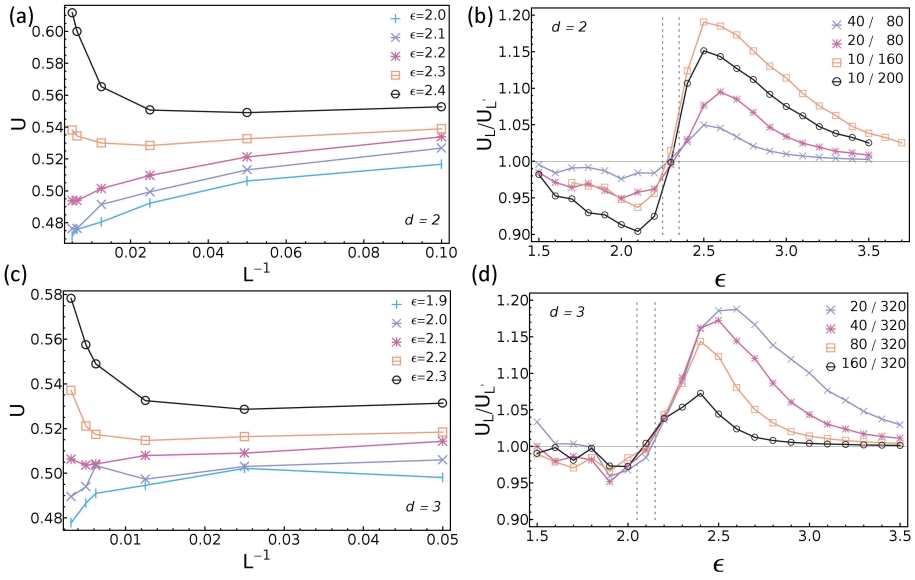


Figure 6.4: Binder cummulant U versus the inverse of chain length L^{-1} around ϵ_0 (a, c), and the ratio between U of different pairs of chain lengths $\{L/L'\} \in \{20/320, 40/320, 80/320, 160/320\}$ versus the adsorptive interaction strength ϵ (b, d) in $d = 2, 3$. The intersection points are located within the pair of vertical dash lines in (b, d).

In order to determine the transition point ϵ_0, we perform the analysis of

the Binder cummulant $U = 1 - \langle N_{ads}^2 \rangle / 3 \langle N_{ads} \rangle^2$ [50, 151]. It is known that, providing the chain length $L \to \infty$, U approaches 2/3 for $\epsilon > \epsilon_0$, and it tends to a nonzero value at ϵ_0 independent of L. Hence, for pairs of different finite chain lengths $\{L, L'\}$, the ratio between the Binder cummulants $U_L/U_{L'}$ should equal to one near the transition point. Fig.6.4 (a, c) shows U as a function of L^{-1} around ϵ_0 for $d = 2, 3$, respectively. U shows different behavior for $\epsilon > \epsilon_0$ and $\epsilon < \epsilon_0$ in both dimensions. This suggests that $\epsilon_0 \sim 2.3$ in $2d$ and $\epsilon_0 \sim 2.1$ in $3d$. The ratio of the Binder cummulants for different pairs of chain lengths versus the interaction strength are plotted in Fig.6.4 (b, d). According to the points, which cross over the horizontal line $U_L/U_{L'}$=1, we found $2.25 < \epsilon_0 < 2.35$ and $2.05 < \epsilon_0 < 2.15$, for $d = 2$ and 3, respectively. But this only gives us a rough range of the transition point.

Since it is reported [155, 153, 151] that the ratio between the perpendicular and parallel components of the mean square radius of gyration $\langle Rg_\perp^2 \rangle / \langle Rg_\parallel^2 \rangle$ should be independent of the chain length L at the transition point in the surface adsorption problem, the dependence of this ratio on the interaction strength for different chain lengths are presented in Fig.6.5. In $2d$, the curves for different L intersect at $\epsilon_0 = 2.2$. But in $3d$, they collapse onto each other at low adsorptive interaction, and do not intersect at one clear point. The difference can be explained if one notice that the flexible polymer can wrap around the rigid polymer in $3d$, but not in $2d$ (due to the dimensionality of the space and the excluded volume interaction between the flexible polymer and the rigid one). Swelling perpendicularly at small ϵ leads to a larger $Rg_\perp$. Because of the same reason, $\langle Rg_\perp^2 \rangle / \langle Rg_\parallel^2 \rangle \sim 1$ for $d = 2$, but $\langle Rg_\perp^2 \rangle / \langle Rg_\parallel^2 \rangle > 2$ for $d = 3$, for weak adsorptive interaction.

We have also measured the dependence of the number of adsorbed monomers N_{ads} on L in the transition region (see Fig.6.6(a, c) for $d = 2, 3$ respectively). In

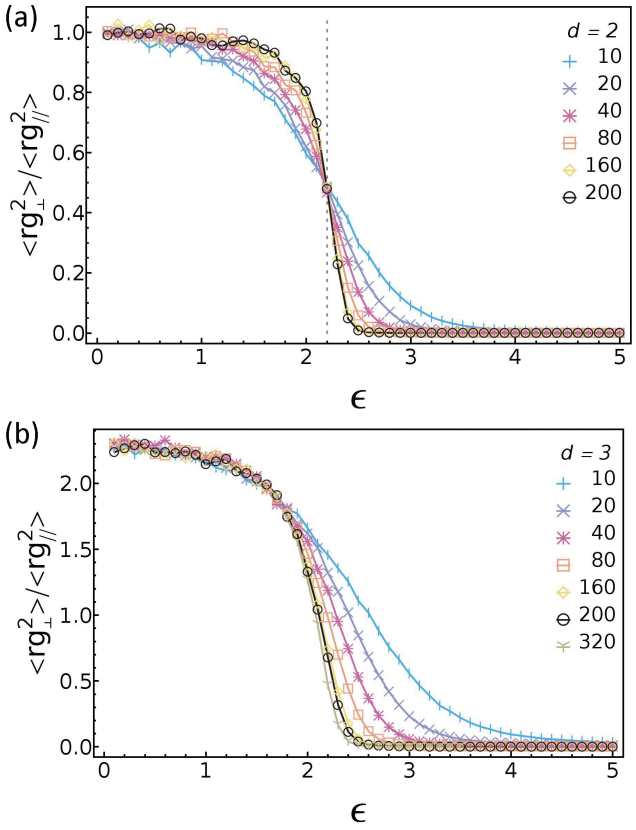


Figure 6.5: The ratio between the perpendicular and parallel components of the mean square radius of gyration $\langle Rg_{\perp}^2 \rangle / \langle Rg_{\parallel}^2 \rangle$ versus ϵ in $2d$ (a) and $3d$ (b). The vertical dash line in (a) indicates the location of the intersection point.

agreement with the scaling analysis, a linear curve in the log-log plot indicates a power law relation $N_{ads} \sim L^{\phi}$ for both dimensions. The exponent values of $\{\phi, \nu\}$ are further calculated by fitting the scaling

$$\begin{aligned} N_{ads} L^{-\phi} =& a_0 + a_1(\epsilon - \epsilon_0) L^{\nu} + \\ & O((\epsilon - \epsilon_0)^2 L^{\nu})] \,, \end{aligned} \tag{6.3}$$

following the method from Luo [151]. In brief, taking $2d$ as an example, N_{ads} at

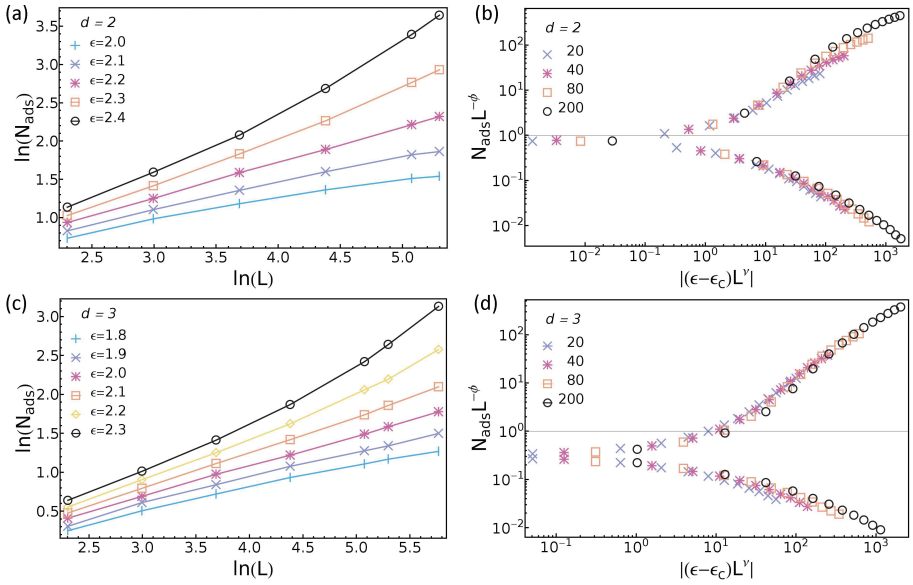


Figure 6.6: Log-log plot of the number of adsorbed monomers N_{ads} versus the chain length L around ϵ_0 (a, c), and the scaling of N_{ads} with $\{\epsilon, \phi, \nu\}$ equals $\{2.20, 0.46, 0.58\}$ in $2d$ (b), and $\{2.05, 0.42, 0.57\}$ in $3d$ (d) for various chain lengths.

$\epsilon \in [2.0, 2.4]$ with a step of 0.01 are obtained from quadratic interpolation from the simulation data at $\epsilon \in \{2.0, 2.1, 2.2, 2.3, 2.4\}$. Then, $\{\epsilon_0, \phi\}$ are determined by a best fit to a power law. ν is the value which minimizes the deviation from the relation $N_{ads}L^{\phi} \sim (\epsilon - \epsilon_0)L^{\nu}$ of the simulation data to a parabolic function. Fig.6.6(b, d) shows the scaled N_{ads} with $\epsilon_0 = 2.20$, $\phi = 0.46$, $\nu = 0.58$ for $d = 2$, and $\epsilon_0 = 2.05$, $\phi = 0.42$, $\nu = 0.57$ for $d = 3$. We can see that all data collapse quite well even for ϵ far from ϵ_0. It is quite interesting to find that our fitted values satisfy $\phi \sim 1 - \nu$, which was proposed by de Gennes [156].

Finally, we compare f_{ads} with the first monomer always grafted to our model (non-grafted polymer) in Fig.6.7(a, c). When the polymer is short, the grafted polymer always has a higher fraction of adsorption than the non-grafted one, but the difference between them diminishes as the polymer gets longer. Hence the above discussion about the phase transition should also apply to the grafted

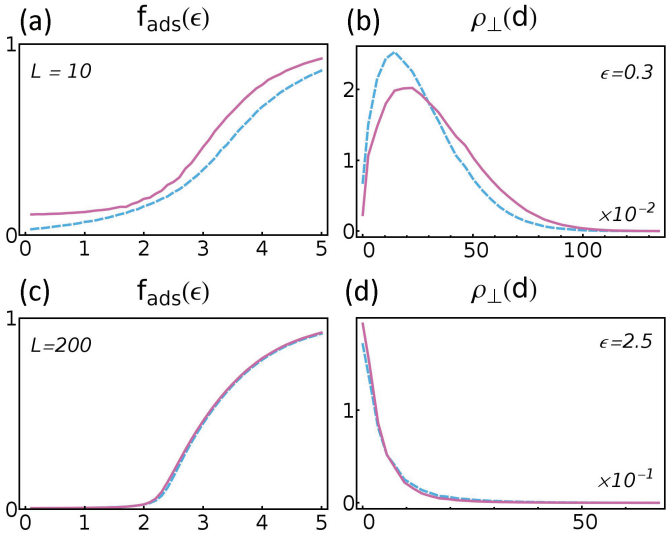


Figure 6.7: Fraction of adsorbed monomers for non-grafted (dash line) and grafted polymer (solid line) with $L = 10, 200$ in (a, c), respectively. The perpendicular monomer density $\rho_\perp$ versus the distance from the monomer to the rigid polymer surface for non-grafted (dash line) and grafted polymer (solid line) with $L = 200$, and $\epsilon = 0.3, 2.5$ in (b, d), respectively.

polymer, providing L approaches infinity. However, deviations can be found if one looks at the perpendicular monomer density $\rho_\perp$ profile for the grafted and non-grafted polymer in Fig.6.7(b, d). At low adsorptive interaction, compared to the non-grafted polymer, the grafted one is expelled further away from the rigid polymer. At high interaction, since most part of the polymer is adsorbed on the surface, this difference disappears.

6.4.2 Winding Properties

Another property of special interest is how the flexible polymer winds or wraps around the rigid one in $3d$. The winding value w is defined as a function of the

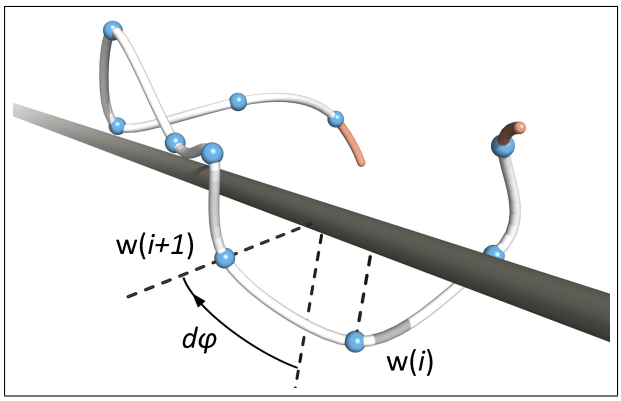


Figure 6.8: The winding value w, as a function of the contour length l along the flexible polymer, varies from $w(i)$ to $w(i+1)$ with $d\varphi$, by which the flexible polymer rotates around the rigid one from monomer i to $i+1$. One turn is counted if $|\Delta w|$ exceeds 2π.

contour length l of the flexible polymer, for $l \in \{1, 2, ..., L\}$. We have

$$w(i+1) = w(i) + d\varphi\,, \tag{6.4}$$

while the flexible polymer rotates around the rigid one from monomer i to $i+1$ by an angle $d\varphi$ (see Fig.6.8). One turn is counted if $|\Delta w| = |w_e - w_s|$ exceeds 2π, where w_s and w_e is the winding value at the head and tail of a segment of the flexible polymer, respectively. Looking along the rigid molecule, the flexible polymer can wind either clockwise ($w > 0$) or anti-clockwise ($w < 0$) with equal probability, and one would expect $\langle w \rangle = 0$. Hence we choose w^2 and plot $\langle w^2/L \rangle$ versus ϵ for various chain lengths in Fig.6.9.

With strong adsorptive interaction, monomers of the flexible polymer are almost adsorbed. A local move parallel to the rigid polymer, from an adsorbing site to a non-adsorbing site, is energetically unfavorable. One can assume that each monomer moves only perpendicular to the rigid polymer stochastically clockwise and anti-clockwise, while still keeps its distance to an adsorbing site

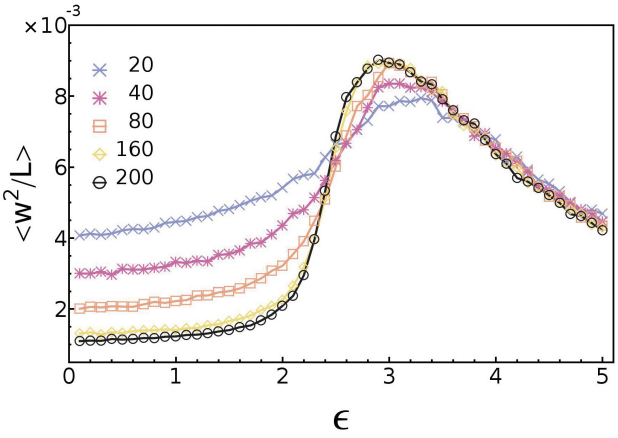


Figure 6.9: The chain length normalized mean square winding value $\langle w^2 \rangle / L$ versus the adsorption energy ϵ for different L.

not larger than $\sqrt{8}$, i.e., still adsorbed. Similar to $\langle \Delta D^2 \rangle \sim t$, where ΔD is the displacement and t is the elapsed time in one dimensional diffusion, this assumption yields that with large ϵ,

$$\langle w^2 \rangle \sim L \tag{6.5a}$$

$$N_{turn} \sim L \tag{6.5b}$$

where N_{turn} is the mean number of turns. For $\epsilon > 3.5$, the curves of $\langle w^2/L \rangle$ for different chain lengths collapse (see Fig.6.9), which validates Eq.6.5(a). We have also plotted the N_{turn} as a function of L at $\epsilon \in \{3.0, 3.5, 4.0, 4.5, 5.0\}$ in Fig.6.10. The linearly fitted dash lines for all these interaction strength confirm the above analysis too.

The flexible polymer is divided into three kinds of segments to further understand the dependence of $\langle w^2 \rangle$ on ϵ. The non-adsorbed successive monomers at the terminals of the chain are called *tail*, in the middle are called *loop*, and the adsorbed successive monomers are called *train* (see Fig.6.1). Given the length

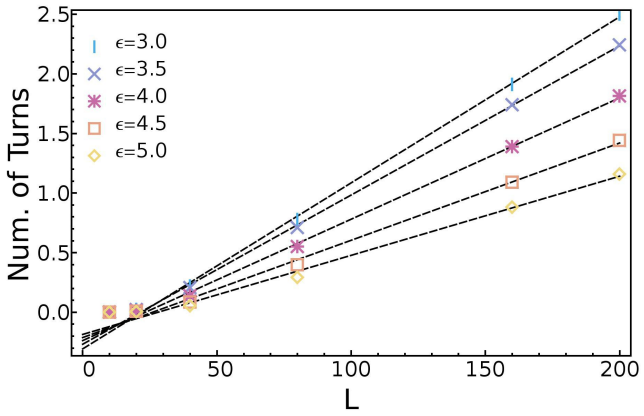


Figure 6.10: The mean number of turns N_{turn} versus the chain length L at strong adsorptive interaction $\epsilon \geq 3.0$. The dash lines are the linear fitted curves.

of a segment is L_s, we define the squared winding value per monomer for this segment as $w^2_{mono} = w^2/L_s$.

Fig.6.11(a) shows the fraction of monomers in tail f_{tail}, loop f_{loop} and train f_{train} as a function of ϵ. With low attractive interaction strength, the tails dominate the polymer. With high attractive interaction strength, the majority of the monomers belong to the trains (see also Fig.6.1). As ϵ increases, f_{tail} or f_{train} changes monotonically, while f_{loop} has a maximum in the range $2.5 < \epsilon < 3.0$.

Furthermore, Fig.6.11(b) presents the mean squared winding value per monomer $\langle w^2_{mono}\rangle$ versus ϵ for three kinds of segments. It shows that $\langle w^2_{mono}\rangle(loop) > \langle w^2_{mono}\rangle(tail) > \langle w^2_{mono}\rangle(train)$ for $\epsilon < 4.0$. Since every monomer in a train is confined on the surface of the rigid polymer and each bond of certain length can not step over a large $d\varphi$, $\langle w^2_{mono}\rangle(train)$ is comparatively small. As for $\langle w^2_{mono}\rangle$ in loop, compared to that in the tail, the additional grafted end impedes the non-adsorbed segment to align parallel to the rigid polymer, which hence results in a larger winding.

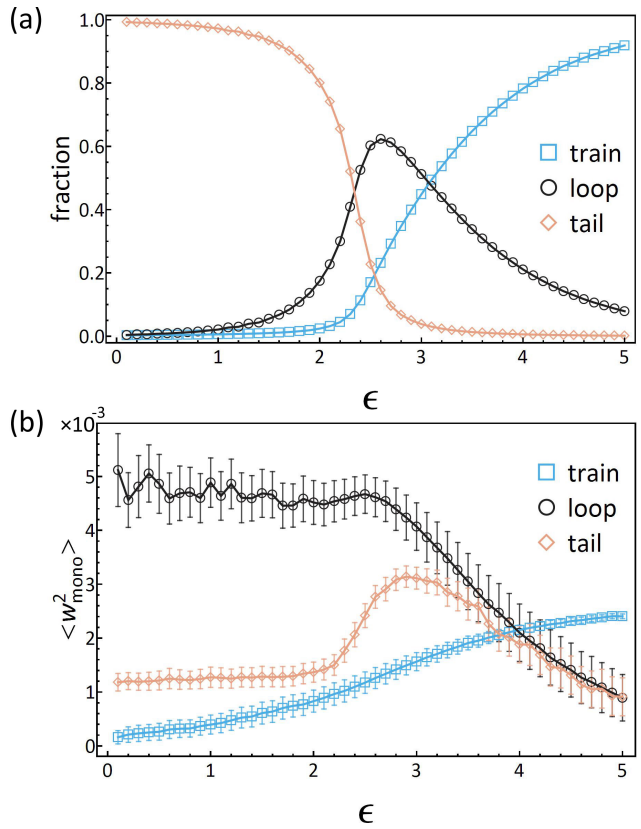


Figure 6.11: The fraction of adsorbed monomers (a) and the mean squared winding value per monomer $< w^2_{mono} >$ (b) in train, loop and tail versus ϵ for $L = 200$.

These two factors together explain why there is a peak for the winding of the whole chain around $\epsilon \sim 2.8$ in Fig.6.9 (see also Fig.6.1), where we have $< w^2/L > \sim f_{train} \times < w^2_{mono}(train) > + f_{loop} \times < w^2_{mono}(loop) > + f_{tail} \times < w^2_{mono}(tail) >$. Finally, we stress that the winding properties analyzed here do not necessarily mean a periodic helical conformation of the flexible polymer wrapping around the rigid one. It has been pointed out that the bending rigidity [146, 147, 158] and weak attraction between non-adjacent monomers of a semi-flexible chain [142] play key roles in forming periodic helical winding on a adsorbing cylinder surface. We have also calculated the periodic correlation

function [142] from conformations for all the studied interaction strength, no ensemble meaningful periodicity is found.

6.5 Conclusion

In this work, we studied a generic polymer-polymer model for the adsorption of a flexible molecule onto a rigid molecule using the Monte Carlo method. In agreement with the theoretical results for the adsorption of a grafted ideal chain, our data show a steeper transition, namely from a non-adsorbed state to an adsorbed state, in $2d$ than in $3d$. Also considering the dependence of the Binder cummulant on the adsorption interaction strength, we conclude that there is a second-order phase transition in two dimension, and a higher-order transition in three dimension. Both the crossing of the Binder cummulant and the ratio of the perpendicular to parallel components of the radius of gyration, indicate the transition point $\epsilon \sim 2.2$ in $2d$, and $\epsilon \sim 2.1$ in $3d$. Further analysis of the scaling of the number of adsorbed monomers with the chain length shows an expected power law relation close to the transition point. In addition, calculation of the winding value of the flexible polymer around the rigid polymer in $3d$ shows that the successive non-adsorbed monomers, which we called *loop*, contribute most to the winding. It leads to an optimum ϵ of medium strength 2.8 for the winding of the whole chain. Here the important role played by the *loop* reminds us of the similar function of the linker peptide of a protein [43]. Taking the mC_2H_2 zinc finger protein wrapping around its target DNA site as an example, usually the C_2H_2 zinc finger domains are bound to the DNA, while the flexible linker peptides between these domains are unbound. Finally it is also shown that, with high interaction strength, the dependence of the winding and the number of turns of flexible polymer on the chain length becomes linear.

In our model, the periodicity of the adsorbing sites on the rigid polymer is

set to 3, which is the integer closest to the priori mean bond length of a flexible polymer in the bond fluctuation model [62]. The resulting transition energy ϵ_0 is larger than that of the adsorption onto a homogeneous rigid polymer. If the periodicity is enlarged, two effects are expected. First, monomers in the flexible chain cannot be adsorbed successively any longer, and the saturation value of f_{ads} will decrease. Second, the transition energy ϵ_0 will increase. These tendencies have been investigated by other studies, such as a Monte Carlo simulation of the adsorption of periodic copolymers at a homogeneous planar substrate [159], and a numeric solution of a directed walk model of a homogeneous polymer adsorbed onto a surface with periodic adsorbing strip pattern [160].

7 Conclusion and Outlook

7.1 A summary of the results

Aiming to better understand how CTCF organizes the chromatin loops in human cells, different models on different scales are developed, simulated and analyzed in this thesis. The findings contribute to new insight into the unbound mC_2H_2 zinc finger proteins and mC_2H_2 zinc finger proteins in complexes with DNA.

In *chapter 4*, we studied the conformational properties of unbound mC_2H_2 zinc finger proteins using multiscale approaches. First, a homology model of the tandem zinc finger domains of the transcription factor CTCF was constructed. All-atom MD simulations showed that single zinc finger is a stable structural unit, independent of the studied environmental conditions. In agreement with the NMR observation of the N-terminal three zinc fingers of TFIIIA, the polypeptide becomes more extended in unbound states than in a DNA-bound state. Next, an atomistic pivoting algorithm, which considers only the excluded volume interaction, was developed to investigate the global conformational characteristics of multi-zinc finger proteins. It showed that as the number of zinc fingers increases, the end-to-end distance distribution gradually changes

its shape, from skewed to the left to skewed to the right. This was explained by using a worm-like chain model. The *effective* bending constraint can be applied not only to multi-zinc finger proteins, but also to other multi-domain proteins connected by short flexible linkers. Finally, a mesoscale peptide model was modified for mC_2H_2 proteins, which is efficient while providing similar conformational properties as those given by atomistic models.

In *chapter 5*, we presented a CG molecular dynamics study on mC_2H_2 zinc finger proteins in complexes with DNA. Using sequence specific "statistical potentials", which were derived from the mC_2H_2-DNA complex structures in the PDB database and tuned by cautious scaling and shifting, we confirmed several experimental key features in our simulation results. The three zinc finger protein Egr-1 slides in the major groove of DNA duplex, and contacts the DNA with the same interface as in the specific crystal structure. The first zinc finger of Egr-1 was found to undergo more intense domain motion and to dissociate from nonspecific DNA when sliding on. What's more, no matter the target DNA motif is CG-rich or AT-rich, mC_2H_2 proteins recognize their target sites and slide away from non-target sites. A further testing performed on six zinc finger domains of a mC_2H_2 protein, also presented a partial binding partial dangling mode found in the experiment.

An application to the eleven zinc finger domains of CTCF with a specific DNA duplex showed that the protein binds only with its central five zinc fingers ZF4∼ZF8. These zinc fingers are embedded in the DNA's major groove, but perhaps not continuously. ZF4 and ZF8 cross the minor groove with the C- and N-terminal linker, respectively. It is also shown that CTCF asymmetrically bends the DNA duplex. It is not yet clear from experiments whether a CTCF or the zinc finger domains of CTCF can organize a DNA loop alone [122, 123]. We didn't find a DNA loop formed by a single CTCF binding in our simulation,

but there still exists other possibilities.

In *chapter 6*, we studied a generic polymer-polymer model for the adsorption of a flexible molecule onto a rigid molecule using the Monte Carlo method. In agreement with the theoretical results for the adsorption of a grafted ideal chain, our data show a steeper transition, namely from a non-adsorbed state to an adsorbed state, in $2d$ than in $3d$. Also considering the dependence of the Binder cummulant on the adsorption interaction strength, we concluded that there is a second-order phase transition in two dimension, and a higher-order transition in three dimension. Both the crossing of the Binder cummulant and the ratio of the perpendicular to parallel components of the radius of gyration, indicate the transition point $\epsilon \sim 2.2$ in $2d$, and $\epsilon \sim 2.1$ in $3d$. Further analysis of the scaling of the number of adsorbed monomers with the chain length shows an expected power law relation close to the transition point. In addition, calculation of the winding value of the flexible polymer around the rigid polymer in $3d$ shows that the successive non-adsorbed monomers, which we called *loop*, contribute most to the winding. It leads to an optimum ϵ of medium strength 2.8 for the winding of the whole chain. Here the important role played by the *loop* reminds us of the similar function of the linker peptides of a protein. Taking the multi-C_2H_2 zinc finger protein wrapping around its target DNA site as an example, usually the C_2H_2 zinc finger domains are bound to the DNA, while the flexible linker peptides between these domains are unbound. Finally it is also shown that, with high interaction strength, the dependence of the winding and the number of turns of flexible polymer on the chain length becomes linear.

7.2 Outlook

Nowadays computing power puts a huge hindrance to study complex systems, like CTCF in complex with DNA, by performing classical atomistic molecular

dynamics simulations for meaningful long simulation time. As an example, in *chapter 4*, tens of thousands of CPU hours are still not long enough to well sample the end-to-end distance distribution of an unbound three-C_2H_2 zinc finger protein. Multiscale approaches are interesting, more important necessary for studying large systems and long-time dynamics. They have been successfully applied in systems like proteins, lipid membranes and other biomolecular systems [161, 162, 163, 164, 165].

There are two popular ways to develop a coarse-grained force field, namely the *top-down* fashion and the *bottom-up* fashion. The top-down approach is to tune the parameter values in the force field so that simulations of the coarse-grained model can reproduce certain properties of interest of some reference systems, such as the 3SPN model (appendix A), the PeptideB model (appendix B), and the cross-parameterization of mC_2H_2-DNA interaction potentials in *chapter 5*.

On one hand, we can further apply this model to other systems. There we only examined CTCF binding conformations at one specific target DNA locus, and it shows that CTCF binds to DNA by using its central zinc fingers. Other possible CTCF binding modes at different DNA loci [38] could be similarly studied. More complex systems, like two CTCFs which contacts a DNA site each, aggregating together and forming a bridged DNA loop, can also be extended in the future.

On the other hand, the derived mC_2H_2-DNA interaction force field has some limitations, which may be further optimized. First, some important factors, like the ionic strength in the solvent, are not considered in the force field. As in the Mullinax and Noid's study [124, 125], structures deposited in PDB database can be regarded as an extended canonical ensemble composed of a collection of canonical ensembles for distinct systems (e.g., different environment's pH

and salt concentration), each at a finite temperature T. Then those variables can be explicitly considered in the force field and further parameterized. Second, the statistical potentials do not contain any long-ranged interaction, such as Coulomb interaction which dominates nonspecific protein-DNA association. This may also influence the motion of a protein searching for its DNA target locus. One can *ad hoc* include long-ranged electrostatic interactions, as Bereau did for the PeptideB model [87].

The bottom-up approach, as the other way, derives the parameter values in a coarse-grained force field from simulation results of a model with higher resolution and reliability, like atomistic molecular dynamics simulation. The main idea is to sample the part of the phase space in the coarse-grained simulation, which is sampled by the atomistic simulation, with the same probability distribution of a *certain* property [166, 167, 168]. Hence depending on the type of the reference property, variant systematic bottom-up strategies have been developed, such as the force-matching algorithm [169, 170, 125, 171, 172], structure-matching algorithm [173, 174, 175, 124] and the relative entropy algorithm [176, 177, 178]. However, the resulting coarse-grained force field is state point dependent (e.g., relying on the temperature at which atomistic simulation is performed) and not necessarily readily transferable. What's more, it does not guarantee that thermodynamic properties of the reference system, like phase behavior, can be well reproduced. Methodological issues and challenges are waiting to be addressed. In any case it would be interesting to develop a mesoscale model for mC_2H_2-DNA complexes using a bottom-up strategy, and compare it with our current model.

A 3SPN model

Simulations of DNA with atomistic details are usually limited to tens of base pairs in length, or ten of nanosecond in time. Depending on the problems of interest, a number of coarse-grained models for DNA on variant scales have been developed. For example, DNA can be simply represented as a rigid or semiflexible rod with charges uniformly distribued along the rod to investigate the effect of electrostatic interactions and molecular stiffness on the distribution of other ionic species on the surface of DNA. It can also be modeled by a bead-spring representation to study the spatail structure of whole genome, where thousands or millions of base pairs are treated as a monomer. However, many phenomena, like binding of proteins, require simulations on length scales between $\sim$ 2 nm and $\sim$ 2 μm with molecular details. Coarse-grained models, in which one nucleotide is represented by n sites, can fulfill both the demand of higher resolution and the demand of reduced computation.

As our choice for simulations of DNA on mesoscale, the *3-site-per-nucleotide* (3SPN) model proposed by Pablo and his colleagues [113, 114, 179] is introduced here (see Figure A.1). Each nucleotide is reduced to three interacting beads, one each for the phosphate group P, the sugar group S and the base group B. Note that $n = 3$ is the minimum number of sites that distinguish major and

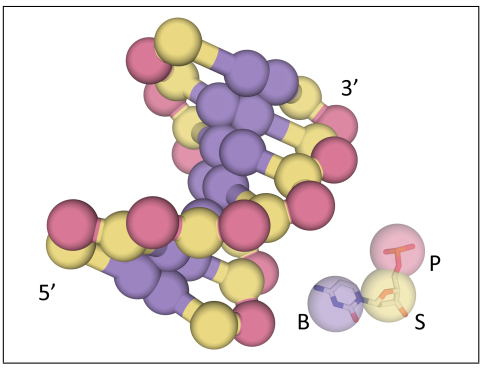


Figure A.1: 3SPN model where three sites, one each for the phosphate group P, the sugar group S and base group B (A,T,C,G), make up a nucleotide. The major and minor grooves are distinguishable.

minor grooves using isotropic potentials. Eight distinct interactions contribute to the accompanying force field,

$$V_{3SPN.1} = V_{bond} + V_{angle} + V_{dihedral} + V_{stack} + V_{bp} + V_{nnat} + V_{elec} + V_{solv}, \quad \text{(A.1)}$$

which are parameterized via reproducing the salt-dependent melting and persistence length of DNA. The first three terms are bonded interactions,

$$V_{bond} = \sum_{i=1}^{N_{bond}} [\kappa_1 (d_i - d_{0i})^2 + \kappa_2 (d_i - d_{0i})^4] \quad \text{(A.2)}$$

$$V_{angle} = \sum_{i=1}^{N_{angle}} \frac{\kappa_\theta}{2} (\theta_i - \theta_{0i})^2 \quad \text{(A.3)}$$

$$V_{dihedral} = \sum_{i=1}^{N_{dihedral}} \kappa_\phi [1 - \cos(\phi_i - \phi_{0i})], \quad \text{(A.4)}$$

which have typical two-, three- and four-body expressions for intramolecular constraints about bonds length, bond angles and dihedral angles. The bond length constant $\kappa_{1,2}$, the bending constant κ_θ and the tersional constant κ_ϕ describe the strengths of these constaints. The equilibrium distances and angles

$\{d_{0i}, \theta_{0i}, \phi_{0i}\}$ are set equal to the values of standard B-form double-strand DNA, which are summarized in Table A.1. Remaining nonbonded pairwise interactions are given by

$$V_{stack} = \sum_{i<j}^{N_{stack}} 4\varepsilon[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] \tag{A.5}$$

$$V_{bp} = \sum_{i=1}^{N_{bp}} 4\varepsilon_{bi}[5(\frac{\sigma_{bi}}{r_{ij}})^{12} - 6(\frac{\sigma_{bi}}{r_{ij}})^{10}] \tag{A.6}$$

$$V_{nnat} = \sum_{i<j}^{N_{nnat}} \begin{cases} 4\varepsilon[(\frac{\sigma_0}{r_{ij}})^{12} - (\frac{\sigma_0}{r_{ij}})^{6}] + \varepsilon & r_{ij} < r_{cut}, \\ 0 & \text{otherwise} \end{cases} \tag{A.7}$$

$$V_{elec} = \sum_{i<j}^{N_{elec}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon(T,C)r_{ij}} e^{-r_{ij}/\lambda_D} \tag{A.8}$$

$$V_{solv} = \sum_{i<j}^{N_{solv}} \varepsilon_s[1 - e^{-\alpha(r_{ij}-r_s)}]^2 - \varepsilon_s \tag{A.9}$$

V_{stack} accounts for the base-stacking interactions (an intrastrand effect) using a G$\overline{o}$-type formula. It applies to all intrastrand native-contact pairs of sites, where the distance between the sites is found within 9 Å. Hence the equilibrium distance σ_{ij} in Equation A.5 is pair dependent.

V_{bp} describes the hydrogen bonding between any complementary base pair (A-T or C-G), both intrastrand and interstrand, which does not participate in V_{stack}. Based on the base pair, it is characterized by an energy constant $\varepsilon_{bi} \in \{\varepsilon_{AT}, \varepsilon_{CG}\}$ and a distance constant $\sigma_{bi} \in \{\sigma_{AT}, \sigma_{CG}\}$.

V_{nnat} is a purely repulsive, excluded volume interaction which has a Weeks-Chandler-Anderson (WCA) form. For mismatched base sites of distance within $r_{cut} = 1.00$ Å, σ_0 is set to $2^{-1/6} \cdot 1.00$ Å. In other cases, $r_{cut} = 6.86$ Å and $\sigma_0 = 2^{-1/6} \cdot 6.86$ Å, where 6.86 Å is a mean pairwise separation in this model.

The last two terms depends on the solvent (NaCl) concentration. The electrostatic contributions from pairs of phosphate groups, which are excluded from

V_{angle}, are included into V_{elec} using Debye-Hueckel theory. The Debye length λ_D, which defines the spatial extend of charge screening, is given by

$$\lambda_D = [\frac{\epsilon_0 \epsilon(T,C) k_B T}{2 N_A e_0^2 I}]^{1/2}, \tag{A.10}$$

where I is the ionic strength of the solution in units of mM ($\mathrm{mol} \cdot \mathrm{m}^{-3}$). The relative dielectric constant $\epsilon(T,C)$, as a function of the temperature T and salt concentration C, is given by

$$\epsilon(T,C) = \epsilon(T) a(C) \tag{A.11}$$

$$\epsilon(T) = 249.4 - 0.788T + 7.20 \times 10^{-4} T^2 \tag{A.12}$$

$$a(C) = 1.000 - 0.2551C + 5.151 \times 10^{-2} C^2 - 6.889 \times 10^{-3} C^3, \tag{A.13}$$

where T and C are measured in K and M respectively. $\epsilon(T)$ is the static (zero-frequency) dielectric constat at temperature T, and $a(C)$ is the salt correction. Together, there equations A.8 and A.10-13, determine the electrostatic interactions.

Finally, V_{solv} implicitely represents many-body effects associated with the arrangement of water molecules during the reversible denaturation of DNA. In Equation A.9, N_{solv} are all possible pairs of interstrand sugar sites. By comparing the melting enthalpies and heat capacities predicted by the simulations and experiments, the interaction strength ε_s, which depends on the salt concentration C and the chain length of DNA N_{nt}, is parameterized as

$$\varepsilon_s \approx \varepsilon_N A_I \tag{A.14}$$

$$\varepsilon_N = 0.504982\varepsilon[1 - (1.40418 - 0.268231 N_{nt})]^{-1} \tag{A.15}$$

$$A_I = 0.474876[1 + (0.148378 + 10.9553C)^{-1}]. \tag{A.16}$$

All values of relevant parameters are collected in Table A.1 and A.2.

Table A.1: Equilibrium distances and angles for bonded interactions. 5' and 3' labels the direction of connections. Hence S(5')-P represents a bond between a phosphate and a sugar belonging to the same nucleotide, whereas S(3')-P represents the bond joining neighboring nucleotides.

Bond	d_0(Å)	Bond Angle	θ_0(degree)	Bond Dihedral	ϕ_0(degree)
S(5')-P	3.899	S(5')-P-(3')S	94.49	P-(5')S(3')-P-(5')S	-154.80
S(3')-P	3.559	P-(5')S(3')-P	120.15	S(3')-P-(5')S(3')-P	-179.17
S-Ab	6.430	P-(5')S-Ab	113.13	Ab-S(3')-P-(5')S	-22.60
S-Tb	4.880	P-(3')S-Ab	108.38	S(3')-P-(5')S-Ab	50.69
S-Cb	4.921	P-(5')S-Tb	102.79	Tb-S(3')-P-(5')S	-33.42
S-Gb	6.392	P-(3')S-Tb	112.72	S(3')-P-(5')S-Tb	54.69
		P-(5')S-Cb	103.49	Cb-S(3')-P-(5')S	-32.72
		P-(3')S-Cb	112.39	S(3')-P-(5')S-Cb	54.50
		P-(5')S-Gb	113.52	Gb-S(3')-P-(5')S	-22.30
		P-(3')S-Gb	108.12	S(3')-P-(5')S-Gb	50.66

Table A.2: Force field parameters for nonbonded interactions and strengths of bonded interactions.

Potential	Parameter	Value(KJ·mol^{-1})	Parameter	Value(Å)
V_{stack}	ε	0.769856	σ_{ij}	Pair-dependent
V_{bp}	ε_{AT}	2.000ε	σ_{AT}	2.9002
	ε_{CG}	2.532ε	σ_{CG}	2.8694
V_{nnat}			σ_0 (mismatch)	$1.00 \cdot 2^{-1/6}$
			σ_0 (otherwise)	$6.86 \cdot 2^{-1/6}$
V_{solv}	ε_s	System-dependent	α^{-1}	5.333
			r_s	13.38
V_{bond}	κ_1	ε		
	κ_2	100ε		
V_{angle}	κ_θ	1400ε		
$V_{dihedral}$	κ_ϕ	28ε		

3SPN model has been applied to explore the dynamics of DNA denaturation, bubbling, bending and hybridization. As a simple example, Figure A.2 shows the number of denatured base pairs of a doulbe strand DNA of 14 bp simulated using 3SPN force field at 300 K (a) and at 350 K (b). While DNA remains double helix structure at 300 K (c-e), it gradually denatured at 350 K

(f-h). More comprehensive comparisons with experimental data are referred to their published articles. It also be noticed that this model is still under vivid updating, from 3SPN.0 [113] to 3SPN.2 [179].

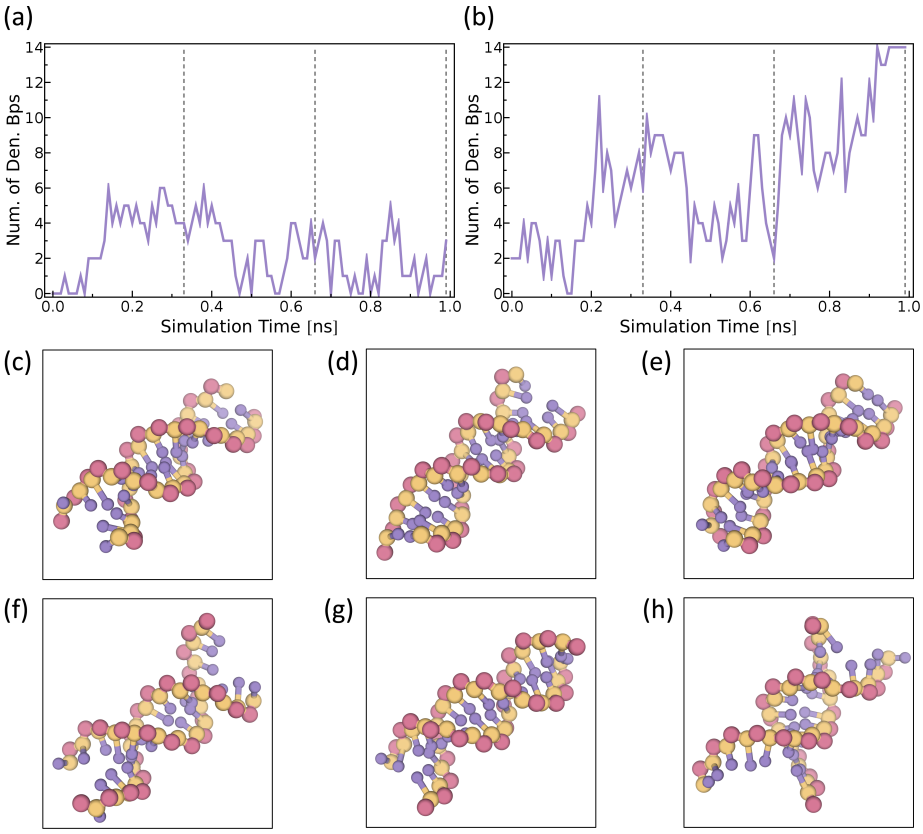


Figure A.2: Double strand DNA of 14 bp simulated at different temperatures using 3SPN.1 model. Number of denatured base pairs versus the simulation time at 300 K (a) and 350 K (b). Conformations of DNA at simulation time $t = \{0.33, 0.66, 0.99\}$ ns with $T = 300$ K (c-e) and $T = 350$ K (f-h), respectively. In (a,b), one base pair is defined denatured if the separation between bases is larger than $(\sigma_{bi} + 2.0)$ Å.

B PeptideB model

Similar to modeling of DNA, the success of atomistic simulations of proteins is limited to available computing power. Coarse-grained models for proteins, on a broad range of length scales, have been developed for many years. For example, the lattice heterogeneous polymer HP model briefly discussed in Section 3.1, and the off-lattice one-bead-per-amino-acid Gō model. They provide many important insights into the protein folding, unfolding, binding to DNA, etc. Another force field of many impressive applications, MARTINI, was proposed by Marrink and his colleagues more recently. On average, four heavy atoms are mapped to one bead except ring-like molecules, and each physiological amino acid is approximately represented by two beads. As initially designed for lipids, MARTINI is parameterized using the partition coefficients between water and a lipid membrane. One constraint underlying Gō model and MARTINI is that the force field biases the simulated protein towards a *pre-defined* native structure or secondary structure. Hence they are not quite suit for studies of proteins with intrisically disordered regions or large domain dynamics (e.g., CTCF). Models of higher resolution could sample more extended phase space of protein structure, and offer a larger probability to succeed in simulating peptide *de novo* folding, aggregation, and proteins of marginal stability. The *peptideB* model,

developed by Bereau and Deserno [87, 180, 181], is our basic mesoscale model for multi-zinc finger proteins.

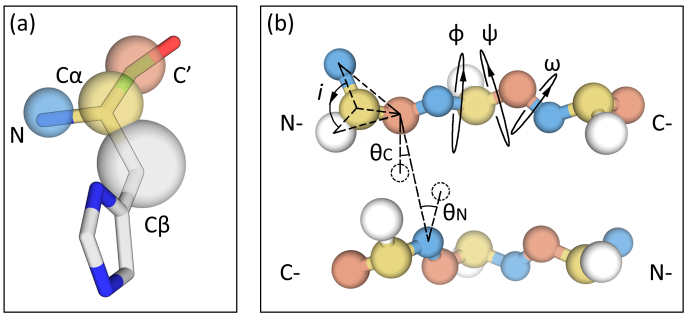


Figure B.1: PeptideB model. (a) One amino acid is constructed using four beads, each representing the amino group N, central carbon C_α, carbonyl group C' and side chain group C_β. (b) Some force field-related angles and dihedral angles. The label i stands for "improper dihedral". Two dashed circles represent two phantom atoms H and O, which is connected to a N and C' bead respectively. Details are explained in the main text.

As shown in Figure B.1(a), in peptideB, one residue is represented by four beads. The backbone is modeled almost atomistically with three beads, each corresponding to the amino group N, central carbon C_α and carbonyl group C'. The side chain group (except glycine) is modeled by a bead located at the position of C_β. The force field is composed of following seven contributions,

$$V_{peptideB} = V_{bond} + V_{angle} + V_{dihedral} + V_{bb} + V_{hp} + V_{hb} + V_{dip}. \tag{B.1}$$

The first three terms are bonded interactions of common expressions,

$$V_{bond} = \frac{1}{2}\kappa_{bond}(d - d_0)^2 \tag{B.2}$$

$$V_{angle} = \frac{1}{2}\kappa_{angle}(\theta - \theta_0)^2 \tag{B.3}$$

$$V_{dihedral} = \kappa_{dihedral}[1 - \cos(\varphi - \varphi_0)]. \tag{B.4}$$

Note that in $V_{dihedral}$ (see also Figure B.1(b)), constraints are only applied to the third backbone dihedral ω ($C_\alpha C' N C_\alpha$) and the improper dihedral ($N C_\alpha C' C_\beta$), which favors *trans* conformations and a local tetrahedron geometry around C_α, respectively. The interaction strengths $\{\kappa_{bond}, \kappa_{angle}, \kappa_{dihedral}\}$ and equilibrium distances and angles $\{d_0, \theta_0, \varphi_0\}$ are listed in Table B.1.

The nonbonded interactions include

$$V_{bb} = \begin{cases} 4\varepsilon_{bb}[(\frac{\sigma_{ij}}{r})^{12} - (\frac{\sigma_{ij}}{r})^{6}] + \varepsilon_{bb} & r \le r_{cut}, \\ 0 & \text{otherwise} \end{cases} \tag{B.5}$$

$$V_{hp} = S_{hp} \times \begin{cases} 4\varepsilon_{hp}[(\frac{\sigma_{C_\beta}}{r})^{12} - (\frac{\sigma_{C_\beta}}{r})^{6}] + \varepsilon_{hp}(1 - \varepsilon'_{ij}) & r \le r_c, \\ 4\varepsilon_{hp}\varepsilon'_{ij}[(\frac{\sigma_{C_\beta}}{r})^{12} - (\frac{\sigma_{C_\beta}}{r})^{6}] & r_c \le r \le r^{hp}_{cut}, \\ 0 & \text{otherwise} \end{cases} \tag{B.6}$$

$$V_{hb} = \varepsilon_{hb}[5(\frac{\sigma_{hb}}{r})^{12} - 6(\frac{\sigma_{hb}}{r})^{10}] \times \begin{cases} \cos^2\theta_N \cos^2\theta_C & |\theta_N|, |\theta_C| < 90^\circ \& r \le r^{hb}_{cut}, \\ 0 & \text{otherwise} \end{cases} \tag{B.7}$$

$$V_{dip} = \kappa_{dip}[(1 - \cos\phi) + (1 - \cos\psi)]. \tag{B.8}$$

V_{bb} describes the excluded volume interactions with a purely repulsive Weeks-Chandler-Andersen (WCA) potential. In Equation B.5, $r_{cut} = 2^{1/6}\sigma_{ij}$, and σ_{ij} is the arithmetic mean of the sizes of two beads involved. It is applied to all backbone-backbone and backbone-side chain bead pairs $i - j$ that are at least three bonds apart.

V_{hp} accounts for the interactions between side chain beads, which are mainly contributed from hydrophobic effect. It is composed of a repulsive WCA potential for small distances and a Lennard-Jones attractive potential at intermediate distances. They are joined at r_c so that both V_{hp} and its first derivative are

continuous. In Equation B.6, ε'_{ij} controlling the strength of attractive potential, obeys the Lorentz-Berthelot mixing rule $\varepsilon'_{ij} = \sqrt{\varepsilon'_i \varepsilon'_j}$, where $\varepsilon'_{i/j}$ are normalized hydrophobicities whose values are listed in Table B.2. The leftmost factor $S_{hp} \in [0, 1]$ is designed to scale the hydrophobic interactions strength depending on the solvent.

V_{hb} describes the hydrogen bond interactions as a function of the separation and relative orientation of the amide and carbonyl groups. In Equation B.7, r is the distance between a N bead and a C' bead, σ_{hb} is an equilibrium distance, θ_N is the angle formed by the atoms HNC' and θ_C is the angle formed by NC'O, where the coordinates of atoms H and O are inferred from backbone beads without explicitly simulation. It favors a geometry of aligned N, H and O atoms (see Figure B.1(b)).

The last item V_{dip} is a crude approximation of the interactions of dipoles formed by bonded carbonyl and amide groups. Taking only the nearest-neighbor interactions into consideration, V_{dip} is determined using the first and second backbone dihedrals, ϕ(C'NC$_\alpha$C') and ψ(NC$_\alpha$C'N) (see also Figure B.1(b)). This contribution is important for balancing the α-helices and β-sheets in the secondary structure contents. Parameters for nonbonded interactions are summarized in Table B.3.

Table B.1: Force field parameters for bonded interactions. For interaction strength constants, $\kappa_{bond} = \kappa_{angle} = 300$, $\kappa_{dihedral} = 67$ for ω, and $\kappa_{dihedral} = 17$ for the improper dihedral, in units of $\varepsilon = k_B T_r = 2.494$ KJ $\cdot$ mol^{-1}.

Bond	d_0(Å)	Bond Angle	θ_0(degree)	Bond Dihedral	φ_0(degree)
NC$_\alpha$	1.455	NC$_\alpha$C$_\beta$	108	ω	180
C$_\alpha$C'	1.510	C$_\beta$C$_\alpha$C'	113	improper	$\mp$120
C'N	1.325	NC$_\alpha$C	111		
C$_\alpha$C$_\beta$	1.530	C$_\alpha$C'N	116		
		C'NC$_\alpha$	122		

Many efforts have been devoted into developing a systematic method to pa-

Table B.2: Normalized hydrophobicities ε_i' for amino acid residue i.

Residue	ε_i'	Residue	ε_i'	Residue	ε_i'	Residue	ε_i'
Lys	0.00	Glu	0.05	Asp	0.06	Asn	0.10
Ser	0.11	Arg	0.13	Gln	0.13	Pro	0.14
Thr	0.16	Gly	0.17	His	0.25	Ala	0.26
Tyr	0.49	Cys	0.54	Trp	0.64	Val	0.65
Met	0.67	Ile	0.84	Phe	0.97	Leu	1.00

Table B.3: Force field parameters for nonbonded interactions. The energy unit $\varepsilon = k_B T_r = 2.494$ KJ $\cdot$ mol^{-1}.

Potential	Parameters	value(ε)	Parameters	Value(Å)
V_{bb}	ε_{bb}	0.02	σ_N	2.9
			σ_{C_α}	3.7
			$\sigma_{C'}$	3.5
			σ_{C_β}	5.0
V_{hp}	ε_{hp}	4.5	σ_{C_β}	5.0
			r_{cut}^{hp}	10
V_{hb}	ε_{hb}	6.0	σ_{hb}	4.11
			r_{cut}^{hb}	8
V_{dip}	ε_{dip}	-0.3		

rameterize a *nice* coarse-grained force field. In general there exist two schemes. The *bottom-up* approach is to derive the parameter values from the simulation results of a model with higher resolution and reliability, like atomistic molecular dynamics simulation and *ab initio* quantum chemistry calculation. The *top-down* approach is to tune the parameter values so that the coarse-grained model can reproduce certain properties of interest of the reference system (also called training set). Following the second scheme, peptideB was parameterized to generate proper local conformations (Ramachandral plot) of tripeptides and to *de nove* fold proteins of native three-helix bundle structure. It also succeeds in aggregating Heptapeptides to form β-sheet.

As an example, Figure B.2 shows the simulated binding of the pepetide KESLV to the syntrophin PDZ domain (PDB code: 2PDZ), which has been studied by Staneva and Wallin using an all-atom Monte Carlo method [182,

183]. The shot peptide is positioned far away from the protein at the beginning of the simulation (b). Once it hits the binding pockect of protein, although with a wrong orientation, it stays there (c). The peptide continually adapts its conformation with the protein until it finds its native binding posture (d). In the subplot (a), the interface nativeness Q quantitatively describes this binding pathway.

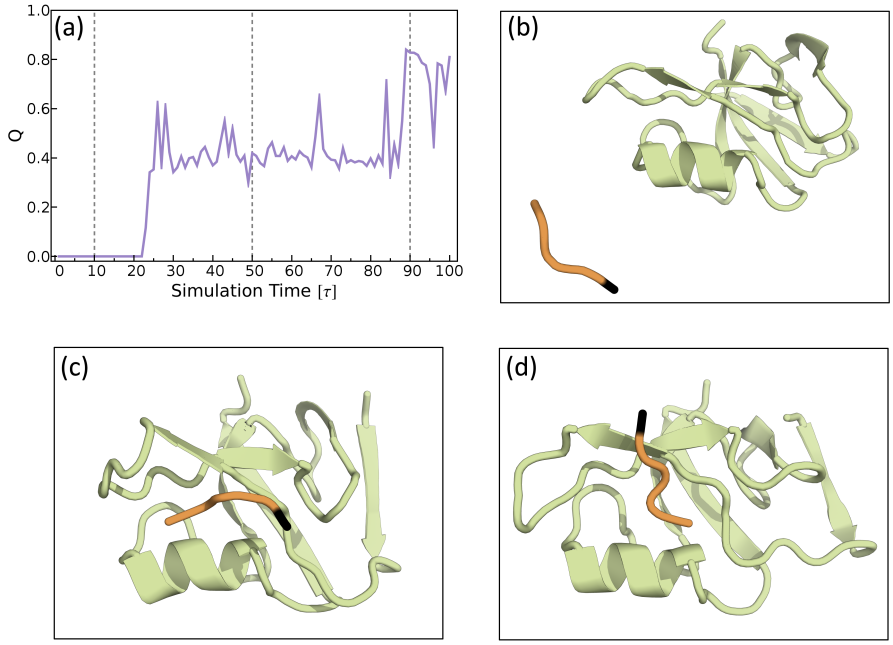


Figure B.2: Simulated binding of the peptide KESLV to the sytrophin PDZ domain. (a) Interface nativeness $Q =< exp[-(r_{ij}^{ref} - r_{ij}^{sim})^2/9] >_{ij}$ versus the simulation time t. Reference structure is the NMR-resolved structure deposited in PDB database (2PDZ). The interface is defined as the pairs of beads, each from the pepetide and from the protein, with a separation less than 7 Å in the reference structure. (b-d) Snapshots at $t = \{10, 50, 90\}\tau$. The protein is colored green, and the peptide is colored orange with a black N-terminus.

C WHAM

The *Weighted Histogram Analysis Method* (WHAM) is a statistical approach to calculate the *Potential of Mean Force* (PMF) from multiple Monte Carlo or melecular dynamics simualtion results. It has been shown to be particularly helpful in parameterizing coarse-grained force field, e.g., 3SPN model and peptideB model.

Here we describe how to determine the PMF from parallel simulations at a series of R inverse temperatures $\beta_i = 1/k_B T_i, i = 1, 2, ..., R$ [92, 93]. The ith simulation produces a histogram $h_i(E)$ composed of N_i records of the internal energy E, i.e., $h_i(E) = \{E_i^1, E_i^2, ..., E_i^{N_i}\}$. According to the WHAM equations, the best estimate for the unnormalized probability of the internal energy E at β_i is

$$P(E, \beta_i) = \frac{\sum_{j=1}^{R} g_j^{-1} h_j(E) e^{-\beta_i E}}{\sum_{j=1}^{R} g_j^{-1} N_j e^{-\beta_j E - f_j}}, \tag{C.1}$$

where $f_j = -\beta_j A_j$ with A_j identical to the (Helmholtz) free energy of the system at β_j, and $g_j = 1 + 2\tau_j$ with τ_j the integrated autocorrelation time for the jth run. And the free energy parameters $\{f_i\}$ are given by the equations

$$f_i = \ln[\sum_E P(E, \beta_i)]. \tag{C.2}$$

For E_i^s, the sth record of E during the ith simulation, notice that $E_i^s \neq E_j^t$ provided $i \neq j$ and $s \neq t$. Then assuming g_i is a constant independent of β_i, C.2 can be formulated as

$$f_i = \ln[\sum_E \frac{1}{\sum_{j=1}^R N_j e^{(\beta_i - \beta_j)E - f_j}}]. \tag{C.3}$$

C.3 is composed of R equations, but there are only $R-1$ independent free energy parameter $\{f_i\}$. In other words, $\{f_i\}$ are determined up to an arbitrary additive constant. Many algorithm, like direct iteration, can be applied to solve these equations.

The next step is to calculate the relative PMF as function of reaction coordinate(s) ξ at a specific inverse temperature β_i. Using the unnormalized probability $P(\xi_k, \beta_i)$ for $\xi_k \in [\xi_k^0 - \Delta\xi, \xi_k^0 - \Delta\xi]$ (with a step $2\Delta\xi$) at T_i

$$P(\xi_k, \beta_i) = \frac{\sum_{j=1}^R g_j^{-1} h_j(\xi_k) e^{-\beta_i E}}{\sum_{j=1}^R g_j^{-1} N_j e^{-\beta_j E - f_j}}. \tag{C.4}$$

the relative $\Delta\text{PMF}(\xi, \beta_i)$ is given by

$$\Delta\text{PMF}(\xi_k, \beta_i) = -\frac{1}{\beta_i} \times \ln[\frac{\sum_E P(\xi_k, \beta_i)}{\max_{i,k}(\sum_E P(\xi_k, \beta_i))}]. \tag{C.5}$$

It should also be noticed that WHAM is a self-consistent way, without any extra parameter, to calculate relative PMF(ξ). The variable ξ can be a vector of $1d$, $2d$ (in chapter 4), $3d$ (in chapter 4), or even higher dimensions.

As an example, Figure C.1 shows an application of WHAM to calculate the relative PMF(ϕ, ψ), where ϕ and ψ are the first and second backbone dihedral of the central residue in tripeptide Gly-Ala-Gly. The simulation was performed using peptideB model at a series of temperatures. During the force field parametrization, Bereau [87] used this heatmap to check whether both

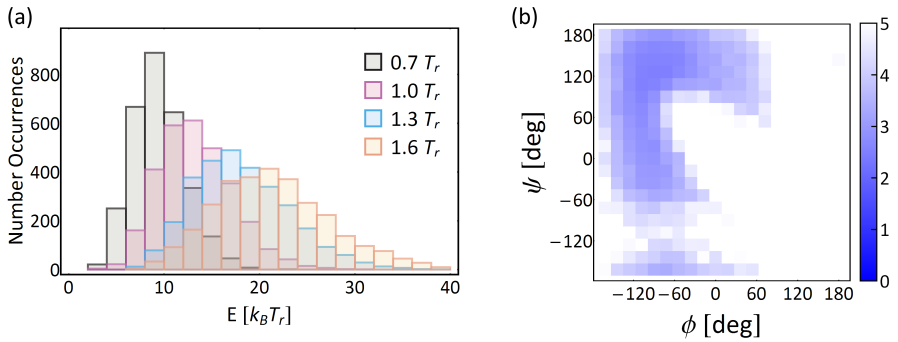


Figure C.1: WHAM applied to coarse-grained simulations of tripeptide Gly-Ala-Gly using peptideB model. (a) Histograms of energies E at $\{0.7, 1.0, 1.3, 1.6\}T_r$. (b) Relative PMF heatmap in units of k_BT_r, compared to the conformation of lowest free energy, as a function of the backbone dihedrals ϕ and ψ in the central residue Alanine.

α-helix (-60°, -60°) and β-sheet (-60°, 130°) are well populated, balanced and connected.

In chapter 4, when the original peptideB model was directly applied to the three-C_2H_2 polypeptides 1tf3_1_3, the first and the third zinc finger quickly aggregated, which is contradictory to the NMR experiment result [75]. To reduce the nonbonded interaction between amino acids, the hydrophobic interaction was scaled by a factor s_{hp}, which should be less than 1. As an application of WHAM in $3d$, proper value of s_{hp} for mC_2H_2 proteins was determined by calculating the standard binding free energy ΔG^0 for a CCHC zinc finger (FOG) binding to a CCCC zinc finger (GATA-1) and comparing the results with experimental value [77].

Based on the structure of the complex GATA-1:FOG in PDB file 1Y0J [77], an initial conformation for a simulation was prepared by assigning a random separation and orientation of FOG relative to GATA-1 without atom clashes. For different s_{hp} values, 50 runs with different initial conformations, each of simulation time 6×10^6 τ, were performed using the same procedue as described

in chapter 4. By aligning the structure of GATA-1 in each snapshot, we obtained 3×10^5 positions $\boldsymbol{r}$ of the zinc ion in FOG. $\boldsymbol{r}$ was taken as a reaction coordinate to calculate the $3d$ PMF $W(\boldsymbol{r})$ for FOG binding to GATA-1. Then using the methodology of Buch et al. [184], ΔG^0 is given by

$$\Delta G^0 = -k_B T \ln (V_b/V_0) - \Delta W. \tag{C.6}$$

V_0 is the standard-state volume (1661 Å^3 for 1 M concentration). The average sampled bound state V_b is an integral over "bound region",

$$V_b = \int_{V_b} \exp(-W(\boldsymbol{r})/k_B T)\, \mathrm{d}\boldsymbol{r}, \tag{C.7}$$

while the "unbound region" is chosen to represent where the $W(\boldsymbol{r})$ is flat in the bulk. ΔW is then the difference between the value of PMF in the bulk and the minimum value in the bulk state, i.e., the depth of PMF.

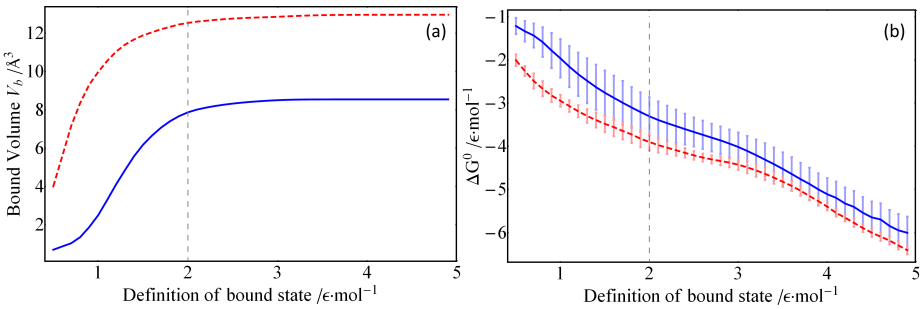


Figure C.2: Dependence of the bound volume V_b (a) and the standard free energy of binding ΔG^0 (b) on the definition of the bound state with the scaling factor s_{hp} of value 0.0 (solid blue line) and of value 1.0 (dashed red line). The errors in (b) are calculated from block average over five sets of data of 6×10^4 τ each.

As Figure C.2 (a) shows, V_b converges as the criterion for bound state increases. If we choose PMF $< 2.0\ \epsilon \cdot \mathrm{mol}^{-1}$ as the definition of bound state, this

procedue yields $\Delta G^0 = -3.31$ for $s_{hp} = 0.0$, and $\Delta G^0 = -3.90$ for $s_{hp} = 1.0$ in units of $\epsilon \cdot \mathrm{mol}^{-1}$. Compared to the experimental value -3.64$\sim$-2.79 $\epsilon \cdot \mathrm{mol}^{-1}$ [77], s_{hp} was set to 0.0 or 0.1.

Bibliography

[1] Dixon J R, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu J S and Ren B 2012 *Nature* **485** 376–380

[2] Zuin J, Dixon J R, van der Reijden M I J A, Ye Z, Kolovos P, Brouwer R W W, van de Corput M P C, van de Werken H J G, Knoch T A, van IJcken W F J, Grosveld F G, Ren B and Wendt K S 2014 *Proceedings of the National Academy of Sciences* **111** 996–1001

[3] Filippova G N, Fagerlie S, Klenova E M, Myers C, Dehner Y, Goodwin G, Neiman P E, Collins S J and Lobanenkov V V 1996 *Molecular and cellular biology* **16** 2802–2813

[4] Phillips J E and Corces V G 2009 *Cell* **137** 1194–1211

[5] Ong C T and Corces V G 2014 *Nat Rev Genet* **15** 234–246

[6] Wikipedia 2007 Hela cells stained with hoechst 33258. [Online; accessed 29-Oct-2014] URL `http://commons.wikimedia.org/wiki/File:HeLa_cells_stained_with_Hoechst_33258.jpg`

[7] MedCell@Yale Euchromatin and heterochromatin. [Online; accessed 29-Oct-2014] URL `http://medcell.med.yale.edu/histology/cell_lab/images/euchromatin_and_heterochromatin.jpg`

[8] Crick F 1970 *Nature* **227** 561–563

[9] Wikipedia 2007 Chromatin structures. [Online; accessed 29-Oct-2014] URL `http://commons.wikimedia.org/wiki/File:Chromatin_Structures.png`

[10] Mateos-Langerak J, Bohn M, de Leeuw W, Giromus O, Manders E M M, Verschure P J, Indemans M H G, Gierman H J, Heermann D W, van Driel R and Goetze S 2009 *Proceedings of the National Academy of Sciences* **106** 3812–3817

[11] Lieberman-Aiden E, van Berkum N L, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie B R, Sabo P J, Dorschner M O, Sandstrom R, Bernstein B, Bender M A, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny L A, Lander E S and Dekker J 2009 *Science* **326** 289–293

[12] Lewin B 2003 *Genes VIII* united states ed ed (Benjamin Cummings) ISBN 0131439812

[13] Zhang Y and Heermann D W 2011 *PLoS ONE* **6** e29225

[14] Zhang Y, Isbaner S and Heermann D W 2013 *Frontiers in Physics* **1** 1–11

[15] Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher M R and Cremer T 2005 *PLoS Biol* **3** e157

[16] Cremer T and Cremer M 2010 *Cold Spring Harbor Perspectives in Biology* **2** 1–22.a003889

[17] Wikipedia 2008 Chromosome conformation capture technology. [Online; accessed 29-Oct-2014] URL `http://commons.wikimedia.org/wiki/File:Chromosome_Conformation_Capture_Technology.jpg`

[18] Farnham P J 2009 *Nat Rev Genet* **10** 605–616

[19] Hakim O and Misteli T 2012 *Cell* **148** 1068 – 1068.e2

[20] Wikipedia 2012 Chromatin immunoprecipitation sequencing. [Online; accessed 29-Oct-2014] URL `http://commons.wikimedia.org/wiki/File:Chromatin_immunoprecipitation_sequencing.svg`

[21] Kim T H H, Abdullaev Z K, Smith A D, Ching K A, Loukinov D I, Green R D, Zhang M Q, Lobanenkov V V and Ren B 2007 *Cell* **128** 1231–1245

[22] Xie X, Mikkelsen T S, Gnirke A, Lindblad-Toh K, Kellis M and Lander E S 2007 *Proceedings of the National Academy of Sciences* **104** 7145–7150

[23] Chen H, Tian Y, Shu W, Bo X and Wang S 2012 *PLoS ONE* **7** e41374+

[24] Bao L, Zhou M and Cui Y 2008 *Nucleic acids research* **36** D83–87

[25] Ziebarth J D, Bhattacharya A and Cui Y 2013 *Nucleic Acids Research* **41** D188–D194

[26] Lobanenkov V V, Nicolas R H, Adler V V, Paterson H, Klenova E M, Polotskaja A V and Goodwin G H 1990 *Oncogene* **5** 1743–1753

[27] Vostrov A A and Quitschke W W 1997 *The Journal of biological chemistry* **272** 33353–33359

[28] Merkenschlager M and Odom D T 2013 *Cell* **152** 1285–1297

[29] Phillips-Cremins J E and Corces V G 2013 *Mol Cell* **50** 461–474

[30] Nativio R, Wendt K S, Ito Y, Huddleston J E, Uribe-Lewis S, Woodfine K, Krueger C, Reik W, Peters J M and Murrell A 2009 *PLoS Genet* **5** e1000739+

[31] Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Singh Sandhu K, Singh U, Pant V, Tiwari V, Kurukuti S and Ohlsson R 2006 *Nat Genet* **38** 1341–1347

[32] Karolchik D, Hinrichs A S, Furey T S, Roskin K M, Sugnet C W, Haussler D and Kent W J 2004 *Nucl. Acids Res.* **32** D493–496

[33] Karolchik D, Barber G P, Casper J, Clawson H, Cline M S, Diekhans M, Dreszer T R, Fujita P A, Guruvadoo L, Haeussler M, Harte R A, Heitner S, Hinrichs A S, Learned K, Lee B T, Li C H, Raney B J, Rhead B, Rosenbloom K R, Sloan C A, Speir M L, Zweig A S, Haussler D, Kuhn R M and Kent W J 2014 *Nucleic Acids Research* **42** D764–D770

[34] Shen Y, Yue F, McCleary D F, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov V V and Ren B 2012 *Nature* **488** 116–120

[35] Barski A, Cuddapah S, Cui K, Roh T Y Y, Schones D E, Wang Z, Wei G, Chepelev I and Zhao K 2007 *Cell* **129** 823–837

[36] Handoko L, Xu H, Li G, Ngan C Y Y, Chew E, Schnapp M, Lee C W H W, Ye C, Ping J L H L, Mulawadi F, Wong E, Sheng J, Zhang Y, Poh T, Chan C S S, Kunarso G, Shahab A, Bourque G, Cacheux-Rataboul V, Sung W K K, Ruan Y and Wei C L L 2011 *Nature genetics* **43** 630–638

[37] Dowen J M, Fan Z P, Hnisz D, Ren G, Abraham B J, Zhang L N, Weintraub A S, Schuijers J, Lee T I, Zhao K and Young R A 2014 *Cell* **159** 374–387

[38] Nakahashi H, Kwon K R, Resch W, Vian L, Dose M, Stavreva D, Hakim O, Pruett N, Nelson S, Yamane A, Qian J, Dubois W, Welsh S, Phair R, Pugh B, Lobanenkov V, Hager G and Casellas R 2013 *Cell Reports* **3** 1678 – 1689

[39] Ohlsson R 2001 *Trends in Genetics* **17** 520–527

[40] Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, Felsenfeld G and Pedone P V 2007 *Journal of Biological Chemistry* **282** 33336–33345

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

[42] Dyson H J and Wright P E 2005 *Nat Rev Mol Cell Biol* **6** 197–208

[43] Laity J H, Dyson H and Wright P E 2000 *Journal of Molecular Biology* **295** 719 – 727

[44] Schneider R, Huang J r, Yao M, Communie G, Ozenne V, Mollica L, Salmon L, Ringkjobing Jensen M and Blackledge M 2012 *Mol. BioSyst.* **8**(1) 58–68

[45] Bernado P and Svergun D I 2012 *Mol. BioSyst.* **8** 151–167

[46] Receveur V, Czjzek M, Schülein M, Panine P and Henrissat B 2002 *Journal of Biological Chemistry* **277** 40887–40892

[47] Schuler B, Lipman E A, Steinbach P J, Kumke M and Eaton W A 2005 *Proceedings of the National Academy of Sciences of the United States of America* **102** 2754–2759

[48] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *The Journal of Chemical Physics* **21** 1087–1092

[49] Madras N and Sokal A 1988 *Journal of Statistical Physics* **50** 109–186

[50] Binder K and Heermann D W 2002 *Monte Carlo Simulation in Statistical Physics* 4th ed (Springer) ISBN 3540432213

[51] Rovere M, Heermann D W and Binder K 1990 *Journal of Physics: Condensed Matter* **2** 7009+

[52] Chang M 2010 *Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms, and Case Studies* (CRC Press) ISBN 978-1-4398-3592-0

[53] Aichinger M and Binder A 2013 *A Workout in Computational Finance* (John Wiley & Sons, Ltd) ISBN 9781119973515

[54] Flory P J 1949 *The Journal of Chemical Physics* **17** 303–310

[55] Bohn M, Heermann D W and van Driel R 2007 *Physical Review E* **76** 051805+

[56] Mann M, Will S and Backofen R 2008 *BMC Bioinformatics* **9** 230

[57] Mann M, Smith C, Rabbath M, Edwards M, Will S and Backofen R 2009 *Bioinformatics* **25** 676–677

[58] Lau K F and Dill K A 1989 *Macromolecules* **22** 3986–3997

[59] Taketomi H, Ueda Y and Gō N 1975 *Int J Pept Protein Res* **7** 445–459

[60] Koga N and Takada S 2001 *Journal of Molecular Biology* **313** 171 – 180

[61] Karanicolas J and Brooks Iii C L 2003 *Journal of Molecular Biology* **334** 309–325

[62] Carmesin I and Kremer K 1988 *Macromolecules* **21** 2819–2823

[63] Deutsch H P and Binder K 1991 *The Journal of Chemical Physics* **94** 2294–2304

[64] Kampmann T A, Boltz H H and Kierfeld J 2013 *The Journal of Chemical Physics* **139** 034903

[65] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark A E and Berendsen H J C 2005 *J. Comput. Chem.* **26** 1701–1718

[66] Limbach H, Arnold A, Mann B and Holm C 2006 *Computer Physics Communications* **174** 704 – 727

[67] Arnold A, Lenz O, Kesselheim S, Weeber R, Fahrenberger F, Roehm D, KoÅąovan P and Holm C 2013 Espresso 3.1: Molecular dynamics software for coarse-grained models. *Meshfree Methods for Partial Differential Equations VI* (*Lecture Notes in Computational Science and Engineering* vol 89) ed Griebel M and Schweitzer M A (Springer Berlin Heidelberg) pp 1–23

[68] Berendsen H J C, Postma J P M, van Gunsteren W F, Dinola A and Haak J R 1984 *Journal of Chemical Physics* **81** 3684–3690

[69] Frenkel D and Smit B 2002 *Understanding Molecular Simulation* 2nd ed (Academic Press) ISBN 978-0-12-267351-1

[70] Hünenberger P 2005 Thermostat algorithms for molecular dynamics simulations *Advanced Computer Simulation* (*Advances in Polymer Science* vol 173) ed Dr Holm C and Prof Dr Kremer K (Springer Berlin Heidelberg) pp 105–149 ISBN 978-3-540-22058-9

[71] Laity J H 2006 *Cys2His2 Zinc Finger Proteins* (John Wiley & Sons, Ltd) chap Cys2His2 Zinc Finger Proteins, pp 1–17 ISBN 9780470028636

[72] Iuchi S 2001 *Cellular and molecular life sciences : CMLS* **58** 625–635

[73] Dyson H J and Wright P E 2004 *Chemical Reviews* **104** 3607–3622

[74] Hyre D E and Klevit R E 1998 *Journal of Molecular Biology* **279** 929 – 943

[75] Brueschweiler R, Liao X and Wright P 1995 *Science* **268** 886–889

[76] Imanishi M, Imamura C, Higashi C, Yan W, Negi S, Futaki S and Sugiura Y 2010 *Biochemical and Biophysical Research Communications* **400** 625 – 630

[77] Liew C K, Simpson R J Y, Kwan A H Y, Crofts L A, Loughlin F E, Matthews J M, Crossley M and Mackay J P 2005 *Proceedings of the National Academy of Sciences of the United States of America* **102** 583–588

[78] Mukhopadhyay R, Yu W, Whitehead J, Xu J, Lezcano M, Pack S, Kanduri C, Kanduri M, Ginjala V, Vostrov A, Quitschke W, Chernukhin I, Klenova E, Lobanenkov V and Ohlsson R 2004 *Genome Research* **14** 1594–1602

[79] Jerabek H and Heermann D W 2012 *PLoS ONE* **7** e37525

[80] Linding R, Jensen L J J, Diella F, Bork P, Gibson T J and Russell R B 2003 *Structure (London, England : 1993)* **11** 1453–1459

[81] Ward J J, Mcguffin L J, Bryson K, Buxton B F and Jones D T 2004 *Bioinformatics* **20** 2138–2139

[82] Linding R, Russell R B, Neduva V and Gibson T J 2003 *Nucleic Acids Research* **31** 3701–3708

[83] Martí-Renom M A, Stuart A C, Fiser A, Sánchez R, Melo F and Sali A 2000 *Annual review of biophysics and biomolecular structure* **29** 291–325

[84] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A and Simmerling C 2006 *Proteins: Structure, Function, and Bioinformatics* **65** 712–725

[85] Pang Y P 2001 *Proteins* **45** 183–189

[86] Madras N and Sokal A 1988 *Journal of Statistical Physics* **50** 109–186

[87] Bereau T and Deserno M 2009 *Biophysical Journal* **96** 405a

[88] Prieto L, de Sancho D and Rey A 2005 *The Journal of Chemical Physics* **123** 154903+

[89] Zandarashvili L, Vuzman D, Esadze A, Takayama Y, Sahu D, Levy Y and Iwahara J 2012 *Proceedings of the National Academy of Sciences* **109** E1724–E1732

[90] Monticelli L, Kandasamy S K, Periole X, Larson R G, Tieleman D P and Marrink S J 2008 *J. Chem. Theory Comput.* **4** 819–834

[91] Takada S, Luthey-Schulten Z and Wolynes P G 1999 *The Journal of Chemical Physics* **110** 11616–11629

[92] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 *Journal of Computational Chemistry* **13** 1011–1021

[93] Bereau T and Swendsen R H 2009 *Journal of Computational Physics* **228** 6119–6129

[94] Seeber M, Cecchini M, Rao F, Settanni G and Caflisch A 2007 *Bioinformatics* **23** 2625–2627

[95] Ryan R F and Darby M K 1998 *Nucleic Acids Research* **26** 703–709

[96] Wilhelm J and Frey E 1996 *Physical Review Letters* **77** 2581–2584

[97] Hou C, Dale R and Dean A 2010 *Proceedings of the National Academy of Sciences* **107** 3651–3656

[98] Li Y, Huang W, Niu L, Umbach D M, Covo S and Li L 2013 *BMC genomics* **14** 553+

[99] Bell A C, West A G and Felsenfeld G 1999 *Cell* **98** 387–396

[100] Bell A C and Felsenfeld G 2000 *Nature* **405** 482–485

[101] Wendt K S and Peters J M M 2009 *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **17** 201–214

[102] Marshall A D, Bailey C G and Rasko J E 2014 *Current Opinion in Genetics & Development* **24** 8 – 15

[103] Papworth M, Kolasinska P and Minczuk M 2006 *Gene* **366** 27–38

[104] Razin S V, Borunova V V, Maksimenko O G and Kantidze O L 2012 *Biochemistry (Moscow)* **77** 217–226

[105] Pabo C O, Peisach E and Grant R A 2001 *Annual review of biochemistry* **70** 313–340

[106] Pavletich N P and Pabo C O 1991 *Science* **252** 809–817

[107] Elrod-Erickson M, Rould M A, Nekludova L and Pabo C O 1996 *Structure* **4** 1171 – 1180

[108] Wolfe S A, Grant R A, Elrod-Erickson M and Pabo C O 2001 *Structure (London, England : 1993)* **9** 717–723

[109] Nolte R T, Conlin R M, Harrison S C and Brown R S 1998 *Proceedings of the National Academy of Sciences* **95** 2938–2943

[110] Persikov A V, Osada R and Singh M 2009 *Bioinformatics (Oxford, England)* **25** 22–29

[111] Persikov A V and Singh M 2014 *Nucleic Acids Research* **42** 97–108

[112] Feinauer C J, Hofmann A, Goldt S, Liu L, Máté G and Heermann D W 2013 Zinc finger proteins and the 3d organization of chromosomes. *Organisation of Chromosomes* (*Advances in Protein Chemistry and Structural Biology* vol 90) ed Donev R (Academic Press) pp 67 – 117

[113] Knotts T A, IV, Rathore N, Schwartz D C and de Pablo J J 2007 *The Journal of Chemical Physics* **126** 084901+

[114] Sambriski E J, Schwartz D C and de Pablo J J 2009 *Biophysical Journal* **96** 1675–1690

[115] Bernstein F, Koetzle T, Williams G, Meyerjr E, Brice M, Rodgers J, Kennard O, Shimanouchi T and Tasumi M 1978 *Archives of Biochemistry and Biophysics* **185** 584–591

[116] Setny P and Zacharias M 2011 *Nucleic acids research* **39** 9118–9129

[117] Setny P, Bahadur R and Zacharias M 2012 *BMC Bioinformatics* **13** 228

[118] Slutsky M and Mirny L A 2004 *Biophysical journal* **87** 4021–4035

[119] Marcovitz A and Levy Y 2013 *The Journal of Physical Chemistry B* **117** 13005–13014

[120] Givaty O and Levy Y 2009 *Journal of Molecular Biology* **385** 1087–1097

[121] Terakawa T, Kenzaki H and Takada S 2012 *Journal of the American Chemical Society* **134** 14555–14562

[122] Arnold R, Burcin M, Kaiser B, Muller M and Renkawitz R 1996 *Nucleic acids research* **24** 2640–2647

[123] MacPherson M J and Sadowski P D 2010 *BMC molecular biology* **11** 101+

[124] Mullinax J W and Noid W G 2009 *The Journal of Chemical Physics* **131** 104110

[125] Mullinax J W and Noid W G 2010 *Proceedings of the National Academy of Sciences* **107** 19867–19872

[126] Radivojac P, Iakoucheva L M, Oldfield C J, Obradovic Z, Uversky V N and Dunker A K 2007 *Biophysical journal* **92** 1439–1456

[127] Uversky V N 2002 *Protein Science* **11** 739–756

[128] Uversky V N and Dunker A K 2010 *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1804** 1231–1264

[129] Sugase K, Dyson H J and Wright P E 2007 *Nature* **447** 1021–1025

[130] Dyson H 2002 *Current Opinion in Structural Biology* **12** 54–60

[131] Zhou H X X and Gilson M K 2009 *Chemical reviews* **109** 4092–4107

[132] Privalov P L, Dragan A I, Crane-Robinson C, Breslauer K J, Remeta D P and Minetti C a A S A 2007 *Journal of Molecular Biology* **365** 1–9

[133] Baskaran D, Mays J W and Bratcher M S 2005 *Chemistry of Materials* **17** 3389–3397

[134] Ramasubramaniam R, Chen J and Liu H 2003 *Applied Physics Letters* **83** 2928

[135] Numata M, Asai M, Kaneko K, Bae A H, Hasegawa T, Sakurai K and Shinkai S 2005 *Journal of the American Chemical Society* **127** 5875–5884

[136] Naito M, Nobusawa K, Onouchi H, Nakamura M, Yasui K i, Ikeda A and Fujiki M 2008 *Journal of the American Chemical Society* **130** 16697–16703

[137] Tu X, Manohar S, Jagota A and Zheng M 2009 *Nature* **460** 250–253

[138] Gigliotti B, Sakizzie B, Bethune D S, Shelby R M and Cha J N 2006 *Nano Letters* **6** 159–164

[139] Nish A, Hwang J Y, Doig J and Nicholas R J 2007 *Nature nanotechnology* **2** 640–646

[140] Zhao Y L and Stoddart J F 2009 *Accounts of chemical research* **42** 1161–1171

[141] Star A, Tu E, Niemann J, Gabriel J C P, Joiner S C and Valcke C 2006 *Proceedings of the National Academy of Sciences* **103** 921–926

[142] Gurevitch I and Srebnik S 2008 *The Journal of chemical physics* **128** 144901

[143] Johnson R R, Johnson A T C and Klein M L 2008 *Nano Lett.* **8** 69–75

[144] Roxbury D, Mittal J and Jagota A 2012 *Nano letters* **12** 1464–1469

[145] Vuzman D, Azia A and Levy Y 2010 *Journal of Molecular Biology* **396** 674–684

[146] Tallury S S and Pasquinelli M A 2010 *The Journal of Physical Chemistry B* **114** 4122–4129

[147] Tallury S S and Pasquinelli M A 2010 *The Journal of Physical Chemistry B* **114** 9349–9355

[148] Rubin R J 1966 *The Journal of Chemical Physics* **44** 2130–2138

[149] Rubin R J 1965 *The Journal of Chemical Physics* **43** 2392–2407

[150] Hanke A 2005 *Journal of Physics: Condensed Matter* **17** S1731

[151] Luo M b 2008 *The Journal of Chemical Physics* **128** 044912+

[152] Eisenriegler E, Kremer K and Binder K 1982 *The Journal of Chemical Physics* **77** 6296–6320

[153] Descas R, Sommer J U and Blumen A 2004 *The Journal of Chemical Physics* **120** 8831–8840

[154] Hegger R and Grassberger P 1994 *Journal of Physics A: Mathematical and General* **27** 4069+

[155] Metzger S, Mueller M, Binder K and Baschnagel J 2003 *The Journal of Chemical Physics* **118** 8489–8499

[156] De Gennes P G 1981 *Macromolecules* **14** 1637–1644

[157] Bhattacharya S, Rostiashvili V G, Milchev A and Vilgis T A 2009 *Phys. Rev. E* **79**(3) 030802

[158] Kusner I and Srebnik S 2006 *Chemical Physics Letters* **430** 84 – 88

[159] Moghaddam M S, Vrbová T and Whittington S G 2000 *Journal of Physics A: Mathematical and General* **33** 4573

[160] Polotsky A A 2012 *Journal of Physics A: Mathematical and Theoretical* **45** 425004

[161] Maupetit J, Tuffery P and Derreumaux P 2007 *Proteins: Structure, Function, and Bioinformatics* **69** 394–408

[162] Cooke I R, Kremer K and Deserno M 2005 *Phys Rev E Stat Nonlin Soft Matter Phys* **72** 011506+

[163] Wang Z J and Deserno M 2010 *The Journal of Physical Chemistry B* **114** 11207–11220

[164] Vuzman D, Polonsky M and Levy Y 2010 *Biophys J* **99** 1202–1211

[165] Globisch C, Krishnamani V, Deserno M and Peter C 2013 *PLoS ONE* **8** e60582

[166] Potestio R, Peter C and Kremer K 2014 *Entropy* **16** 4199–4245

[167] Rühle V, Junghans C, Lukyanov A, Kremer K and Andrienko D 2009 *J. Chem. Theory Comput.* **5** 3211–3223

[168] Rühle V and Junghans C 2011 *Macromolecular Theory and Simulations* **20** 472–477

[169] Izvekov S and Voth G A 2005 *The Journal of Physical Chemistry B* **109** 2469–2473

[170] Noid W G, Liu P, Wang Y, Chu J W, Ayton G S, Izvekov S, Andersen H C and Voth G A 2008 *The Journal of Chemical Physics* **128** 244115+

[171] Mullinax J W and Noid W G 2009 *Phys. Rev. Lett.* **103**(19) 198104

[172] Mullinax J W and Noid W G 2010 *J. Phys. Chem. C* **114** 5661–5674

[173] Tschöp W, Kremer K, Batoulis J, Bürger T and Hahn O 1998 *Acta Polymerica* **49** 61–74

[174] Reith D, Pütz M and Müller-Plathe F 2003 *Journal of computational chemistry* **24** 1624–1636

[175] Lyubartsev A P and Laaksonen A 1995 *Physical Review E* **52** 3730–3737

[176] Shell M S 2008 *The Journal of Chemical Physics* **129** 144108+

[177] Chaimovich A and Shell M S 2009 *Phys. Chem. Chem. Phys.* **11** 1901–1915

[178] Chaimovich A and Shell M S 2010 *Physical Review E* **81** 060104+

[179] Hinckley D M, Freeman G S, Whitmer J K and de Pablo J J 2013 *The Journal of Chemical Physics* **139** 144903

[180] Bereau T, Wang Z J and Deserno M 2014 *The Journal of Chemical Physics* **140** 115101

[181] Bereau T and Deserno M 2014 *The Journal of Membrane Biology* 1–11

[182] Staneva I and Wallin S 2009 *Journal of Molecular Biology* **393** 1118 – 1128

[183] Staneva I and Wallin S 2011 *PLoS Comput Biol* **7** e1002131

[184] Buch I, Giorgino T and De Fabritiis G 2011 *Proceedings of the National Academy of Sciences* **108** 10184–10189