Computational modeling and automation techniques to study biomolecular dynamics

Shantanu Sharma

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry and Biophysics.

Chapel Hill
2009

Approved by:

Prof. Nikolay V. Dokholyan, Advisor

Prof. Brian D. Strahl, Committee Chair

Prof. Aziz Sancar, Reader

Prof. Jason D. Lieb, Reader

Prof. Garegin A. Papoian, Reader

Abstract

Shantanu Sharma: Computational modeling and automation techniques to study biomolecular dynamics.
(Under the guidance of Prof. Nikolay V. Dokholyan.)

Physically-principled computational modeling and automation techniques have emerged as potent methodologies in exploring biomolecular dynamics and generating experimentally-testable hypotheses. In this dissertation, we develop a set of simulation automation techniques and present results on case studies of biomolecular simulation. Nucleosomes form the fundamental building blocks of eukaryotic chromatin. We use multiscale modeling and discrete molecular dynamics simulations to investigate the dynamics of the Xenopus laevis nucleosome core particle, the fundamental unit of chromatin. Histone tails are flexible and are poorly resolved in X-ray crystal structures. We probe how molecular-level dynamics of the histone tails, core histones and associated DNA mediate chromatin stability at the scale of single-nucleosomes. Based on the positional fluctuations of core histone residues, we postulate cold sites, a set of core histone residues essential for stabilizing the Xenopus laevis nucleosome core particle. We explore changes in the biophysical stability of mono-nucleosomes by designing mutations in core histones and using Medusa, a high-throughput computational technique to explore changes in mononucleosomal stability resulting from point mutations. The presence of centromere-specific H3 variant histone (Cse4) in centromere-specific nucleosomes defines the kinetochore locus. However, structural details of the centromere-specific nucleosomes remain to be completely understood. We construct a homology model of the Saccharomyces cerevisiae centromeric nucleosome and generate a biophysically-principled C-loop model for elongation of Saccharomyces cerevisiae kinetochore. We present simulation automation techniques by means of two web-based servers: iFold (http://iFold.dokhlab.org) and iFoldRNA (http://iFoldRNA.dokhlab.org). iFold enables automated simulations of protein folding, unfolding using discrete molecular dynamics. iFoldRNA enables ab initio RNA structure prediction using replica-exchange discrete molecular dynamics simulations. We also demonstrate

rapid and accurate three-dimensional structure prediction of over 150 RNA molecules. We used all-atom molecular dynamics simulations to study the mechanistic and structural differences between two anticancer therapeutics - cisplatin and oxaliplatin. Our simulations suggest that the cisplatinated- and oxaliplatinated- DNA cause differential effects on the dynamics and bending propensities of adducted DNA. This study suggest a role of differential bending propensities in the efficacies of oxaliplatin and cisplatin. In summary, the research presented in this dissertation helps us understand the mechanisms of biomolecular interactions at atomic and mesoscale levels. This dissertation adds to scientific knowledge by a set of methodologies for exploring the dynamics of protein and RNA molecules. Physically-principled simulations of the nucleosome core particle yield experimentally-testable hypotheses on chromatin structure and function.

Dedicated to my parents, my Professors, my wonderful colleagues and all other friends at UNC.

Acknowledgments

I express my sincerest gratitute to the entire University of North Carolina Biochemistry faculty members and my colleagues in the department. My professors and colleagues helped me in some of the most difficult moments of my doctoral research. I am most indebted to my thesis advisor Prof. Nikolay V. Dokholyan, for his supervision and guidance in all my research endeavors. I sincerely thank Prof. Brian Strahl, Prof. Stephen Chaney, Prof. Diane Pozefsky, Prof. Kerry Bloom, Prof. Brenda Temple and Prof. Feng Ding for their support in my research collaborations. I owe my sincere thanks to my dissertation committee members, Prof. Aziz Sancar, Prof. Brian Strahl, Prof. Jason Lieb, Prof. Garegin Papoian, for their guidance throughout my doctoral research. I am grateful to Prof. Feng Ding for his relentless support and guidance over the years. I thank Sagar, Yiwen, Shuangye, Tamas, Huifen, Kyle, Barry, Peng, Adrian, and all other members of the Dokholyan Group whose names I have missed. I thank the UNC Information Technology Services Team, especially Ruth and Steven for their support. I owe my special thanks to Kyle for graciously agreeing to share the dinner with me at the Carolina Brewery on Dec 4th, 2008. I would also like to acknowledge the support of my undergraduate thesis advisor, Prof. Somenath Biswas, and my undergraduate research mentors, Dr. Ramanathan Sowdhamini and Dr. Narendra Karmarkar for motivating me to pursue academic research. Finally, I would like to acknowledge the resolute support of my family throughout my research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1. 2D-PMF: Two-Dimensional Potential of Mean Force
2. 3D: Three Dimensional
3. AT: Automation Techniques
4. CBDCA: Cis-diammine-1,1-rcyclobutanedicarboxylatoplatinum(II)
5. CG: Coarse Grained
6. CM: Computational Modeling
7. CP: Cis-diamminedichloroplatinum(II)
8. CP-DNA: Cisplatinated DNA
9. DMD: Discrete Molecular Dynamics
10. DNA: Deoxyribonucleic Nucleic Acid
11. MC: Monte Carlo
12. MM: Multiscale Modeling
13. N: Native Contacts
14. NCM: Normalized Covariance Matrix
15. NCP: Nucleosome Core Particle
16. NDB: Nucleic-acid Data Bank
17. NN: Near-native Contacts
18. OX: Trans-R,R-1,2-diaminocyclohexaneoxalatoplatinum(II)
19. OX-DNA: Oxaliplatinated DNA

20. PCA: Principal Component Analysis

21. PDB: Protein Data Bank

22. PT: Parallel Tempering

23. RNA: Riboxy Nucleic Acid

24. RX: Replica Exchange

25. tu: Time Units

List of Symbols

1. ΔG_{sb} Free energy of salt bridge interactions

2. ε DMD unit energy, Hydrogen-bond potential energy

3. ε/k_B DMD unit temperature

4. k_B Boltzmann constant

Chapter 1

Introduction

1.1 Computational Biology

Proteins, Deoxyribonucleic Nucleic Acids (DNAs) and Riboxy Nucleic Acids (RNAs) constitute the fundamental biological macromolecules. One of the most fundamental constraints in investigating biology at the molecular levels is that of length and time scales at which biomolecular phenomena can be probed. Numerous in vivo and in vitro assays are designed to elucidate mechanistic insights to biological phenomena at molecular details. Physically-principled computational modeling (CM) and automation techniques (AT) have emerged as potent methodologies to explore the atomic-scale dynamics of fundamental biomolecules. Biophysical simulation techniques facilitate research using in vivo and in vitro experiments by providing experimentally-testable hypotheses. In this dissertation, we develop a set of computational modeling and automation techniques applied towards generating experiemntally-testable hypotheses. In the following chapters, we discuss several case studies of physically-based simulations. Multiscale modeling (MM) techniques are used to explore biomolecular dynamics at mesoscale levels.

1.2 Case studies on computational modeling

1.2.1 Structure and dynamics of eukaryoric chromatin

Nucleosomes form the fundamental building blocks of eukaryotic chromatin. Subtle modifications of constituent histones tails mediate chromatin stability and regulate gene expression. Although the structure of the nucleosome core particle is identified at atomic resolution, even the most fundamental knowledge regarding effects of histone variants on chromatin organization and stability remains controversial. A structural understanding of gene activation and aberrations necessitates probing how arrays of nucleosomes organize the formation of higher order structures. Chapter 2 focuses on investigating the dynamics of fundamental unit of eukaryotic chromatin, the nucleosome core particle using multiscale computational modeling and discrete molecular dynamics. We report a multiscale model of the Xenopus laevis nucleosome core particle, using a simplified model for rapid discrete molecular dynamics simulations and an all-atom model for detailed structural investigation. Using the simplified structural model, we perform equilibrium simulations of a single nucleosome at various temperatures. We further reconstruct all-atom nucleosome structures from simulation trajectories.

We find that histone tails bind to nucleosomal DNA via strong salt-bridge interactions over a wide range of temperatures, suggesting a mechanism of chromatin structural organization whereby histone tails regulate inter- and intranucleosomal assemblies via binding with nucleosomal DNA. We identify specific regions of the histone core H2A/H2B-H4/H3-H3/H4-H2B/H2A, termed "cold sites", which retain a significant fraction of contacts with adjoining residues throughout the simulation, indicating their functional role in nucleosome organization. Cold sites are clustered around H3-H3, H2A-H4 and H4-H2A interhistone interfaces, indicating the necessity of these contacts for nucleosome stability. Essential dynamics analysis of simulation trajectories shows that bending across the H3-H3 is a prominent mode of intranucleosomal dynamics. We

postulate that effects of salts on mononucleosomes can be modeled in discrete molecular dynamics by modulating histone-DNA interaction potentials. Local fluctuations in nucleosomal DNA vary significantly along the DNA sequence, suggesting that only a fraction of histone-DNA contacts make strong interactions dominating mononucleosomal dynamics. Our findings suggest that histone tails have a direct functional role in stabilizing higher-order chromatin structure, mediated by salt-bridge interactions with adjacent DNA.

The kinetochore is the protein-DNA complex at eukaryotic centromeres that functions as the attachment site for spindle microtubules. In Saccharomyces cerevisiae, the centromere spans 120 bp, there is a single microtubule per kinetochore, and the entire spindle is composed of 16 kinetochore microtubules plus four interpolar microtubules from each pole. A structural model of the Saccaromyces Cerevisiae centromeric nucleosome, termed C-loop, is developed in Chapter 4. In chapter 3 we use a high-throughput computational protocol to test the role of histone mutagenesis on the stability of eukaryotic nucleosomes.

1.2.2 Automation of protein folding and unfolding simulations

Discrete molecular dynamics simulations enable rapid exploration of protein conformational dynamics. The iFold server enables a web-based service for automated simulation of protein dynamics using discrete molecular dynamics. The service is freely accessible to the scientific community at http://iFold.dokhlab.org. iFold supports long timescale simulations of protein folding, thermal denaturation, thermodynamic scan, simulated annealing and p_{fold} analysis. In iFold simulations, proteins are modeled as coarse-grained two-bead/residue model with structure-based Gō-interactions between amino acids. Chapter 6 presents the research on automation of protein folding and unfolding simulations using the iFold server.

1.2.3 Automation of RNA tertiary structure prediction and folding thermodynamics

The iFoldRNA server (http://iFoldRNA.dokhlab.org) automates rapid tertiary structure prediction and probing folding thermodynamics of RNA molecules. Replica-exchange discrete molecular dynamcis simulations are used to predict RNA structure using a simplified three-bead/nucleotide model. The predicted RNA structures have $\leq$ 5Å RMSD from the experimentally observed structures. Thermodynamic analyses of RNA molecules can be performed using the Folding Thermodynamics module of the iFoldRNA server. Chapter 5 discusses folding of RNA molecules using replica-exchange discrete molecular dynamisc simulations. Chapter 6 presents the research on ab initio RNA structure prediction and folding thermodynamic analyses using the iFoldRNA server.

1.2.4 Molecular dynamics simulation of DNA adducted with Platinum-based therapeutics

The two Platinum-based therapeutics cisplatin and oxaliplatin are among the most efficacious anticancer drugs. However, the precise molecular level mechanism of tissue specificity and of cisplatin vs. oxaliplatin remains to be completely understood. Mismatch repair proteins, DNA damage-recognition proteins and translesion DNA polymerases discriminate between Pt-GG adducts containing cis-diammine ligands (formed by cisplatin (CP) and carboplatin) and trans-RR-diaminocyclohexane ligands (formed by oxaliplatin (OX)) and this discrimination is thought to be important in determining differences in the efficacy, toxicity and mutagenicity of these platinum anticancer agents. In chapter 7, we present all-atom molecular dynamics simulations of a dodecamer DNA adducted with cisplatin and oxaliplatin. Our simulations yield mechanistic hypotheses regarding the function and tissue-specificity of the two anticancer drugs. We postulate that these proteins recognize differences in conformation and/or confor-

mational dynamics of the DNA containing the adducts. We previously determined the NMR solution structure of OX-DNA, CP-DNA and undamaged duplex DNA in the 5'-d(CCTCAGGCCTCC)-3' sequence context and have shown the existence of several conformational differences in the vicinity of the Pt-GG adduct. Here we have used molecular dynamics simulations to explore differences in the conformational dynamics between OX-DNA, CP-DNA and undamaged DNA in the same sequence context. Twenty-five 10 ns unrestrained fully solvated molecular dynamics simulations were performed starting from two different DNA conformations using AMBER v8.0. All 25 simulations reached equilibrium within 4 ns, were independent of the starting structure and were in close agreement with previous crystal and NMR structures. Our data show that the cis-diammine (CP) ligand preferentially forms hydrogen bonds on the 5' side of the Pt-GG adduct, while the trans-RR-diaminocyclohexane (OX) ligand preferentially forms hydrogen bonds on the 3' side of the adduct. In addition, our data show that these differences in hydrogen bond formation are strongly correlated with differences in conformational dynamics, specifically the fraction of time spent in different DNA conformations in the vicinity of the adduct, for CP- and OX-DNA adducts. We postulate that differential recognition of CP- and OX-GG adducts by mismatch repair proteins, DNA damage-recognition proteins and DNA polymerases may be due, in part, to differences in the fraction of time that the adducts spend in a conformation favorable for protein binding.

1.3 Summary

In chapter 8 we present the conclusion of this dissertation. Collectively, the research presented in this dissertation provides several case studies on applications of computational modeling and automation techniques to explore biomolecular interaction and dynamics at atomic levels and generate experimentally-testable hypotheses.

Chapter 2

Multiscale modeling of nucleosome dynamics

2.1 Introduction

Most eukaryotic DNA is associated with histone proteins to form highly compact structures. This complex of DNA and histones, called nucleosomes (Kor74), forms the fundamental repeating subunit of chromatin. Such hierarchical packaging of DNA is of fundamental importance to eukaryotic organisms. The central core of nucleosomes, called the nucleosome core particle (NCP) is composed of a histone octamer having four pairs of the core histone proteins: H2A/H2B-H4/H3-*H3/H4-H2B/H2A* wrapped around by 1.65 turns of 147 base pairs of nucleosomal DNA (LMR$^+$97; HHTB00) (italicized and non-italicized core histones refer to structurally distant sets of H2A/H2B-H4/H3 tetramers). The NCP crystal structure has been reported at 1.9 Å resolution (Protein DataBank: 1kx5) (DSL$^+$02). The structure of the constituent eight core histones is also well resolved, and consists of many basic Arginine and Lysine residues (Fig. 2.1). It is also known that the N-terminal histone tails are flexible in the crystal structure (LMR$^+$97; WSL01; JA01) and their reversible post-translational modifications (such as methylation, acetylation, phosphorylation, ubiquitination and ADP-ribosylation) trigger specific functions critical for nucleosome stability (SA00; FWA03). The high resolution

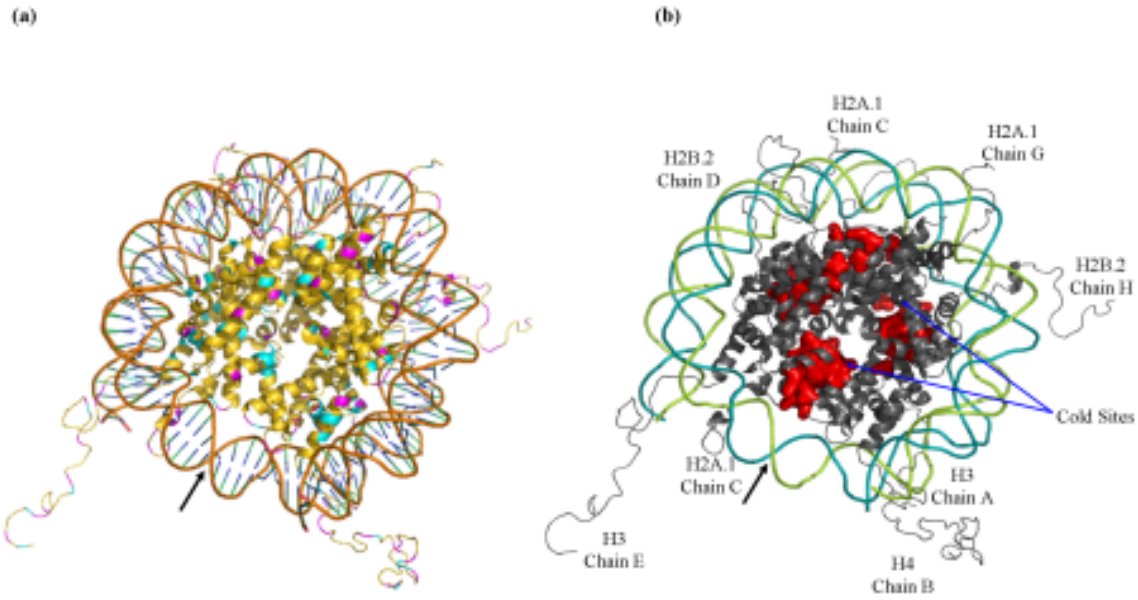

Figure 2.1: (a) Basic residues in the nucleosome core particle. Crystal structure of the nucleosome core particle (Protein DataBank ID 1kx5). Basic residues, lysine (dark gray) and arginine (black), present in the histone octamer assembly are shown. Strong salt-bridge interactions are formed between basic histone side chains and the phosphate backbone of nucleosomal DNA. (b) Snapshot of cold sites in the nucleosome. Cold sites are those histone residues that maintain more of their contacts throughout simulations than other residues. In discrete molecular dynamics simulations of histone-octamer assembly, we observe that these sites are present within the core of the nucleosome. A and B are aligned for comparison of cold sites against regions rich in basic lysine and arginine residues. Bold arrow indicates the nucleosomal dyad axis of symmetry.

crystal structures of the nucleosome core particle shows histone tails as posessing extended random coils topology, lacking secondary structure (LMR$^+$97; HHTB00). While proximal end of H2A histone tails passes through the minor groove DNA between superhelical gyres, the distal ends are spatially segregated from the nucleosomal DNA (LMR$^+$97; DSL$^+$02). Although many *in vitro* and *in vivo* assays demonstrate the role of conserved histone tail modifications in transcriptional regulation (SA00; Ber02), a structural understanding of how dynamics of histone tail mediates chromatin organization is poorly understood.

The organization and packaging of nucleosomes into discrete domains (such as eu-

chromatin and heterochromatin in eukaryotic cells) has been an intriguing puzzle for many years (KL99; ZLv98; vZ96). Yet, the precise control of chromatin structure is essential for all DNA-templated processes such as replication, recombination, repair and transcription (ZR01; Gru97). The heart of the chromatin structure is the NCP, whose structure has been solved at the atomic resolution (DSL$^+$02) and is composed of DNA and histone proteins. Many biophysical experiments have provided insights on the higher-order chromatin organization (LYR$^+$94). Coarse-grained molecular dynamics simulations may be used to explain the complex dynamics and the nature of internucleosome interactions mediating higher-order chromatin structure.

Although the primary sequence as well as the tertiary structure of histone proteins is highly conserved across genomes (RSKC05), a key element missing in our understanding of chromatin structure and function has been the lack of information pertaining to the histone tails, which are dynamic, resolved in many crystal structures and disordered (LR98). Increasing evidence indicates a fundamental role for histone tails and their covalent modifications in higher-order chromatin organization.

However, how these tails and their post-translational modifications contribute to the packaging and organization process of the chromatin is not well understood, but likely involves a combination of activities including the control of nucleosome stability, nucleosome-nucleosome interaction and the precise recruitment of protein machineries that organize discrete chromosomal domains. Chromatin organization is a highly complex process involving multiple steps and layers of regulation. Combining high resolution computational modeling of nucleosomes with long timescale discrete molecular dynamics (DMD) simulation enables a detailed understanding of the complex and dynamic nature of nucleosomes.

Molecular dynamics approaches have provided important insights into our understanding of the dynamics of proteins (KP90; DLD05; DBSS98; Dok06; DBB$^+$03b; LW75; PCC$^+$95) and nucleic acids (PCC$^+$95; CY00; CK00). Theoretical studies of chromatin

fiber and the NCP have been performed using coarse-grained (CG) physical models (Sch03; Sch06). Computational approaches that employ conventional all-atom molecular dynamics simulations (CCD$^+$05) using molecular mechanics and quantum mechanics force fields (PCC$^+$95; PC03) provide detailed information on the local dynamics of molecules. However, because of the large size of the NCP ($> 16,850$ heavy atoms in mono-nucleosome crystal structure) and the vast dimensionality of feasible conformations of the nucleosome, all-atom molecular dynamics simulations have severe limitations on the time scales and length scales on which the dynamics of nucleosomes can be studied (DD05). An alternative approach for improving the conformational sampling efficiency is using simplified structural models of protein and DNA. In these simplified models, amino acids and nucleotides are coarse-grained to the level of effective particles (beads), where each bead represents the center of mass or geometric centroid of a group of atoms. Local fluctuations among atoms constituting the beads are ignored and the interaction potentials between these beads are derived from the native crystal structure of NCP.

In this study, we examine the dynamics of NCP using fixed temperature DMD simulations. We further investigate the structural dynamics of our model nucleosome and the determinants of nucleosome stability using essential dynamics analysis, inter-histone and histone-DNA contacts found in DMD trajectories. The results presented here show that, histone tails in mono-nucleosomes form strong salt-bridge interactions with adjacent nucleosomal DNA, suggesting their direct functional role in stabilizing higher order chromatin structure. We identify a small fraction of histone core residues, termed cold sites, exhibiting significantly low fluctuations in multiple constant temperature simulations. We postulate a functional role of cold sites in mediating nucleosome stability. Our simulations with high resolution show interactions of distal ends of histone tails with the nucleosomal DNA due to formation of hydrogen bonds between the terminal Lysines/Arginines and DNA phosphates. Also, based on our simulations, we report the existence of cold sites - residues mediating the structural stability of the nucleosome

core particle.

2.2 Materials and Methods

2.2.1 Geometric description of model histone octamer

Amino acids in the histone octamer assembly are modeled by two effective beads per residue (DDB$^+$02a): Cα bead representing the coordinates of backbone α-carbon atoms and Cβ bead representing the coarse-grained coordinates of side chain β-carbon atoms (For glycine, the Cα and Cβ beads coincide) (PDU$^+$04; DBB$^+$03a). Amino-acids are assigned an index (i), corresponding to its position in the sequence starting from the N terminus (i = 1) to the C-terminus (i = N, number of residues). The geometry of histones is modeled by four types of bonds: (i) covalent bonds between Cαi and Cβi, (ii) peptide bonds between Cαi and C(αi$\pm$1) (iii) theoretical bonds between Cαi and Cαi$\pm$1), and (iv) effective bonds between Cαi and Cαi$\pm$2). Effective bond lengths for bond types (iii), (iv) are determined by computing the standard deviation of distances between carbon pairs in 103 representative globular proteins obtained from Protein DataBank as described in (DDB$^+$02a).

2.2.2 Geometric description of model nucleosomal DNA

Each nucleotide in the 147 bp DNA fragment is modeled as three-beads (Fig. 2.2): one bead for the sugar, phosphate and base, respectively. These beads represent the effective coordinates of sugar, phosphate and base portions of the nucleotide. The sugar bead S_i of the i^{th} nucleotide is positioned at the centroid of its constituent C1', C2', C3', C4' and O4' atoms, the phosphate bead P_i at the centroid of P, O1P, O2P and O5P atoms and the base bead B_i is positioned at the centroid of N1, C2, N3, C4, C5 and C6 atoms. The average bond length parameters of beads and their standard deviations of nucleosomal DNA were obtained from the available high resolution NCP structure

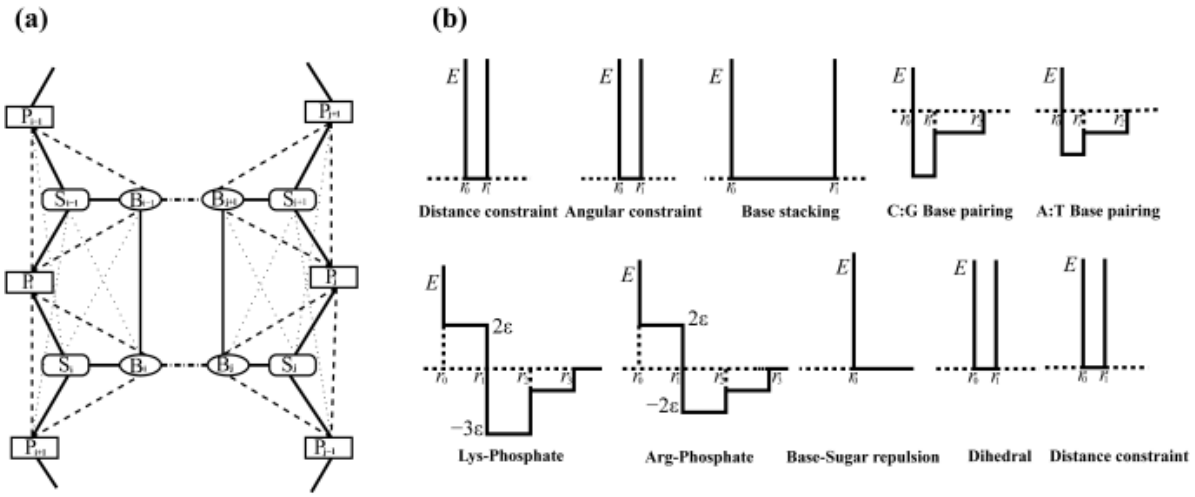

Figure 2.2: (a) Interactions between the model DNA beads. Consecutive nucleotides in the DNA double helix are shown. The sugar, phosphate and base beads are indicated by S_i, P_i and B_i, respectively. Covalent interactions between these beads are shown by (dark solid lines) and noncovalent interactions by dashed lines: angular constraints (dotted lines), basepairing (dash-dotted line), base stacking (thin solid lines), and dihedral constraints. (b) Interaction potentials used in the DNA model. Each interaction is modeled as a coarse-grained square-well potential. Lysine-phosphate and arginine-phosphate interactions have a weaker long-range attractive shoulder to simulate the effects of solvent-mediated hydrogen bonds. Steric repulsion between sugar and base beads is modeled as hardcore repulsion. Constraints for distance, angular, dihedral, and base stacking are infinitely high potential wells whose breadth depends on the nature of the interacting species. The depths of potential wells for A-T and C-G basepairs are proportional to their corresponding strength of interaction: 2:3. The radial separations (such as r_0, r_1, and r_2) at which these constraints are applicable are based on mean separation and fluctuations of corresponding beads observed in the crystal structure of the nucleosome core particle.

(DSL$^+$02). The structural parameters used in the model are listed in Table 2.1.

2.2.3 Simulation potentials

Interactions in the histone octamer assembly CG model are modeled as theoretical pairwise interactions. These interactions include covalent bonds between Cβi and CCβi, peptide bonds between Cαi and Cα(i+1), angular constraints between Cβi and Cαi(i+1), and dihedral constraints between Cαi and Cα(i+2) beads. These additional bonds model angular and dihedral constraints between side chains and backbones. Theoretical bonds are used mimic the tetrahedral constraints of amino-acids and the planar constraints of peptide bonds. Permanent bonds are realized by infinitely high potential wells, while hard-core repulsions are modeled by infinitely high square-well potentials (PDU$^+$04; DBD05).

In nucleosomal DNA, covalent interactions between sugar, phosphates, and bases are modeled as infinitely high potential wells: $Vi, j = 0$ if ri, j lies within one standard deviation ($\sigma_{i,j}$) of the mean bond length ($D_{i,j}$) between beads i and j and $Vi, j = \infty$, otherwise. Noncovalent interactions are modeled as discrete attractive/repulsive potentials: DNA base-pairing interactions are attractive and the base-stacking interactions along the chain incur steric repulsions. The interactions between core histones and the DNA are modeled as nonspecific electrostatic attractions between the basic side chains of lysine and arginine residues and the acidic phosphates present in the DNA backbone. Interactions between histone amino acids and nucleotide bases are dominated by direct and solvent (water)-mediated electrostatic attractions (hydrogen bonds) between basic arginine and lysine side chains and DNA backbone phosphates (DSL$^+$02). Repulsive interactions of nucleosomal DNA with acidic and nonpolar amino acids are ignored to mimic $G\bar{o}$-like interactions. Interaction potentials used for the DNA model are shown in Fig. 2.2.

The structural parameters required for the model are mean bond lengths and stan-

(a) Distance constraints:

Interacting Pair	r_0 (Å)	r_1 (Å)
$S_i - P_i$	4.0	4.4
$S_i - P_{i+1}$	3.8	4.0
$A_i - S_i$	4.9	5.1
$C_i - S_i$	3.8	4.0
$G_i - S_i$	4.9	5.0
$T_i - S_i$	3.8	3.9

(b) Angular constraints

$A_i - P_i$	7.1	8.2
$C_i - P_i$	5.9	6.7
$G_i - P_i$	7.1	8.3
$T_i - P_i$	5.9	6.7
$P_i - P_{i+1}$	6.4	7.0
$S_i - S_{i+1}$	5.3	5.7
$A_i - P_{i+1}$	7.5	8.0
$C_i - P_{i+1}$	6.9	7.4
$G_i - P_{i+1}$	7.5	8.0
$T_i - P_{i+1}$	6.8	7.3

(c) Dihedral constraints

$P_i - S_{i+1}$	9.0	9.7
$A_i - S_{i+1}$	5.6	6.5
$C_i - S_{i+1}$	6.3	7.5
$G_i - S_{i+1}$	5.7	6.8
$T_i - S_{i+1}$	6.0	7.0
$S_i - P_{i+2}$	8.9	9.4
$S_i - A_{i+1}$	5.9	7.2
$S_i - C_{i+1}$	5.0	5.7
$S_i - G_{i+1}$	5.9	6.9
$S_i - T_{i+1}$	5.4	6.2

Table 2.1: Set of DMD constraints and interaction radii used to model the nucleosomal DNA. S_i, P_i, A_i, C_i, G_i and T_i represent the i^{th} nucleotide's sugar, phosphate, adenine-base, cytosine-base, guanine-base and thymine-base beads, respectively. r_0, r_1, r_2, r_3 are the corresponding interaction radii used in the model. K_j and R_j denote the side chain beads of j^{th} lysine and arginine amino acids, respectively.

(a) Base stacking

$A_i - A_{i+1}$	3.4	4.7
$A_i - C_{i+1}$	3.6	4.7
$A_i - G_{i+1}$	3.4	5.8
$A_i - T_{i+1}$	3.5	4.4
$C_i - A_{i+1}$	4.1	6.8
$C_i - C_{i+1}$	3.8	6.6
$C_i - G_{i+1}$	3.9	6.7
$C_i - T_{i+1}$	3.9	6.6
$G_i - A_{i+1}$	3.6	4.8
$G_i - C_{i+1}$	3.8	6.9
$G_i - G_{i+1}$	3.4	5.1
$G_i - T_{i+1}$	3.6	4.4
$T_i - A_{i+1}$	4.5	6.4
$T_i - C_{i+1}$	3.7	5.0
$T_i - G_{i+1}$	4.2	6.8
$T_i - T_{i+1}$	3.8	5.3

(b) Base pairing

Interacting Pair	r_0 (Å)	r_1 (Å)	r_2 (Å)
$A_i - T_j$	5.4	5.8	6.3
$C_i - G_j$	5.5	5.8	6.0

(c) Lysine/Arginine-Phosphate attraction

Interacting Pair	r_0 (Å)	r_1 (Å)	r_2 (Å)	r_3 (Å)
$P_i - C_\beta$	3.3	4.0	6.0	8.0

(d) Base-Sugar repulsion

Interacting Pair	r_0 (Å)
$A_i - S_j$	8.3
$G_i - S_j$	8.3
$C_i - S_j$	10.3
$T_i - S_j$	10.3

Table 2.2: Set of DMD constraints and interaction radii used to model the nucleosomal DNA. S_i, P_i, A_i, C_i, G_i and T_i represent the i^{th} nucleotide's sugar, phosphate, adenine-base, cytosine-base, guanine-base and thymine-base beads, respectively. r_0, r_1, r_2, r_3 are the corresponding interaction radii used in the model. K_j and R_j denote the side chain beads of j^{th} lysine and arginine amino acids, respectively.

dard deviations for each pair of the modeled interactions (Fig. 2.2). The structure of nucleosomal DNA is remarkably different from that found in nonhistone protein-DNA complexes or canonical B DNA (RD03). These structural parameters, which have been derived from the high-resolution NCP structure, are used in all the simulations (LMR$^+$97). The numerical values of each of the r_0, r_1, r_2, and r_3 interaction radii used in the model are listed in Tables 2.1 and 2.2. In this model, hydrogen bond interactions are not amino-acid specific. The hydrogen-bond potential energy, ε is considered as the energy unit. The DMD potential for purine-pyrimidine interactions are scaled as -3ε and -2ε for G-C and A-T basepairing, according to the relative strengths of these basepairs: three hydrogen bonds are formed in the G-C basepair, whereas two hydrogen bonds are formed in the A-T pair. Similarly, the interactions between DNA and lysine/arginine are also scaled according to the number of potential hydrogen bonds formed with the DNA.

2.2.4 DMD algorithm

Using simplified square-well potentials as interparticle interactions, we apply the discrete time molecular dynamics simulations approach to study the dynamics of the NCP (DD05; PDU$^+$04; ZK97). In this approach, the beads move with a constant velocity until an elastic or inelastic collision occurs. Upon collision, the interaction potential of the beads changes, thereby changing the kinetics of colliding beads. DMD maintains a set of possible collisions and the current state of each bead. After each collision, DMD updates the set of possible collisions and the state for colliding beads. It then determines the pair of beads undergoing the earliest subsequent collision. Since every collision needs to update the state of only the colliding pair of beads, this approach samples a vast dimensionality of histone and DNA conformations. DMD simulations of two-bead-per-residue protein models may be performed using the iFold server (http://iFold.dokhlab.org) (SDN$^+$06). Each 1×10^5 time units (t.u.) of NCP simulations

takes $\approx$ 30 days on a single 2.4-GHz Intel Pentium IV-based workstation.

2.2.5 Essential dynamics of the nucleosome core particle

The essential dynamics of a multiparticle system separates large concerted structural rearrangements from irrelevant uncorrelated fluctuations (ALB93). In this method, we compute the normalized covariance matrix (NCM):

$$NCM(i,j) = \frac{< r_i - < r_i >> . < r_j - < r_j >>}{\sqrt{<< r_i.r_i > - < r_i > . < r_i >> . << r_j.r_j > - < r_j > . < r_j >>}}$$

Here, r_i and r_j represent the cartesian coordinates of the i^{th} and j^{th} beads, respectively, and the bracketed values represent averages over the entire trajectory. The covariance matrix describes the correlation of the positional fluctuations of α carbon beads for the histone core and of phosphate beads for nucleosomal DNA. To verify the validity of calculated correlation coefficients, we perform covariance analysis over two $3{\times}10^4$ t.u. of nonoverlapping DMD subtrajectories and observe that the two covariance matrices thus obtained are nearly identical to the matrix obtained for the entire trajectory. We then diagonalize the covariance matrix of fluctuations of atoms (beads in the case of the coarse grained model) along the trajectory, yielding eigenvectors as directions in the 3N-dimensional subspace (where N represents the total number of Cα and P beads in the system). Most of the topological fluctuations cluster in correlated motions in a subspace of a few degrees of freedom, whereas the other degrees of freedom represent independent uncorrelated fluctuations. The eigenvalues are a measure of the mean-squared fluctuations of the constituent beads along the corresponding eigenvectors, and are computed using the QL algorithm of Numerical Recipes in C (HPAT88). The eigenvalues are sorted in descending order, and the corresponding first eigenvalue represents the largest topological fluctuation and a majority of the fluctuations are restricted to first few eigenvectors.

2.2.6 Heavy-atom reconstruction of histone, DNA conformations

Using the reconstruction procedure, the CG model (two-bead) trajectories of histones obtained from DMD simulations are transformed into a heavy-atom representation (N, C, O, CA, and CB). The method of heavy-atom reconstruction used for histones is described in Ding et al. (DPCD06). A four-bead representation of each residue is generated by adding N and C' atoms into the simulated two-bead (Cα-Cβ) model. The conformation of this four-bead model was then relaxed to the lowest energy state and the secondary-structure elements were refined using short discrete molecular simulations. The side-chain and backbone oxygen structures are added according to the most stable (Cα, Cβ, N, C') conformation. Backbone and side-chain rotamers are optimized using Monte Carlo-based simulated annealing procedure using the Dunbrack and Cohen backbone-dependent rotamer library (DC97).

DNA reconstruction is used to generate a heavy-atom trajectory of DNA from three-bead trajectories generated by DMD. For each nucleotide present in the crystal structure of nucleosomal DNA, we generate coordinates of the corresponding sugar (S_i), phosphate (P_i), and base (B_i) beads, and the coordinates of preceding sugar (S_{i-1}) and succeeding phosphate (P_{i+1}) beads, forming a five-bead nucleotide conformation template $T_i = [S_i, P_i, B_i, S_{i-1}, P_{i+1}]$. We then classify these conformation templates according to nucleotide type: adenine, cytosine, guanine, or thymine, yielding a library of CG nucleotide conformations present in the native state. Each snapshot of the simulation trajectory is then reconstructed as follows: for the jth nucleotide of the snapshot, the target conformation $T_j = [S_j, P_j, B_i, S_{j-1}, P_{j+1}]$ is structurally superimposed with each of the templates of corresponding nucleotide type using the Kabsch algorithm (KS83). The template $T_k = [S_k, P_k, B_k, S_{k-1}, P_{k+1}]$ minimizing root-mean-square deviation with target T_j is chosen and the rotation matrix $R_{k \to j}$ transforming template T_k to target structure T_j is computed. This rotation matrix R is then applied to the crystal structure

coordinates of the nucleotide corresponding to the k^{th} template to yield the heavy-atom structure of the j^{th} nucleotide of the snapshot.

2.2.7 Analysis of conserved contacts: interhistone and histone-DNA contact frequencies

Frequencies of interhistone and histone-DNA contacts reveal key contacts conserved in the course of simulation. We define two histone residues to be in contact if the separation between their corresponding Cβ beads is $\leq$7.5 Å. Mean frequencies of interhistone contacts are evaluated by averaging the contacts formed over the entire simulation trajectory. A comparison of mean frequencies of interhistone contacts against the contacts present in the native state reveals key histone-histone interactions conserved in the simulation, thereby ascertaining the flexibility of the contact. We propose that interactions having high frequencies of histone-DNA contacts in the constant-temperature simulations specify key interactions responsible for nucleosomal stability. A large fraction of histone-DNA contacts in the NCP are solvent-mediated salt bridges and hydrogen bonds (Wid01) between the backbone phosphate of DNA and basic histone side chains. Thus, we define a contact between histone residue and DNA nucleotide if the separation between the corresponding Cβ and phosphate beads is $\leq$11.5 Å. We plot the frequencies of histone-DNA contacts formed in constant-temperature DMD simulations performed over a range of temperatures.

2.2.8 Estimation of the DMD simulation timescales

Evolution of DMD trajectories does not require Verlet integration; rather, it computes iterative solutions of the ballistic equations of motion under soft square-well potential. Longer timescales are accessible by DMD simulations due to integration of available degrees of freedom in CG models and use of soft square-well potentials/implicit solvation

in DMD simulations. The classical equipartition principle divides thermal motions of nucleosomes into translational, rotational, and vibrational degrees of freedom. However, high-frequency vibrations, such as hydrogen vibrations, are typically uncoupled from the mean-field dynamics of the system. The effective mean-field interactions in CG models reduce the classical degrees of freedom.

Nielson et al. (NLSK04) have estimated the effective timescales accessed by CG models for dimyristoyl phosphatidylcholine in water (13 beads/molecule), calibrating diffusion coefficients for CG models against all-atom simulations and experiments. There is a 100-fold increase in timescales for translational and rotational diffusions between atomistic versus CG simulations. The CG model for DNA incorporates nucleotides as three beads/nucleotide (≈20 heavy atoms and 14 hydrogen atoms, 1:11 reduction) and proteins are coarse-grained as two beads/residue (≈10 heavy atoms and 12 hydrogen atoms, 1:11 reduction). We estimate a 100-fold increase in translational and rotational diffusional timescales for CG nucleosomes. In addition, the use of soft square-well potentials in simulations allows another 10- to 100-fold increase in simulation timesteps (MP02). In the worst-case scenario, there is at least a three-orders-of-magnitude reduction in time steps due to the use of soft potentials and implicit solvation in CG models.

To incorporate the effect of time-step discreteness we compute the fundamental DMD time unit along with the effective scaling Formula 1 due to reduced degrees of freedom in the CG model. The timescales for CG models in DMD are given by: [TCG] ≈ sCG[L]$\sqrt{[M]/[E]}$. For NCP simulations, the effective mass of each coarse-grained bead is ≈100 g/mol, the unit length of simulation is 1 Å, and the DMD unit energy is 1 kcal/mol. This leads to the estimate: [TCG] ≈ 0.5 ns. Thus, one time unit in CG-DMD simulations of nucleosomes corresponds to ≈0.5 ns of physical time. Consequently, our coarse-grained NCP and histone octamer simulations of 1×10^5 t.u. correspond to simulating dynamics for roughly 50 μs of experimental time. Zhou et al. (ZK99) have

also used a coarse-grained model to investigate the timescales of CG simulations. The authors have translated the simulation timescales to physical timescales by comparing the dynamics observed in their simulations to experimental results. They conclude that the reduced simulation time unit corresponds to 1 ns of physical time; thus, our time unit is also of the order computed by Zhou et al. (ZK99).

2.3 Results

To understand the role of DNA in nucleosome stability, we first describe the comparison of DMD simulations of NCP with simulations of the histone octamer complex. We then describe essential dynamics analysis of the NCP to study the large-scale dynamics of nucleosomes. We then present the frequencies of interhistone and histone-DNA contacts and analyze contacts mediating nucleosome stability. Throughout this study, the temperature is measured in DMD units of energy ε divided by Boltzmann's constant, ε/k_B (see 2.2). The reduced temperature 0.7 ε/k_B corresponds to approximately the ambient temperature (T_{amb} = 300 K).

2.3.1 DMD simulations of nucleosomes display cold sites in the nucleosome core

Our simulations of the histone octamer complex reveal that in the absence of nucleosomal DNA, histone tails are highly mobile in nature, and often adopt random-coil conformations. We study the equilibrium behavior of the histone octamer complex and NCP by measuring the heat capacity and the average potential energy as a function of temperature. Based on our constant-temperature simulations, we define the unfolding temperature of the histone octamer assembly to be $T_f = 0.8\varepsilon/k_B$. For nucleosomes, we performed constant-temperature DMD simulations over a temperatures range $T = 0.1 - 2.8$ for $1{\times}10^5$ t.u. At each sampled temperature, we start with the

native-state (crystal structure) conformation and perform DMD simulations for 5×10^4 t.u. simulation to equilibrate the system, followed by an additional 5×10^4 t.u. for recording the simulation trajectory (see Materials and Methods). The dependence of average potential energy and related heat capacity versus temperature for the NCP is shown in Fig. 2.3. We find that the NCP folding temperature is 0.92 ε/k_B. The heat capacity is computed from the relation $Cv =< (\delta E)^2 > /T^2$

In simulations performed at low temperatures ($T = 0.1 - 0.5\varepsilon/k_B$), the core histone octamer is rigid, while the histone tails are flexible. Under high-temperature conditions, the histone octamer is destabilized, thereby contributing to increased flexibility of histone tails. We characterize specific regions present in the histone core (cold sites) (Fig. 2.1), where residues retain a majority of their contacts throughout simulations compared to other residues (trajectory-normalized contact frequency > 0.7). We find that many cold sites are composed of hydrophobic residues, clustered as domains of five of more adjacent residues, and are present in the core of the nucleosome. Large fractions of cold sites are clustered in the interface between H3-H3 histones formed by the C-terminal helices of H3 histone fold domains: (His-113A, His-113E, Ala-114A, Ala-114E, Leu-126A, Leu-126E, Ala-127A, Ala-127E, Arg-131A, Arg-131E, and Ile-130E). Interhistone interactions between the H3-H3 interfaces are essential for fastening the two H2A/H2B-H4/H3 NCP tetramers. This H3-H3 interface thereby mediates the dynamics of these halves of the NCP. Cold sites are also localized between the two H4-H2A and H4-H2A interfaces: (Thr-96B, Leu-97B, Tyr-98B, Gly-99B, Val-100G, Thr-101G, Ile-102G, Ala-103G) and (Thr-96F, Leu-97F, Tyr-98F, Val-100C, Thr-101C, Ile-102C, Ala-103). As opposed to the H3-H3 histone fold, the interface of each of these two domains is formed by small stretches of parallel interhistone beta-sheets (H4-H2A and H4-H2A).

We posit that these cold sites are essential for the stability of histone octamer complex and that the presence of clusters of cold sites at interhistone interface suggests that

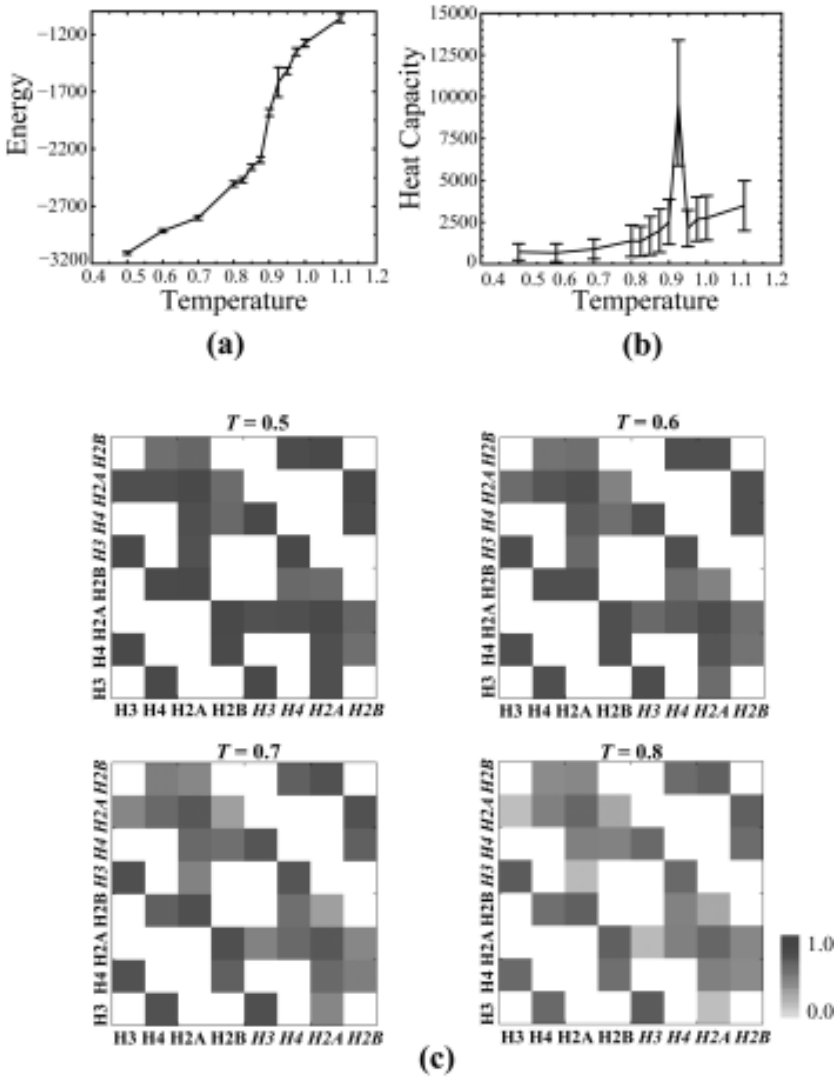

Figure 2.3: Thermodynamics of the nucleosome core particle. The dependence on temperature of the energy, E, is shown. The error bars represent a standard deviation of energy fluctuations. (b) Variation of the constant-volume heat capacity, C_v, of the nucleosome core particle with temperature. The error bars are the standard deviation of C_v fluctuations. NCP unfolding occurs at temperature $T = 0.92$. (c) Temperature dependence of frequencies of interhistone contacts. Histone chains making contacts in constant-temperature simulations of the histone-octamer complex are shown for temperatures $T = 0.3$-0.8. Frequencies of interhistone contacts are color-coded from white (contact frequency of 0) to black (contact frequency of 1). Under high-temperature conditions, the frequencies of contact are significantly reduced; however, novel interhistone contacts are not observed, indicating the absence of intranucleosome domain swap.

domains containing cold residues may have a direct functional role in stabilizing the corresponding interactions at interhistone interfaces. During transcription elongation, the interactions at these interhistone cold sites may need to be weakened in part, resulting in significant destabilization of the nucleosomes. We find less abundance of cold sites in the histone fold domains of H2A, H2B, and H4 histones, suggesting that within the NCP, the globular histone fold domains of these histones have relatively greater flexibility than H3 histones. The frequency of contacts between H3-H2A, H3-H2B interfaces is low, indicating that these contacts have weak interactions in the NCP. The enhanced rigidity of selective interhistone interfaces by cold sites also suggests an order of histone release during nucleosome dissociation, whereby weakly interacting H2B and H2A histones are dislodged before the release of strongly bound H3, H4 histones.

2.3.2 Essential dynamics of nucleosome and histone octamer assembly

To elucidate the global dominant motions within the NCP, we use essential dynamics approach (ALB93) on DMD trajectories to generate the principal components of nucleosome dynamics during constant temperature simulations. In this approach, the collective concerted fluctuations in the NCP are projected onto the principle-components subspace (see Materials and Methods). In the principal-components subspace, eigenvectors and eigenvalues of the covariance matrix represent the direction and amplitudes, respectively, of the essential motions of nucleosome. We perform separate essential dynamics analyses on the DMD trajectories of the histone octamer and the NCP. We find that in both cases, the largest principal component of nucleosome dynamics corresponds to flagellar motions of flexible histone tails. This behavior is conserved over a wide range of temperatures examined ($T = 0.1 - 1.1$). Normalized correlation maps depict correlations between motions of all pairs of histone-histone, histone-DNA, and DNA-DNA beads (cf. Materials and Methods). By comparing the normalized correlation maps of

the NCP and the core histone octamer, we find that in the absence of nucleosomal DNA, dynamics of intrahistone residues are strongly correlated, whereas dynamics of interhistone residues are largely uncorrelated. The subsequent component of histone-octamer dynamics consists of bending of the two [H3-H4-H2A-H2B] tetramers relative to each other about the H3-H3 interface. This observation is consistent with our previous result: the cold sites found at the H3-H3 interface mediate large-scale dynamics of the NCP.

Temperature dependence of the normalized correlation map of the nucleosomes over a range of temperatures ($T = 0.1 - 1.2\varepsilon/k_B$) demonstrates that in the presence of DNA, the fluctuation of histone tails is suppressed by hydrogen-bonded interactions with nucleosomal DNA. Within the histone octamer core, the H3 and H4 core histones belonging to the same H2A/H2B-H4/H3 histone tetramer undergo mutually correlated dynamics, whereas these motions are uncorrelated with the motion of H2A, H2B histones and H3 and H4 histones belonging to the other H2A/H2B-H4/H3 tetramer. However, the two H2A and H2B histone pairs belonging to the same histone tetramer are mutually correlated. This result suggests that the dynamics of the two histone tetramer halves of the NCP are largely uncoupled with each other; however, their constituent histones have strongly correlated dynamics.

We find that in high-temperature simulations, nucleosomal DNA collapses into the histone octamer assembly. Because of base-pairing interactions, relative motions between the DNA strands are strongly correlated with each other. Under high temperature regimes (T = 0.8-1.2 ε/k_B), the motion of DNA is significantly anti-correlated with motion of H2A and H2B histones and is largely uncorrelated with motion of H3 and H4 histones. This finding suggests presence of fluctuating DNA-histone interactions formed with H2A and H2B, which may impart conformational flexibility and thereby assist in stabilizing the NCP. Also, absence of correlated motions between H2A/H2B and H3/H4 heterodimers at elevated temperatures shows that the contacts between the two heterodimers are weakened under these conditions.

2.3.3 Contact frequencies reveal key interhistone and histone-DNA interactions

We calculate contact frequencies for all histone-DNA contacts formed at a range of temperatures from DMD simulations. The contact frequency map shows frequencies of histone-DNA interactions formed, averaged over the simulation trajectory. We have generated the map of contact frequencies for histone-DNA contacts formed during simulations at temperatures T = 0.1, 0.8, and 1.2 (Fig. 2.4). We find that under low-temperature conditions (T = 0.1), fewer histone-DNA contacts are formed relative to high-temperature conditions (T = 0.8 and 1.2), which is a characteristic of higher conformational rigidity of DNA under low-temperature conditions. A high frequency of histone-DNA contacts specifies key salt-bridge interactions persistent in the constant-temperature DMD simulations. We observe that a large number of histone-DNA contacts are long-range interactions and the frequencies of intrahistone contacts decreases monotonically as the temperature is increased from 0.1 (below unfolding temperature), through 0.9 ε/k_B ($\approx T_f$) to 1.2ε/k_B (above the unfolding temperature). We also find that all long-range contacts have frequencies close to zero. Our coarse-grained histone-DNA interaction potentials mimic the first-order simplification of specificity among amino acid-base interactions as demonstrated by Luscombe et al. (LLT01; LT02). We have generated a detailed map illustrating frequencies of inter- and intrahistone contacts formed in the presence of DNA. We find that in the presence of DNA, there is an increase in the number of interhistone contacts formed, suggesting that in the NCP, histones are tightly embraced by nucleosomal DNA.

A plot of per-nucleotide fluctuations of the two nucleosomal DNA strands observed in constant-temperature DMD simulations performed at temperatures T = 0.1-0.8 ε/k_B (folded state) is shown in Fig. 2.5. Fluctuations are computed as standard deviations of phosphate beads, relative to their initial conformation, averaged over the last 3×10^4 t.u. of the total time (1×10^5) of simulation trajectory. Nucleotides having low fluc-

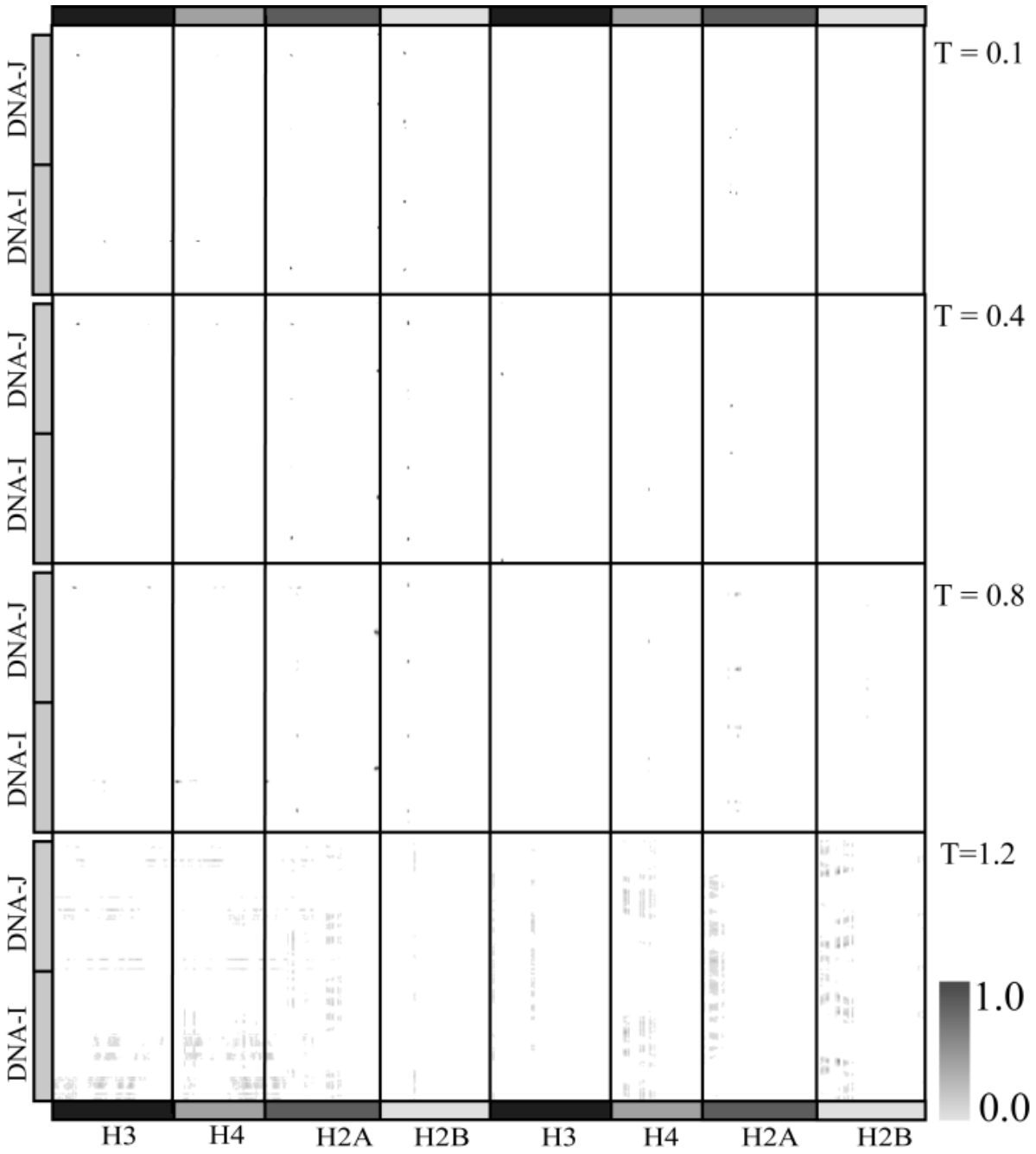

Figure 2.4: Temperature dependence of histone-DNA contact frequencies. The frequencies of contact are color-coded from white (contact frequency of 0.0) to black (contact frequency of 1.0). Few contacts are populated in the low-temperature condition (T = 0.1), indicating that the nucleosome core particle is energetically trapped in a local energy minimum. Under high temperature conditions, T = 0.4 and 0.8, we observe a substantial increase in highly frequent histone-DNA contacts formed. Under very high temperature conditions (T = 1.1), the nucleosome core particle is unfolded and the frequency of histone-DNA contacts becomes uniform over the contact-space.

tuations ($\leq$ 1.0Å RMSD) impart structural rigidity to the nucleosomal DNA and are conformationally constrained by interactions with the histone octamer assembly. These fluctuations in nucleosomal DNA are specific to the DNA sequence context and significant variations in the magnitude of fluctuations among neighboring nucleotides are observed in our simulations. DNA fragments making strong contacts with neighboring nucleotides have low mean fluctuations (Fig. 2.5). The dynamics of the two DNA strands are cross-correlated with each other over the range of temperatures used in the simulation; however, the extent of correlation is diminished at the elevated temperature (T = 1.2 ε/k_B), indicating that the standard Watson/Crick basepairing is conserved in simulations performed at elevated temperatures. The crystallographic temperature factors for corresponding phosphorus atoms in the two strands are also shown (Fig. 2.5). We observe that these experimentally observed temperature factors correlate with the extent of fluctuations of the phosphate beads. Localized sequence specificity of these DNA fluctuations is essential for the nucleosome positioning code (SFMC$^+$06) and is expected to be a conserved feature for nucleosomal DNA across genomes.

2.3.4 Modulating DMD histone-DNA interaction potentials simulates salt effects

The electrostatic environment of the NCP is known to have a significant effect on its dynamics (GP02; WK98). We performed constant-temperature simulations of mononucleosomes with varying strengths of interaction between histone side chains and DNA over a range of temperatures. In our model, DNA backbone phosphates have attractive (electrostatic) interactions with histone lysine and arginine residues (see Materials and Methods). A coarse-grained phase-space plot of histone-DNA interactions is shown in Fig. 2.6. Since salt and ionic environment are known to mediate the stability of histone-DNA contacts (OW87; YMvH89; MLV$^+$02; MLDL03), varying the strengths of these histone-DNA contacts qualitatively models the variations in salt concentration

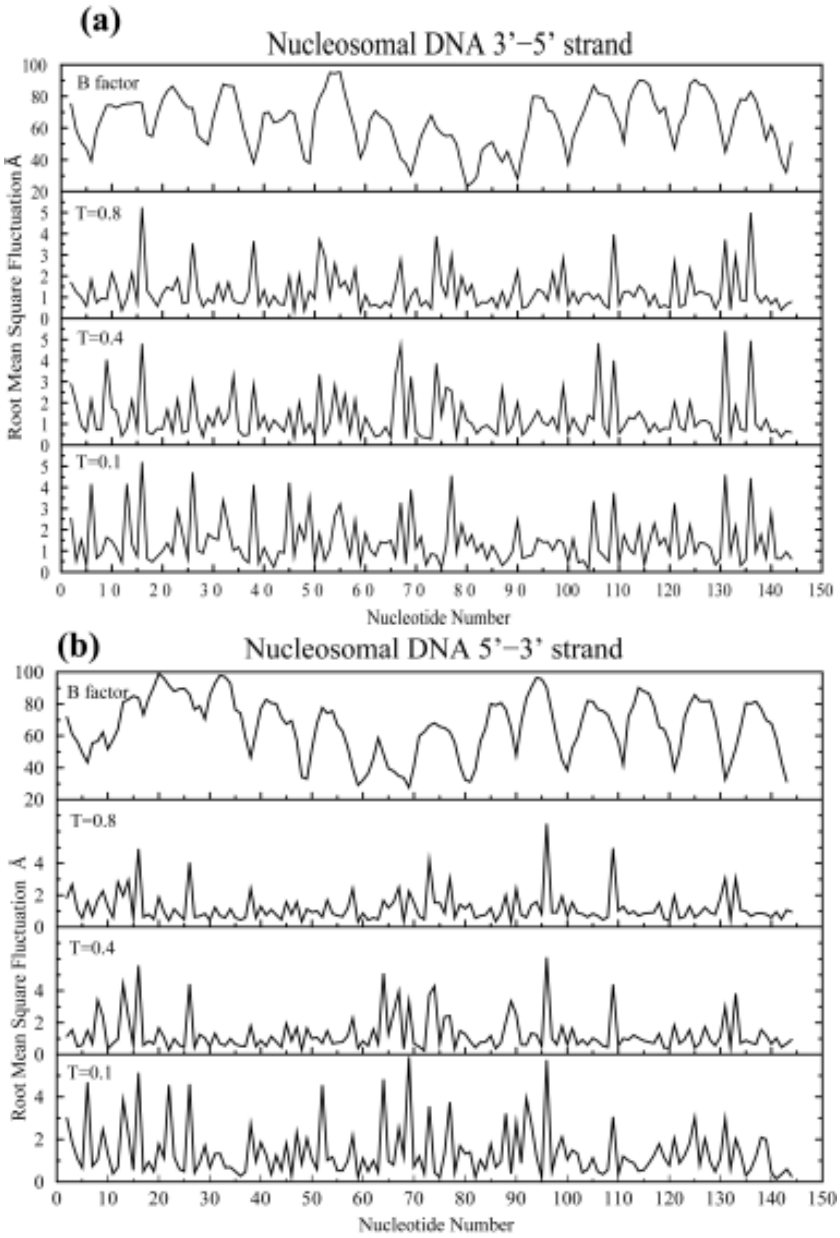

Figure 2.5: Fluctuations in nucleosomal DNA conserved over a range of temperatures. Root-mean-square deviations (in Angstroms) with respect to the initial conformation of the phosphate beads along the two nucleosomal DNA strands: 3'-5' (a) and 5'-3' (b) are shown at temperatures $T = 0.1$, 0.4, and 0.8 ε/k_B. Crystallographic temperature factors for corresponding phosphorus atoms derived from the crystal structure (Protein Data-Bank 1kx5) are also shown. The extent of these fluctuations is observed by averaging the fluctuations in the last $3{\times}10^4$ t.u. of the corresponding constant-temperature DMD simulation trajectories. The fluctuations in nucleosomal DNA are sequence-specific and are correlated (correlation coefficient 0.55) over the shown range of temperatures (0.1-0.8 ε/k_B). These crystallographic temperature factors are also correlated with these fluctuations in phosphate beads. The simulation data is averaged over three independent constant-temperature DMD simulation runs performed at temperatures from 0.1 to 0.8 ε/k_B.

around the nucleosome. Calibration of histone-DNA interaction strengths enables us to recapitulate the effects of salt on the stability of NCP (MLV$^+$02; MLDL03). We have performed high-temperature simulations (T = 1.2-2.8 ε/k_B) to estimate the NCP phase-space behavior. We observed that the ensemble of NCP conformations found at T = 1.2 ε/k_B was qualitatively similar to the conformational ensemble at significantly higher temperatures, with progressively increased amounts of DNA-end fraying. However, temperatures above T = 1.2 ε/k_B are physically unrealistic and are unlikely to be attained under normal physiological conditions.

The number of histone-DNA contacts formed in the simulations enumerates the stability of the NCP. We observe that under conditions of low interaction strength, the interactions between histones and DNA are predominantly local and consist of weak salt bridges, which are often disrupted by thermal fluctuations or in simulations at higher temperatures. Under conditions of moderate attraction between histones and DNA, ($\varepsilon_{His-DNA}$ = 0.4-1.2 k_BT_{amb}), the NCP is stable over a range of temperatures. Under very large values of histone-DNA interaction strengths ($\varepsilon_{His-DNA}$ = 1.2-2.8 k_BT_{amb}), interhistone interactions become significantly weaker than histone-DNA interactions. In this phase, the wrapped nucleosomal DNA entwines and presses against the histone octamer assembly. We characterize $\varepsilon_{His-DNA} = 0.8\ k_BT_{amb}$ as the histone-DNA interaction strength for the physiologically relevant scale of histone-DNA contacts.

2.4 Discussion

In our simulations, we observe persistent salt-bridge interactions formed between histone tails and nucleosomal DNA at low temperatures, whereas thermal fluctuations disrupt these contacts in high-temperature conditions. These interactions include direct (short-range) and solvent-mediated (long-range) contacts and mediate the stability of the NCP. Experimental evidence suggests that histone tails are highly mobile in the NCP

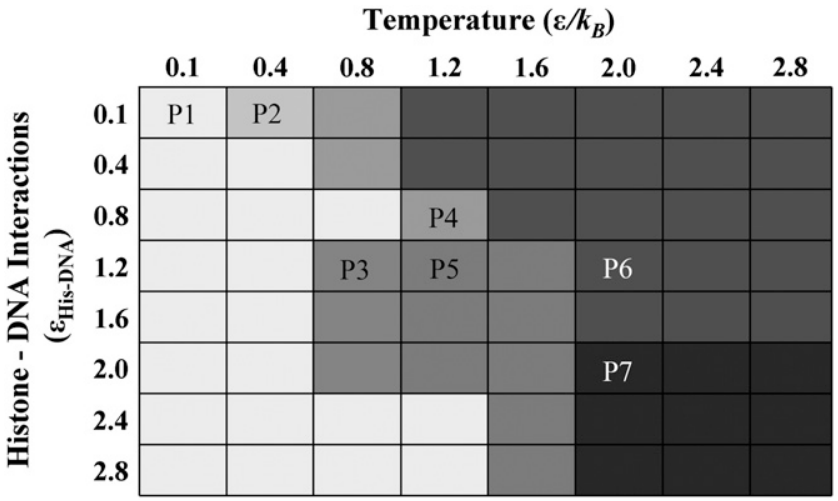

Figure 2.6: Phase space of the nucleosome core particle. This coarse-grained phase space of the nucleosome core particle is generated by performing discrete molecular dynamics simulations of the nucleosome core particle over a range of simulation temperature (T = 0.1-2.8 ε/k_B, step size of 0.4 ε/k_B) and also over a range of histone-DNA interaction potentials: $\varepsilon_{His-DNA}$ = 0.1-2.8 k_BT_{amb}, with step size ε = 0.4 k_BT_{amb}. Interactions between lysine side chains and DNA backbone phosphate correspond to -3 $\varepsilon_{His-DNA}$, whereas interactions between arginine side chains and DNA backbone phosphate correspond to -2 $\varepsilon_{His-DNA}$. Independent sets of constant-temperature DMD simulations are performed at each of these interaction potentials and over this range of temperatures to obtain the nucleosome behavior under varying physical conditions. Differences in grayscale intensities (labeled P1 to P7) correspond to variations in equilibrium NCP conformations observed over the range of temperatures, and histone-DNA contact potentials. P1), Histone tails bind to DNA; NCP is stable. P2), DNA segregated from histones; histone octamer is intact. P3), Nucleosomal end-fraying: terminal DNA base-pairing is disrupted, histone octamer is intact. P4), Nucleosome intact, histone tails bind DNA; H2A histones extruded from the NCP and the H3-H3 interhistone interactions stabilizing the NCP are disrupted. P5), DNA ends unfurl, collapse on histones; histone octamer unfolds. P6), DNA base-pairing is segregated and the histone octamer unfolds. P7), DNA collapses onto the histone octamer, and the H3-H3, H3-H4, and H2A-H2B contacts unfold.

and often adopt random-coil structures (DSL$^+$02). Luger et al. (LR98) have proposed that histone tails essentially assemble nucleosomes into chromatin fibers. Transcriptional regulation by covalent histone tail modifications is well established in literature (JA01; Ber02; ZR01). Many histone acetyltransferases are identified as transcriptional coactivators (OSR$^+$96). In our NCP simulations, we observe that the histone tails are highly mobile and they bind to proximal nucleosomal DNA. These strong electrostatic interactions between DNA phosphates and histone tails are conserved even in simulations performed at elevated temperatures. This finding supports our hypothesis that histone tails may have a direct structural role in stabilizing higher-order chromatin structure, and suggests a structural mechanism for stabilizing higher-order chromatin organization (CASC00), whereby histone tails bind to neighboring nucleosomal DNA and this binding is destabilized by histone tail modifications leading to nucleosome remodeling. Our findings of strong interactions between histone tails and nucleosomal DNA are in agreement with the work done by Ausio et al. (ADvH89), where the authors report that under physiologically relevant salt conditions ($\leq$ 0.7 M NaCl), nucleosomal tails play a significant role in maintaining the thermal stability of mononucleosomes. We attribute this effect to the strong electrostatic interactions between histone tails and DNA. In our simulations, the histone tails do not behave as random coils, in polymer theoretic sense (LGK78). However, we report the absence of extended secondary/tertiary structures for histone tails in our DMD simulation trajectories. Our simulations suggest that binding of elongated histone tails to nucleosomal DNA imparts structural stability to the tails. During chromatin condensation, the histone tails may adopt an ordered secondary structure and make structurally conserved interactions with adjacent nucleosomes.

In the context of long nucleosomal arrays, with a large conformational flexibility accessible to histone tails, the propensity to form internucleosomal histone-DNA interactions competes against intranucleosomal histone-DNA interactions. This equilibrium between intra- and internucleosomal interactions is attributed to the stabil-

ity during condensation of chromatin fibers. Studies of small-angle x-ray scattering (SAXS) for NCPs also support our results: SAXS data from Mangenot et al. (MLV$^+$02) shows salt-dependent binding of histone tails to neighboring nucleosomes. Zheng et al. (ZH03; ZH04; ZLHH05) investigated inter- and intranucleosomal interactions in a model dinucleosomal array and observed that upon salt-dependent folding and oligomerization of nucleosomes, H3 tail interactions reorganize to engage in primarily internucleosome interactions. Positively charged N-terminal tails of H3 histones are longest among all core histone tails, and therefore have significantly greater conformational flexibility available to form strong interactions with neighboring nucleosomes and its own nucleosomal DNA. Preferential binding to neighboring nucleosomes at higher salt concentrations is expected due to weakening of intranucleosome histone tail-DNA interactions. Predictions of Zheng et al. (ZH03; ZH04; ZLHH05) that alterations in H3 histone tail interactions may elaborate different structural and functional states of chromatin is also in agreement with our simulations: Low-temperature DMD simulation trajectories show stabilization of nucleosome upon binding of histone tails with nucleosomal DNA, while under higher-temperature conditions, thermal fluctuations hinder histone tails binding to DNA, leading to a higher energy conformation.

Evidently, below the folding transition temperature (i.e., $T = 0.5$-$0.875\ \varepsilon/k_B$), DMD simulations achieve equilibration by 4×10^4 t.u., whereas the simulations performed above the folding temperatures are equilibrated by 6.5×10^4 t.u. Subtle equilibrium between intramolecular interactions and entropic contributions is sufficient to fold a protein into its specific tertiary structure. The approach used in our model, i.e., DMD simulations of two-bead/residue models, has previously been applied successfully to study the protein-folding transition-state ensemble of the C-Src SH3 domain (DDB$^+$02a) and the amyloidogenesis mechanism for Src SH3 domain proteins (DDB$^+$02b). Similar DMD approaches have been extensively applied to successfully capture the essential elements of structural stability, encompassing important biological processes: pro-

tein folding, unfolding, and aggregation (DD05; PDU$^+$04; DBD05; UCD$^+$04; BDB$^+$04; DJD05; KDD03). Thus, we posit that our DMD model captures relevant elements of structural stability for the NCP. Coarse-graining the structural details of NCP results in underestimation of histone side-chain entropic contributions. Due to the enormous structural complexity of nucleosomes, an accurate estimation of side-chain entropy is unfeasible. The coarse-grained approximation in the DMD model allows us to sample longer-timescale conformations at the expense of structural detail. By generating an all-atom representation of corresponding coarse-grained trajectories (using the heavy-atom reconstruction method, cf. Materials and Methods), we can generate atomic-resolution trajectories for NCP, giving an estimate of the loss of entropic contributions in CG models.

CG models have integrated degrees of freedom in comparison to the corresponding all-atom models. We have used the coordinates of Cα and Cβ atoms as our beads for investigating the dynamics of corresponding histone residues. Hence, in our unified-atom model, the masses of these beads (mC) are independent of the nature of the histone residue. Accurate estimation of physical timescales from CG models is difficult (NLSK04). We have provided a method to scale the DMD-simulation timescales to experimental timescales (Materials and Methods). Due to coarse-graining of the system, intrabead fluctuations occurring at small timescales, such as hydrogen vibrations, are not manifested in DMD simulations. Notably, DMD simulations of the NCP enable us to sample longer conformational dynamics of NCP, accessing experimentally relevant timescales (BS01; LLBW05) with near atomic-resolution detail.

We postulate a functional role of cold sites in the nucleosomes. The interface between H2A-H4 and H2A-H4 β-sheets are found to be rich in clusters of cold sites, and make stable interactions throughout the DMD simulations. Santisteban et al. (SAMS97) showed that disruption of H2A-H4 and H2A-H4 β-sheet interactions by H4-Y98G mutants leads to disruption of H2A-H3-H4, H2A-H3-H4 molecular clusters and H4-H2B interactions,

thereby causing nucleosome dissociation. Wood et al. (WNL$^+$05) showed the presence of significant H2A-H4 and H2A-H4 β-sheet interactions in their high-resolution (1.9-Å) crystal structure of the histone octamer assembly. Based on solvent-accessibility analysis and residue conservation, they postulated that the region of histone octamer binding transcription elongation factors and other histone-binding compounds involved in transcription is present in the beta-sheet interaction region. Thus, studies by Santisteban et al. (SAMS97) and Wood et al. (WNL$^+$05) also support a functional role of H2A-H4, H2A-H4 cold sites in transcription elongation.

We also observe that the magnitude of local structural fluctuations in nucleosomal DNA is sequence-dependent in nature. Although the extent of these fluctuations increases monotonically with temperature; the sequence dependence is conserved across a wide range of temperatures (correlation coefficient of 0.55 for T = 0.1, 0.4, and 0.8 ε/k_B), as shown in Fig. 2.5. Local fluctuations in nucleosomal DNA vary significantly along the DNA sequence, suggesting that only a fraction of histone-DNA contacts make strong interactions and dominate the dynamics of nucleosomal DNA. This observation is in agreement with the observations by Luger et al. (LMR$^+$97), where x-ray crystal structure of the NCP shows 14 contact points between DNA and the histone.

Histone-DNA interactions have previously been studied using DNase I digestion (CIS$^+$04), protein-DNA cross-linking (MGA$^+$98; SLV88), and immunoprecipitation (KA99). In our constant-temperature DMD simulations, under low-temperature conditions (T = 0.1 ε/k_B), histone tails form few contacts with nucleosomal DNA, whereas at higher temperatures (T = 0.4-0.8 ε/k_B), we observe frequent contacts formed between the C-terminus of one H2A (chain C) and the dyad axis of nucleosomal DNA (Fig. 2.4). These results support earlier experimental work (UBGB94) where the authors used covalent protein-DNA cross-linking experiments to demonstrate that in the absence of linker DNA, the C-terminal domain of histone H2A contacts the dyad axis, and showed the ability of the H2A C-terminal domain to rearrange. These interactions of positively

charged histone tails with negatively charged nucleosomal DNA stabilizes the histone tails and its secondary structure may change from random coils to alpha-helices, which is consistent with previously reported results on a similar increase in α-helical content upon acetylation of histone tails (WMLA00). Other dominant histone-DNA interactions include contacts with H2B histone. Under very high temperature conditions ($T = 1.2$ ε/k_B), the histone octamer assembly is unfolded and the DNA basepairing is lost at the termini. We find that the frequency of histone-DNA contacts found in the native state is significantly reduced, with contacts largely interspersed. We expect that transient histone-DNA hydrogen-bond interactions which have low contact frequency in DMD simulations will be weaker and contribute less to nucleosome stability.

The elastic nature of free DNA has been characterized using several biophysical experiments (SFB92; SCB96; CW04) and theoretical models (MS02; MS95; LT99; LT04; BSD06). Bending properties of DNA have been extensively studied for prokaryotic (PMd97) as well as eukaryotic (WM00; Dic98) cells. Protein-induced DNA bending (OGL$^+$98) is shown to be necessary for transcription activation (BE99). It is proposed that the intrinsic curvature and flexibility of nucleosomal DNA mediates nucleosome stability (ABD$^+$99). In our simulations, we observe that due to the preferential attraction of DNA strands toward core histones, significant bending deformations are observed in nucleosomal DNA. Strong electrostatic interactions with histones stabilize the bent state of DNA, and this conformation persists throughout the simulation. However, in mononucleosomes, elongated conformations are preferred for base-pairings present at the ends of nucleosomal DNA. Stopped-flow FRET experiments demonstrate spontaneous unwrapping of nucleosomal DNA (LLBW05; LW04). Recent work by Li et al. (LLBW05) demonstrates rapid rates for unwrapping ($\approx$4 s^{-1}) and rewrapping (20-90 s^{-1}) of nucleosomal DNA from the histone octamer assembly. Our simulations suggest that for nucleosome remodeling, the resulting rate-limiting step of nucleosomal DNA unwrapping is mediated in part by spontaneous disruption of these interactions between

histone tails, core histones and nucleosomal DNA. The kinetics of nucleosomal transcription, which occurs at longer timescales (≈1.4 kb/min (SO91)), is limited by the rate of DNA rewrapping (LLBW05). The aggregate number of local contacts formed between histone tails and nucleosomal DNA, which enhance DNA rewrapping, may mediate the kinetics of transcription at nucleosome-rich DNA fragments.

CG models using effective potentials are also known to reproduce gyration-radius and distribution functions of constituent CG variables over a wide range of temperatures (FTD02). Our approach of performing DMD simulations on near-atomic-resolution CG models of the NCP, with heavy-atom trajectory reconstruction, may be extended to simulations of naturally occurring variant nucleosomes (having a Cse4-containing H3-variant or Htz1-containing H2A-variant) for deciphering the functionality of histone variants (MH03; FRLT04; AH02a; AH02b). Our multiscale modeling methodology is also applicable to exploring dynamics of dinucleosomes, and other higher-order nucleosomal arrays having linker histones to explore histone tail modifications (WHMA01), and dynamics of linker DNA (vZ96) on the stability of NCP and the higher-order organization of chromatin structure. Existing approaches for modeling the NCP using coarse-grained electrostatic models (BS01; SZS05) and Monte Carlo simulations of the chromatosome particle (KBO00) have been successful in predicting gross chromatin dynamics. All-atom simulations of simplified nucleosome models lacking histone tails (BZ02; Bis05) have yielded important insights into higher-order chromatin organization. However, the presence of histone tails is critical in ascertaining the structural organization of chromatin fibers (JA01; SA00). Complementing models of higher-order chromatin dynamics with our higher-resolution DMD simulations based on NCP structural models, a detailed insight on large-scale chromatin structure and dynamics is accessible.

2.5 Conclusions

In summary, using DMD simulations, we show that our simplistic model recapitulates the stability and simulates the dynamics of NCP for experimentally relevant timescales. We find that in our simulations of mononucleosomes, histone tails form strong salt-bridge interactions with nucleosomal DNA, that suggests their direct role in forming higher-order chromatin structure. Based on constant-temperature discrete molecular dynamics simulations, we find that bending across the H3-H3 interface is a prominent mode of nucleosome dynamics. The dynamics of the NCP is dominated by histone tails with subsequent normal modes composed of large-scale interhistone motions. Analysis of frequencies of histone-DNA contacts formed in constant-temperature DMD simulations shows persistent contacts formed with C-terminal H2A and the nucleosomal dyad axis, thereby suggesting functional roles of the H2A C-terminal domain. We determine a coarse-grained phase space of the NCP under altering potentials of histone-DNA interactions. Our approach of amalgamating rapid conformation sampling techniques like DMD with coarse-grained models of nucleosome may be useful for analyzing the effects of histone variants and the effects of DNA sequence on nucleosome positioning.

Chapter 3

Exploring core histone residues essential to nucleosome stability

3.1 Introduction

Chromatin structure controls accessibility to genomic DNA and thereby regulates all DNA-templated processes such as transcription, replication, DNA repair and recombination. An accurate understanding of the processes involved in regulation of chromatin structure and the mechanism of DNA accessibility is of fundamental importance in molecular biology, as providing mechanistic insights into chromatin-associated processes will advance our understanding of human biology and diseases, including cancer (HB01). The nucleosome, composed of DNA wrapped around histone proteins, is the basic unit of eukaryotic chromatin. The nucleosome consists of two copies of each of the four histone proteins: H2A, H2B, H3, and H4 forming the histone octamer core. Nearly 147 base pairs of eukaryotic DNA are wrapped around each histone octamer to form the nucleosome core particle (NCP).

The crystal structure of *Xenopus laevis* and *Saccharomyces cerevisiae* nucleosome core particle are available at 1.9 Å (DSL$^+$02) and 3.1 Å (WSL01), respectively. However, the effects of dynamics of core histones and nucleosomal DNA on macromolecular thermodynamic stability are still poorly understood. Discrete molecular dynamics (DMD)

simulations of the *Xenopus laevis* NCP have suggested a functional role of conserved core histone interactions (SDD07). Over the course of evolution, the primary sequence of core histones proteins has remained highly conserved, especially for H3 and H4 histones (MF80). While covalent modifications to the histone tails are known to regulate chromatin organization and function (SA00), the necessity of a conserved histone core residues in contributing to the stability of nucleosomes and the organizational state of chromatin is poorly understood. Here, we probe the role of *Saccharomyces cerevisiae* H3 core histone residues in mediating chromatin stability both in silico by means of computational design of single point mutations, and in vivo using a yeast-based viability assay. We report that cold sites in the nucleosome core, which are predicted to mediate nucleosomal stability, are sensitive to point mutations and result in cell lethality, while hot site mutations, which are predicted to further enhance nucleosome stability, exhibit a viable phenotype. Our results suggest distinct roles for core histone residues in mediating chromatin stability.

3.2 Discrete molecular dynamics simulations of nucleosomes

We used equilibrium DMD simulations to explore the dynamics of *Xenopus laevis* mononucleosomes (SDD07). These simulations reveal a set of core histone residues, termed cold sites, exhibiting persistent interactions with adjoining cold site histone residues in constant temperature DMD simulations (trajectory-normalized contact frequency greater than 0.7). We observe that the cold sites are clustered in the H3-H3, H4-H2A inter-histone surfaces. Further, the results also suggested that the cold site interactions are essential for regulating the stability of the NCP, thereby mediating chromatin states (SDD07). To further examine the importance of histone cold sites, we used the Eris protocol (YDD07) with Medusa force-field (DD06) to estimate the thermodynamic stability

of point mutations in the *Saccharomyces cerevisiae* H3 core histone cold sites.

Cold sites in the *Xenopus laevis* nucleosome core particle were identified in Sharma et al (SDD07). Cold sites for the *Saccharomyces cerevisiae* nucleosome are identified using constant temperature discrete molecular dynamics simulation of the corresponding *Saccharomyces cerevisiae* NCP (Protein DataBank: 1id3). In contrast to the crystal structure of *Xenopus laevis* nucleosome (PDB: 1kx5), histone tail structure is absent in the crystal structure of the *Saccharomyces cerevisiae* NCP (PDB: 1id3). Insight II molecular modeling software (http://www.accelrys.com) was used to model the structure of missing histone residues in the N-terminal histone tails in the *Saccharomyces cerevisiae* NCP.

Despite the lack of tertiary structure in the core histone tails, the *Saccharomyces cerevisiae* and *Xenopus laevis* NCP core histone folds have exceedingly high homology in the tertiary structure of core octamer, where cold sites are present (Root Mean Square Deviation of ≤ 0.45Å between Protein DataBank 1id3 chain A and Protein DataBank 1kx5 chain A, computed using PyMOL).

We designed point mutations in the H3 cold site loci disrupting persistent intra-nucleosomal salt-bridge interactions, thereby causing a loss in nucleosome stability. Specifically, point mutations H3:H113A, H3:A114Y, H3:L130A and H3:L126A were predicted to have a significant loss of cold site interactions (Fig. 3.1). These predicted mutations were investigated *in vivo* by testing the viability of cells with mutant nucleosomes. Each of the predicted destabilizing cold-site mutations was found to exhibit an unviable, no-growth phenotype; while the control, wild type nucleosome exhibit a viable, normal growth phenotype.

Furthermore, we investigate hot sites in the core nucleosomal histones, wherein mutations are predicted to enhance the stability of nucleosomes. We explored the thermodynamic stability of all possible point mutations in the *Saccharomyces cerevisiae* H3 core histone using Medusa (DD06). Specifically, we predicted mutation P67T and F55S

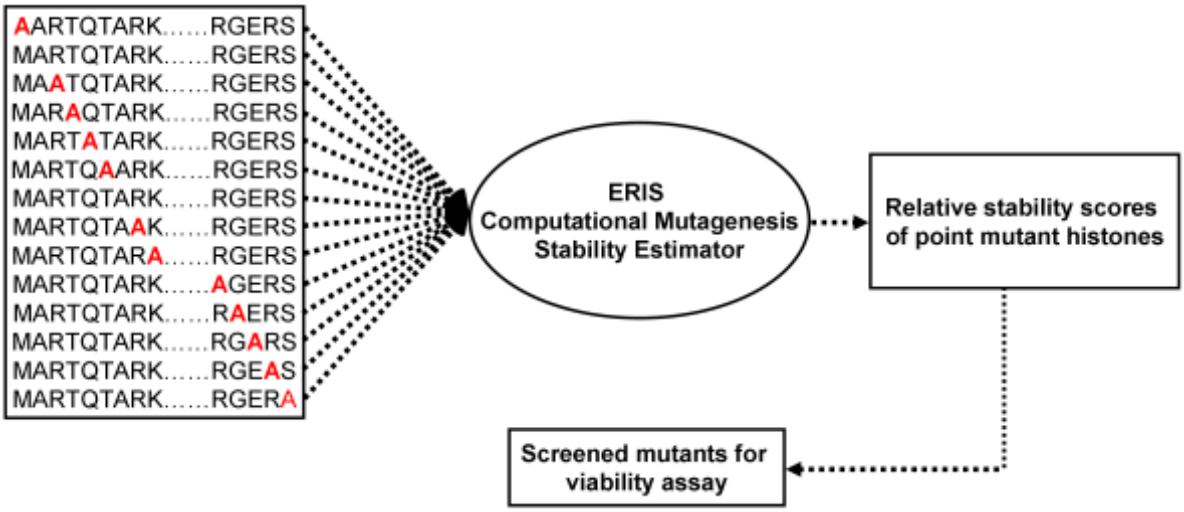

Figure 3.1: Schema of the *Saccharomyces cerevisiae* H3 histone mutant screening protocol. The Eris high-throughput protein mutagenesis stability estimator was used to computationally score all possible mutations in the *Saccharomyces cerevisiae* H3 histone. Point mutations were screened which lead to greatest loss in relative stability of the H3 core histone relative to the wild type H3 histone.

in H3 histone to enhance the stability of the wild-type nucleosome core via novel side-chain mediated salt-bridge interactions. The 5-FOA plating assay confirmed that the H3:P67T hot-site mutation as well as the wild-type nucleosome have a normal growth phenotype, while each of the cold-site mutation exhibits no-growth phenotype (Fig. 3.2).

3.3 Materials and Methods

3.3.1 Nucleosome mutation selection

Mutations identified as disrupting cold site interactions in histone H3 that were predicted to result in significant loss in stability of the nucleosome core particle are as follows: H113A, A114Y, L130A and L126A. We used the Medusa computational protein design toolkit (DD06) and the Eris protocol (YDD07) for high-throughput screening of all possible mutations in H3 histones residues in the nucleosome core particle. Medusa uses a physical force field with an atomic resolution model to predict thermodynamic

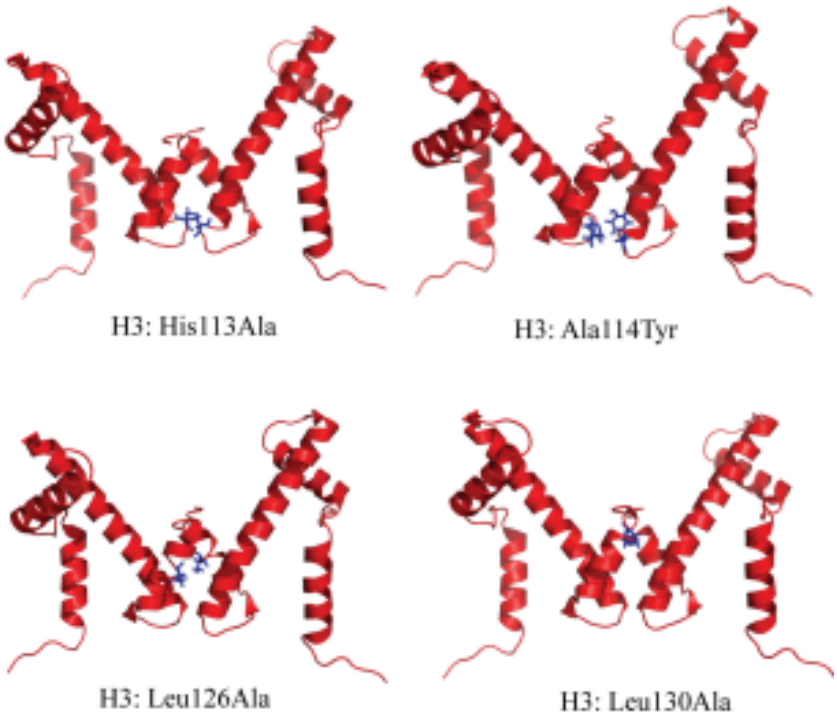

Figure 3.2: Structure of mutant H3 histone dimers disrupting cold-site interactions. Point mutations His113Ala, Ala114Tyr, Leu126Ala Leu130Ala in *Saccharomyces cerevisiae* H3 histone were found to disrupt cold-sites interactions of corresponding H3 histone residues.

stability ($\Delta\Delta G$) of point mutations (YDD07). Medusa gave relative scores of change in free energy imparted by each point mutation in the H3 core. Point mutations P67T and F55S in the H3 core histone result in greatest stability to the nucleosome core particle. These mutations were selected as hot site mutations in the H3 core histones.

3.3.2 Yeast strains, plasmids and histone shuffling

The *Saccharomyces cerevisiae* histone H3-H4 shuffle strain WZY42 (ZBE$^+$98), which harbors the deletions of both gene copies of H3 and H4 while maintaining a wild type (WT) copy of H3 and H4 (copy II: HHT2-HHF2) on a plasmid, was cultured in either rich YPD media or selective SC media as needed. This strain was used to select for cells containing site-directed mutations of H3 (see below) that were created on a H3 and H4 plasmid bearing the Trp marker. Colonies from the SC-Trp plates were streaked onto 5-FOA-conatining plates to select for cells containing only the mutated Trp plasmid.

3.3.3 Site-directed mutagenesis

Site-Directed Mutagenesis was carried out as described by Stratagene's QuikChange Site-Directed Mutagenesis Kit with the exception of omitting the mineral oil overlay and allowing the DpnI restriction enzyme to digest the parental super coiled double stranded DNA for up to 6 hours. Oligonucleotide primers were designed by making the desired mutations to the wild-type *Saccharomyces cerevisiae* gene sequence of H3 found in the Saccharomyces Genome Database sequence for HHT2/YNL031C and ordered from Operon Biotechnologies, Inc. All point mutations were introduced into a H3/H4 Trp+ plasmid (ZBE$^+$98) that was sequenced for accuracy and then transformed into WZY42 using the One-Step Yeast transformation protocol by Chen et al. (CYK92).

3.4 Results

We tested mutations rescuing cold-site interactions in mutant nucleosomes by introducing a second mutation in addition to the original cold site mutation that would compensate for the salt bridge interactions disrupted by the cold site mutations. A114Y+H113G, H113A+D123E, L130A+L126TY and L126A+R129E were identified as the double mutations compensating the lost cold site interactions. We predicted viability for these rescue mutants in H3 histone. However, the combination of the original cold site and new rescue mutation was found to be lethal to *Saccharomyces cerevisiae* cells (data not shown). This result suggests that nucleosome stability is very sensitive to the specific inter-histone interactions and may not be restored by incorporating a second rescue mutation. Hot-site mutations that are predicted to further stabilize the nucleosome included H3:P67T and F55S. The hot-site mutant H3:P67T showed viability while the H3:F55S mutation was lethal. We tested the cells consisting of the viable H3:P67T mutant histone for heat sensitivity at 37 °C, however the P67T mutant shows no-growth defect. The viability of H3:P67T mutant H3 histone and lethality of H3:H113A, H3:A114Y, H3:L130A and H3:L126A cold site mutations are in agreement with the proposed functional role of cold sites. However, the lack of growth phenotype and viability of H3:F55S necessitate an accurate understanding of the effects of these mutations on chromatin-associated cellular processes, such as DNA replication and repair which potentially influence cellular viability.

3.5 Discussion

Nakanishi et al. (NSD$^+$08) used site-directed mutagenesis to develop the Scanning histone mutagenesis with alanine (SHIMA) library of alanine point mutations at all residue positions in the *Saccharomyces cerevisiae* core histone proteins. The SHIMA library suggested the following set of core histone residues to be essential for cellular

viability: Histone H3: Tyr41, Leu48, Ile51, Gln55, Glu97, His113, Arg116, Thr118, and Asp123. Comprehending the function of conserved core histones is of significant interest in chromatin biology. In this work, we explore the role of dynamically conserved residues in stabilizing intra-nucleosomal interactions. Mutations at cold site residues result in disruption of salt-bridge interactions within the core histone octamer. Here, we demonstrate that mutations at specific cold site residues lead to cellular lethality. This work underscores the importance of cold site residues in stabilizing chromatin *in vivo*.

Chapter 4

Homology modeling of the Saccharomyces cerevisiae centromeric nucleosome

4.1 Introduction

The kinetochore is the protein-DNA complex at eukaryotic centromeres that functions as the attachment site for spindle microtubules. In budding yeast, the centromere spans 120 bp, there is a single microtubule per kinetochore, and the entire spindle is composed of 16 kinetochore microtubules plus four interpolar microtubules from each pole. There are > 65 different proteins at the kinetochore, organized in at least six core multimeric complexes (MTS03). A spindle checkpoint network monitors the state of attachment and tension between the microtubule and chromosome. We present a model for the path of DNA in the kinetochore.

Replicated sister centromeres become maximally separated by 600-800 nm in metaphase (PMSB01). Separation progressively decreases along chromosome arms such that sister chromatids are tightly juxtaposed at 10 kb from the centromere (PMSB01). The molecular glue linking sister chromatids, cohesin, is recruited to a 20-50 kb region surrounding the centromere at 3- to 5-fold higher levels than centromere-distal locations (BK99). A major paradox is the accumulation of cohesin at regions of separated sister DNA strands. A second problem is the nature of the mechanical linkage coupling DNA

to a dynamic microtubule plus-end. This linkage must resist detachment by mitotic forces while sliding along the polymerizing and depolymerizing microtubule lattice.

4.2 Materials and methods

4.2.1 Homology modeling of Cse4-containing nucleosome

A structural model of the centromeric nucleosome core particle was determined by substituting H3 histones in the known crystal structure of the S. cerevisiae nucleosome core particle (WSL01), with a homology model ($SPY^{+}95$) of Cse4, and replacing the palindromic 146 bp nucleosomal DNA sequence with centromeric DNA sequence (KFH00). Cse4 was modeled by combining homology models of its histone-fold domain (HFD, residues 132-229) and essential N-terminal domain (END, residues 1-66). In the modeled Cse4 variant nucleosome structure, basic lysine and arginine residues in the N-terminal domain form globular structures and interact with negatively-charged DNA backbone phosphates present at the junction between linker and nucleosomal DNA (Fig. 4.1). This is a major divergence from nucleosome core particle with histone H3, where the N-terminal histone tails are largely unstructured random coils. The highly-charged Cse4 tails are clustered at the exit and entry sites of the nucleosome, where they may restrict the mobility of Cse4 nucleosomes, as well as promote intramolecular cohesion by bending linker DNA. The Cse4 nucleosome is unique in that it is flanked by very broad (30-50 bp) hypersensitive nuclease cleavage sites (BC82). This linker DNA is susceptible to nucleolytic attack in the transition zone between the Cse4-containing nucleosome and H3-containing chromatin. Thus, Cse4 may stabilize the nucleosome core and direct the path of the DNA as it enters and exits the nucleosome. The Cse4 nucleosome represents a physiochemical interruption in chromatin organization, contributing to the unique assembly features dictated by this chromosomal locus. The model structures of Cse4 and centromeric nucleosome core particle have been deposited in the protein data bank at

http://www.rcsb.org (PDB ID code 2FSB and 2FSC, respectively).

4.3 C-loop model of Saccharomyces cerevisiae centromere

We propose that pericentric chromatin is held together via intramolecular cohesion (Fig. 4.2), similar to a foldback structure proposed for the fission yeast centromere (PA05). In contrast to fission yeast, the budding yeast core centromere (120 bp DNA wrapped around a specialized nucleosome containing two molecules of the centromere-specific histone H3 variant, Cse4) and flanking chromatin may adopt a cruciform configuration in metaphase.

Centromeric DNA is sharply bent around the Cse4 nucleosome by the CBF3 protein complex (PTH$^+$99), forming the apex of the putative centromere-loop (C-loop). The C-loop would be approximately 22 nm in diameter (twice the diameter of a nucleosome) and held together through intramolecular cohesin bridges (Fig. 4.2). To account for the measured distance between replicated sister centromeres, a transition zone 7-8 kb from the centromere-specific nucleosome marks the conversion from intra- to inter-molecular bridges. 7-8 kb of DNA wound 1.65 times around the histone octamer is approximately 300-400 nm long (2.3 μm of B-form DNA, or 7- fold nucleosomal compaction). The proposed intramolecular linkage is therefore consistent with the appearance of separated centromeres, the apposition of DNA markers 10 kb from the centromere, and the increased concentration of cohesin at the centromere. Two alternative forms of cohesin have recently been proposed (HMK05), perhaps reflecting the different substrates dictated by centromere-flanking chromatin vs. chromosome arms.

The budding yeast centromere is unique in having a single Cse4-containing nucleosome (MYG$^+$98). We derived a structural model of the centromeric nucleosome to evaluate whether the path of DNA around the nucleosome core particle is compatible

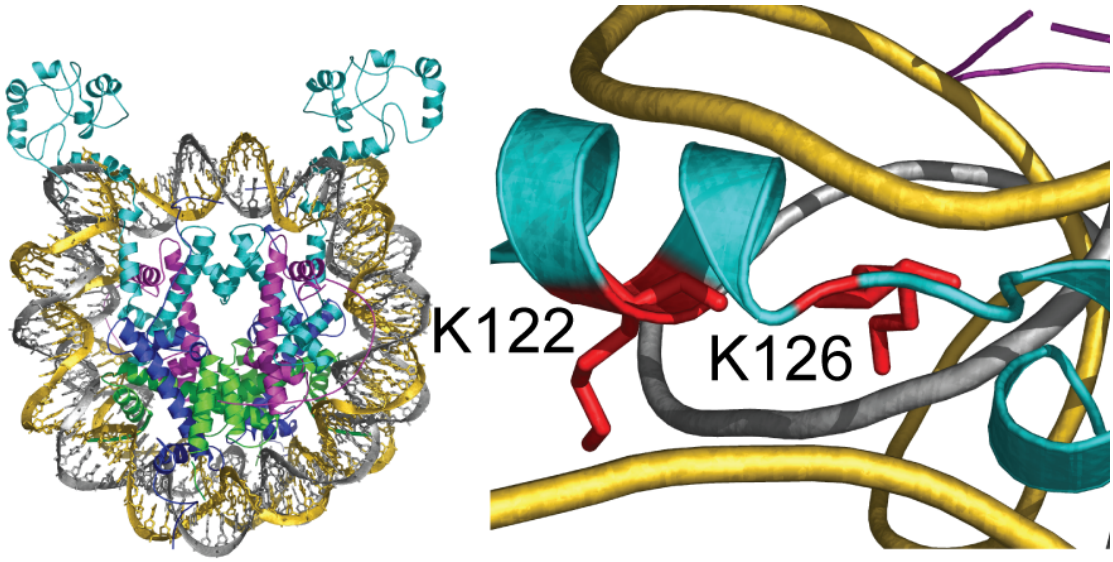

Figure 4.1: A structural model of the Cse4 nucleosome. (Left) Cse4 (cyan), DNA (gold/silver), and histones H2A (blue), H2B (green), H4 (pink), are shown. Cse4 is modeled by combining homology models of the histone-fold domain (HFD, residues 132-229) and essential N-terminal domain (END, residues 1-66). Sequence alignments of the two domains are based on the consensus alignment predicted by the 3D-Jury Structure Prediction meta-server. 3D-Jury predicts HFD, END domains have most significant homology with histone H3 (PDB: 1ID3-A) and the N-terminus of 1,6 phosphofructokinase (PDB: 1BIF), respectively, and no significant homolog is predicted for the intermediate region (residues 67-131). The combined Cse4 model was energy-minimized using the molecular dynamics simulation procedure of the program Insight-II (Accelrys Software Inc). Evaluation of the resulting Cse4 model using Verify3d indicated that molecular geometry and stereochemistry of the Cse4 model is of good quality. The histone-fold domain of Cse4 is structurally superimposed with the known crystal structure of H3 histones present in S. cerevisiae nucleosome core particle (PDB: 1ID3-A, 1ID3-E) to yield the model structure of centromeric nucleosome core particle. (Right) Zoomed-in view of Cse4 interactions with centromeric DNA. Lys122 and Lys126 (red) of modeled Cse4 interact with the termini of nucleosomal DNA (gold/silver).

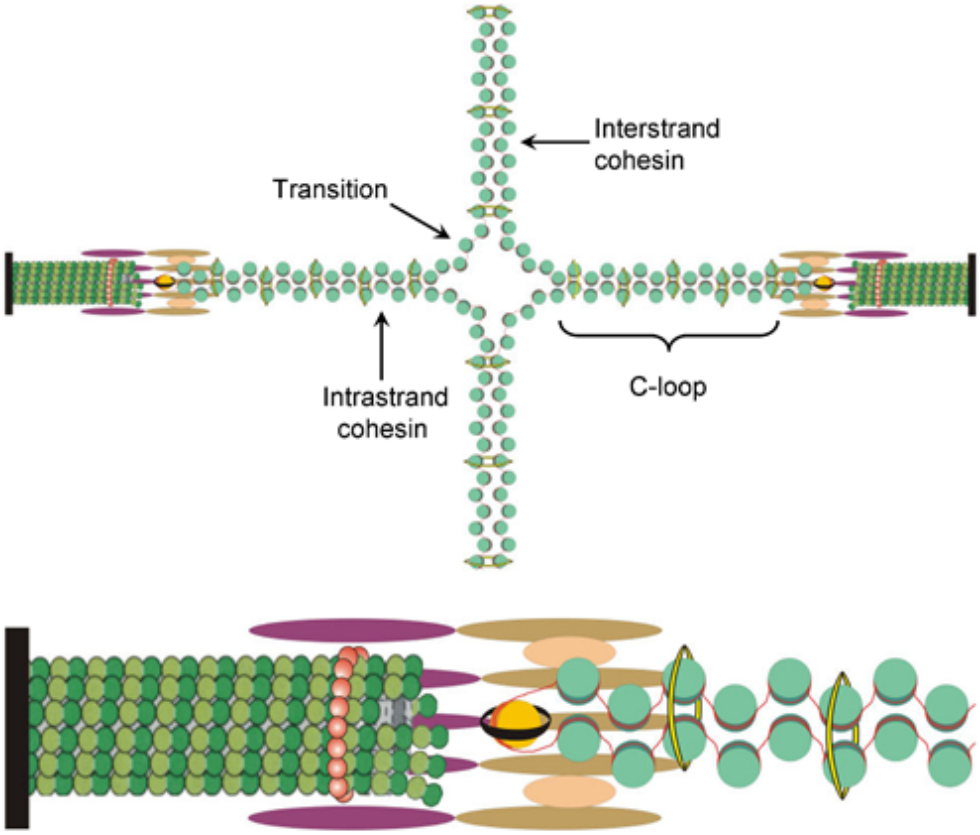

Figure 4.2: Proposed structure for centromere DNA in the kinetochore. (Top) Bi-oriented sister chromatids adopt a cruciform structure. Centromere-flanking chromatin is held together by intrastrand cohesin bridges, and chromosome arms by interstrand cohesin bridges. The transition between these two regions in budding yeast is mobile and on average 7 kb from the centromere core. (Bottom) The Cse4-containing nucleosome (orange circle) and flanking nucleosomes (green circles) are proximal to the microtubule plus-end. The microtubule (left) is encompassed by the Dam1 ring (pink) (MDSH05; WvSW^{+}05) and elongated Ndc80 rods (purple) (WSH05). Binding of CBF3 complex (black), bends centromere DNA $\approx 55°$ (PTH^{+}99), forming a C-loop of chromatin held together by intrastrand cohesin (yellow rings). Additional kinetochore complexes (Coma and Mind in tan and blue, respectively) are proposed to link CBF3 and the C-loop to Ndc80, Dam1, and other linker complexes at the microtubule plus-end.

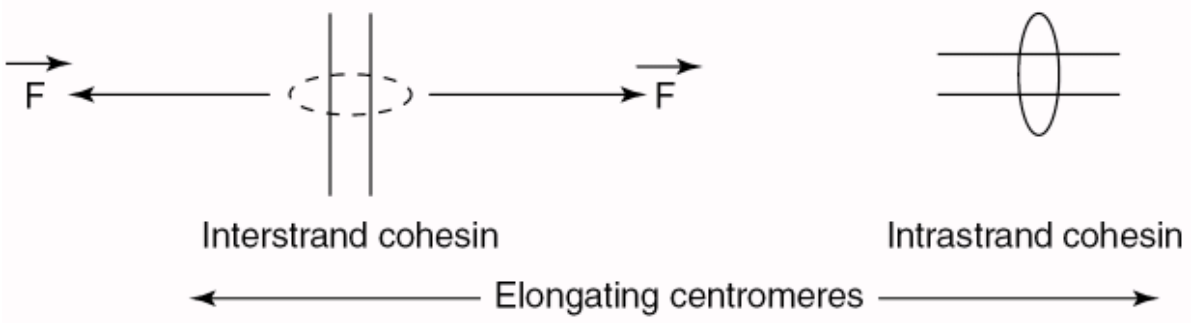

Figure 4.3: Comparison of interstrand and intrastrand cohesin forces. Interstrand cohesin forces act parallel to the microtubule axis and perform work when Cse4-containing nucleosomes move away from the chromosomal axis. Intrastrand cohesin forces act perpendicular to the direction of movement of Cse4-containing nucleosomes, and do not perform work. Intrastrand cohesin progressively clamps elongating intrachromatid pairs while interstrand cohesin rings supply the restoring force opposing microtubule-based forces.

with the C-loop (Fig. 4.3). The highly charged Cse4 tails are clustered at the exit and entry sites of the nucleosome, where they may restrict the mobility of the nucleosome as well as promote intramolecular cohesion by bending linker DNA. Thus Cse4, together with CBF3, may stabilize the nucleosome core and direct the path of the DNA as it enters and exits the nucleosome.

4.4 Qualitative estimates of forces on the C-loop

Cohesin rings have a large diameter (≈ 400Å) relative to mononucleosomes (≈ 100Å). We hypothesize that interactions between spatially separated inter-chromatid strands are dominated by strong salt-bridges formed between basic (lysine/arginine) side-chains of elongated histone tails and phosphate (DNA) of the sister chromatid strands. We have modeled the distribution of inter vs. intrastrand cohesion by the increase in these stable salt-bridge contacts formed in the C-loop (Fig. 4.4). Fluctuations in the number of effective inter-chromatid interactions (σ_n) along the chromosome arm axis result in fluctuations in separation between the ends proximal to microtubules, σ. As intrastrand cohesins accumulate, they progressively tether flanking nucleosomes and fa-

cilitate elongation of the C-loop. In the course of this elongation, the work done by microtubule-based forces (F) is $F\sigma$ and the loss of free energy due to disruption of inter-chromatid salt bridges is $\sigma_n \Delta G_{sb}$ where ΔG_{sb} is the free energy of breaking a single inter-chromatid salt bridge. The work done by microtubule forces must compensate for the loss of favorable salt bridges and the energy associated with thermal fluctuations: $F = \Delta n \Delta G_{sb} + k_B T$ However, fluctuations in centromere separation caused by microtubule forces Fσ far exceed the thermal fluctuations $k_B T$: $\sigma \approx \sigma_n \Delta G_{sb}/F$. Assuming each nucleosome in the 10-kb stretch around the centromere makes a single dominant salt bridge with the sister chromatid, then Δn is ≈100 (10 kb/200 bp per nucleosome x 2 = 100). Surface salt-bridge interactions between lysine-rich histone tails, as well as ΔG_{sb}, depend on salt concentration (TLW04), and are on the order of 10-60 kcal/mol for DNA-integration host factor complexes (HTSR01). If we estimate ΔG_{sb} for inter-chromatid interactions as 20 kcal/mol, our model predicts that a F ≈ 20 pN force will lead to large positional fluctuation ($z \approx 500$ nm) of the distal end of the C-loop relative to the chromosome axis. A single microtubule has been estimated to generate at least 10 pN of force (Nic83; GMAM05). Thus, the range of force generated by the microtubule is on the order of that required to alter the position of the transition zone, and hence the distance between ends of the C-loop.

4.4.1 Predictions from centromeric nucleosome model

This model predicts that Cse4 (a CENP-A homolog) is proximal to the microtubule plus-end. The CENP-A homologs in D. melanogaster (CID) and C. elegans (HCP-3) face poleward on the mitotic chromosome 8 and 9 (HMAM01). However, unlike a single Cse4-containing nucleosome in budding yeast, CENP-A nucleosomes are interspersed with blocks of histone-H3 nucleosomes (BSK02). The degree of DNA bending as DNA enters and exits the canonical CENP-A nucleosome (Fig. 4.1) may dictate whether single or multiple CENP-A nucleosomes comprise the kinetochore. CENP-A is highly divergent

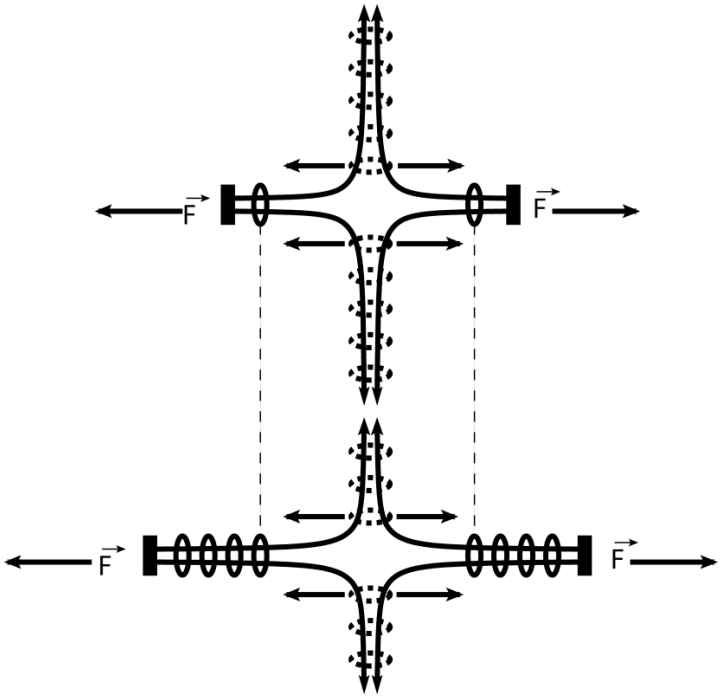

Figure 4.4: Positional instability of the C-loop. We propose that cohesins (rings) form complexes on sister chromatids in both lateral (interstrand) and longitudinal (intrastrand) directions relative to the direction of microtubule-based forces (arrows). Intrastrand cohesins clamp the C-loop, facilitating its elongation and movement of centromere ends (black rectangles). Forces (F vector displacement) from attached microtubules act predominantly in the lateral direction, destabilizing interchromatid cohesins. Fluorescence imaging techniques demonstrate that centromere reassociation during mitosis is infrequent (v = 0.4% of experiment time) and predict elongation of the ends proximal to microtubules with a velocity of $\approx 1\ \mu m/min$. This observation suggests that the dynamic equilibrium between disruption of interchromatid cohesion and formation of intrachromatid cohesin tethers is shifted towards the latter process (compare schematics above and below), and predicts a higher density of intrachromatid cohesin bridges along the flanking chromatin. Addition of intrastrand cohesins progressively tethers the region between transition zone and centromere ends, thereby facilitating lateral elongation of centromere ends. The cohesion-free region is fluctuating around the same mean value (denoted by dashed lines), governed by the balance between the cohesive forces and microtubule-induced forces. From the frequency and magnitude of separation previously observed between sister kinetochores in live cells, we estimate the stabilization caused by the kinetics of conversion of inter- to intra-chromatid cohesin tethers as $\Delta\Delta G_{elong}$ = -RT ln(l/v) $\approx$ -3.5 kcal mol^{-1}. We expect the actual stabilization to be larger than $\Delta\Delta G_{elong}$ due to the limits in resolution of fluorescence microscopy. Thus, centromeric cohesin complexes may have a direct functional role in stabilizing the elongating centromere instead of producing an opposing force against pulling by microtubules. Upon loss of force the sister centromeres are predicted to return to the lowest free energy state, that of interstrand sister chromatid cohesion.

(HAM01), indicating potential changes in its molecular architecture in different species.

Several specialized chromosome domains are organized into loop structures, including the T-loop of telomere DNA (WSd04) and the DNA loops that characterize lampbrush chromosomes. Evidence for a centromeric DNA loop can be found in a deletion analysis of dicentric chromosomes ((KBB$^+$94) and J.A. Brock, unpublished observations), which undergo a breakage-fusion-bridge cycle leading to chromosome rearrangements, with the predominant outcome of losing an entire centromere and flanking DNA. Deletions arising from two DNA double-strand breaks within the C-loop are consistent with these findings. Thus, similar to the 8-kb deletion blocks of T-loops at the telomere (WSd04), *in vivo* deletions that remove large domains of one centromere from dicentric chromosomes are indicative of loss of a structural element.

4.5 Results and discussions

A corollary of the model is that the tip of the C-loop may be mobile relative to the chromosome axis (Fig. 4.4). A change in the position of the transition zone relative to the centromere-specific nucleosome will alter the position of the C-loop's distal end. The C-loop tip will migrate toward the transition zone tip as interstrand cohesion is favored, and away from the transition zone as intrastrand cohesin is favored. The range of force generated by the microtubule is on the order of that required to alter the transition zone position and hence the spatial position of the C-loop (Fig. 4.4). We predict that change in the position of the C-loop tip will coincide with change in the position of kinetochore microtubule plus-ends (GPS$^+$05). Thus, while the mechanisms are completely different, both 'ends' of the C-loop and the kinetochore microtubule are dynamic, a feature of the kinetochore that may contribute to the tension-based surveillance system.

Inducing mammalian cells to enter mitosis with unreplicated genomes has allowed dissection of the kinetochore's subunit structure (ZMB91). Each of the 25-30 microtubule-

binding sites in a mammalian kinetochore can be detached from the chromosome and still maintain an autonomous structure that includes DNA (ZMB91). These data suggest the mammalian kinetochore is comprised of a repeating DNA-protein structural unit that is autonomous in its ability to form a C-loop and bind single or multiple microtubules. The C-loop may insert into a cylindrical kinetochore structure that encompasses both DNA and the microtubule. The C-loop predicted by our model in S. cerevisiae would thus represent the fundamental unit of the kinetochore across phylogeny.

Chapter 5

Ab initio RNA structure prediction using discrete molecular dynamics

5.1 Introduction

The central dogma of molecular biology ascribed fundamental importance to RNA molecules in transcription and translation. Both coding and noncoding RNA molecules are now known to possess much greater variety of biological functions (Edd01; HS06) than what was suggested by the central dogma. During the last two decades, significant developments have lead to new insights in the importance of RNA in many post-transcriptional and post-translational processes. Discoveries of ribozymes and a variety of small RNAs with novel biological functions have highlighted RNA as a ubiquitous molecule in cellular processes (DD01). To perform their biological functions, many RNA molecules adopt well-defined tertiary structures. The RNA conformational dynamics determines how often these functionally important conformations appear in the course of RNA's life and, therefore, modulate its functional activity. Hence, there is a rejuvenated interest in accurate *ab initio* prediction of three-dimensional (3D) structure and dynamics of RNAs (SYKB07).

Currently, RNA folding tools are mainly focused on predicting RNA secondary structure (Mat06). Computational tools for RNA secondary structure prediction, such as

Mfold (Zuk03) and Vienna RNA (Hof03), are successful in predicting the RNA base pairing loci, thereby predicting the secondary structure organization. Using a dynamic programming approach (Edd04), secondary structures are inferred by scoring nearest-neighbor stacking interactions with adjacent base pairs (Mat06). However, these analyses based on base-pairing and base-stacking interactions ignore 3D steric hindrances in scoring putative secondary structures of RNA. The explicit modeling of the 3D structure might prohibit unfeasible tertiary structures of RNA. Cao and Chen designed a simplified diamond-lattice model for predicting folded structure and thermodynamics of RNA pseudoknots (CC06). This approach quantitatively predicts the free energy landscape for sequence-dependent folding of RNA pseudoknots, in agreement with experimental observations (CC06). However, due to the lattice constraints and the dynamic issues associated with predefined Monte Carlo moves (Bau87), this approach is inadequate to study the folding dynamics of RNAs. Several other computational tools were developed for RNA 3D structure prediction (for review, see (SYKB07)). These methods either use comparative modeling of RNA sequences with known structures or utilize known secondary and tertiary structural information from experiments in interactive modeling (MTG$^+$91; MGC93; SYKB07). Therefore, novel automated computational tools are required to accurately predict the tertiary structure and dynamics of RNA molecules. Recently developed knowledge-based approaches using assembly of trinucleotide torsion-angle libraries (DB07) are successful in predicting RNA structures for small globular RNA fragments ($\leq$30 nucleotides [nt]). However, RNA molecules often do not adopt globular topologies, such as the L-shaped tRNA. Enhanced prediction accuracy for longer RNA molecules is attainable by using physically principled energy functions and using an accurate sampling of RNA conformations.

Here, we introduce a discrete molecular dynamics (DMD) (DD05) approach toward *ab initio* 3D RNA structure predictions and characterization of RNA folding dynamics using simplified structural models. In contrast to the traditional molecular dynamics

simulations, which are computation-intensive and hence expensive in probing RNA folding dynamics over long time scales, the DMD algorithm provides rapid conformational sampling (DD05). It is demonstrated in numerous studies that the DMD method is suitable for studying various properties of protein folding (CDN$^+$08) and protein aggregation (DD05), and for probing different biomolecular mechanisms (DD05; SDN$^+$06; SDD07). Here, we extend this methodology to the RNA folding problem. We simplify the RNA structural model by using a bead-on-a-string model polymer with three coarse-grained beads: phosphate, sugar, and base, representing each nucleotide (see Materials and Methods; Fig. 5.1). We include the base-pairing, base-stacking, and hydrophobic interactions, the parameters of which are obtained from experiments. The coarse-grained nature of the model, as well as the efficiency of the conformational sampling algorithm, enables us to rapidly explore the possible conformational space of RNA molecules.

5.2 Materials and Methods

5.2.1 Discrete molecular dynamics

A detailed description of the DMD algorithm can be found elsewhere (DBSS98). Briefly, interatomic interactions in DMD are governed by stepwise potential functions. Neighboring interactions, such as bonds, bond angles, and dihedrals, are modeled by infinitely high square well potentials. During a simulation, an atom's velocity remains constant until a potential step is encountered, where it changes instantaneously according to the conservations of energy, momentum, and angular momentum. Simulations proceed as a series of such collisions with a rapid sorting algorithm employed at each step to determine the following collision.

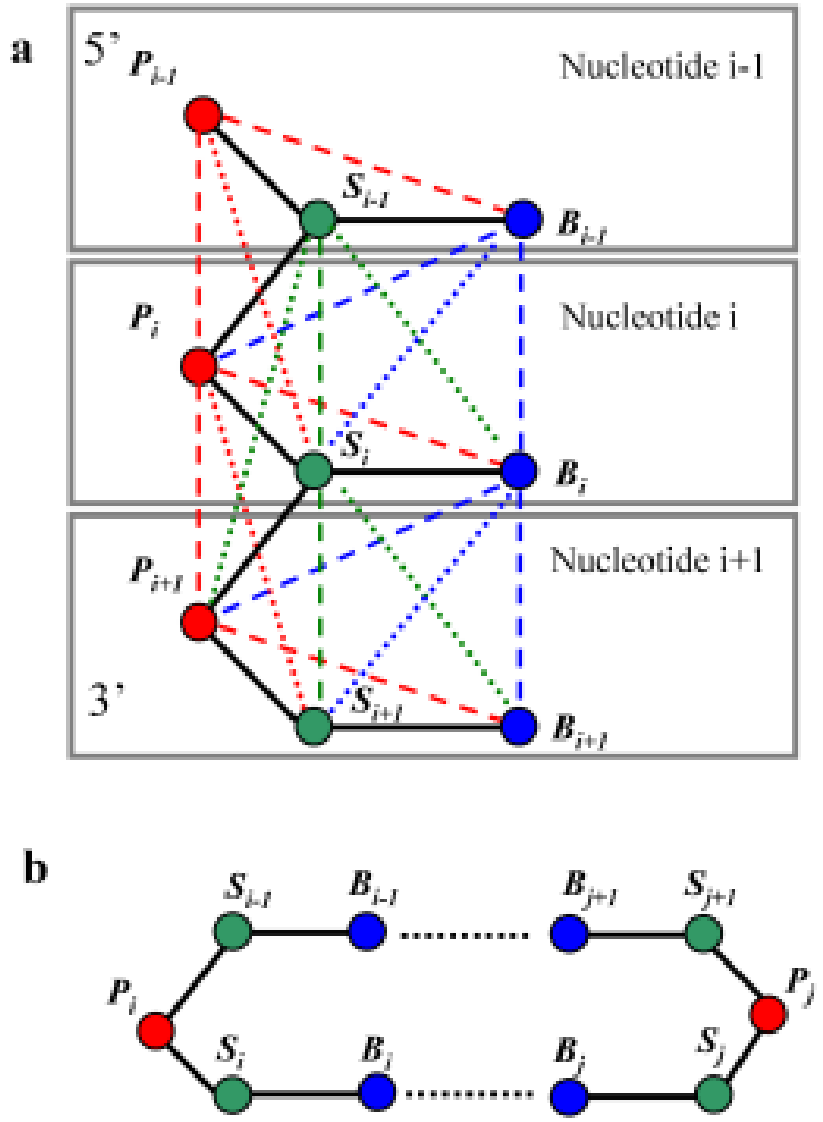

Figure 5.1: Coarse-grained structural model of RNA employed in DMD simulations. (A) Three consecutive nucleotides, indexed i-1, i, i+1, are shown. Beads in the RNA: sugar (S), phosphate (P), and base (B). (Thick lines) Covalent interactions, (dashed lines) angular constraints, (dashed-dotted lines) dihedral constraints. Additional steric constraints are used to model base stacking. (B) Hydrogen bonding in RNA base pairing. (Dashed lines) The base-pairing contacts between bases $B_{i-1} : B_{j+1}$ and $B_i : B_j$. A reaction algorithm is used (see Materials and Methods) for modeling the hydrogen bonding interaction between specific nucleotide base pairs.

5.2.2 The simplified RNA model

We approximate the single-stranded RNA molecule as a beads-on-a-string polymer, with each bead corresponding to either sugar (S), phosphate (P), or nucleo-base (B) moieties, thus making three beads for each nucleotide (Fig. 5.1). Beads P and S are positioned at the center of mass of the corresponding phosphate group and the five-atom ring sugar. For both purines (adenine and guanine) and pyrimidines (uracil and cytosine), we represent the base bead (B) as the center of the six-atom ring. The neighboring beads, which are either inter- or intranucleotides, are constrained to mimic the chain connectivity and the local chain geometry (Fig. 5.1). The types of constraints include bonds (solid lines), bond angles (dashed lines), and dihedrals (dot-dashed lines). The parameters for the bonded interactions mimic the folded RNA structure and are derived from a high-resolution RNA structure database (MARR03). The nonbonded interactions are crucial to model the folding dynamics of RNA molecules. In our model, we include the base-pairing (A-U, G-C, and U-G), base-stacking, short-range phosphate-phosphate repulsion, and hydrophobic interactions, which are described below as well as in the parameterization procedure.

5.2.3 Base pairing

In the folding of RNA molecules, the complementary hydrogen bonding interactions between nucleotides, base pairing, are the key interactions. We use the reaction algorithm to model the hydrogen-bonding interaction between specific nucleotide base pairs. The details of the algorithm can be found in (DBB$^+$03a). Briefly, to mimic the orientation-dependent hydrogen-bond interaction, we introduce auxiliary interaction beside the distance-dependent interaction between donor and acceptor (Fig. 5.1). For example, once the two nucleotides (e.g., A-U, G-C, or U-G, represented as B_i and B_j in Fig. 5.1) approach the interaction range, we evaluate the distances between S_iB_j and S_jB_i, which define the orientations between these two nucleotides. If the distances

satisfy the predetermined range, we allow the hydrogen bond to be formed, and forbid its formation otherwise.

5.2.4 Phosphate-phosphate repulsion

Phosphates are negatively charged and usually repel each other. To account for the repulsion, we assign repulsion between phosphate groups. Due to the strong screening effect of water and ions, we use the Debye − Hückel model to account for the electrostatic repulsion between phosphates. We discretize the continuous potential with a step-wise function with a step of 1 Å and the cutoff distance of 10 Å.

5.2.5 Hydrophobic interactions

Buried inside the double-helix, the bases are hydrophobic in nature. We include a general attraction between all bases. Due to the coarse-graining feature of our model, the assignment of attraction between bases results in overpacking (e.g., the symmetrically attractive tends to form close packing). In order to avoid this artifact, we introduce an effect energy term to penalize the overpacking of bases: $E_{overpack} = dE\Theta(n_c - n_{max})$. Here, $\Theta(x)$ is a step function, which adapts the value of x if x is positive and zero; otherwise, n_c is number of contacts, and n_{max} is the maximum number of contacts; dE is the repulsion coefficient. Using a cutoff of 6.5 Å, we sample the available RNA structures from NDB and find that n_{max} corresponds to 4.2.

5.2.6 Base stacking

A close examination of stacking interactions from available crystal structures suggests the following salient features: (1) Stacking interactions are usually short-ranged as in close packing; (2) each base has a stacking valence of 2; i.e., a base does not make more than two stacking interactions; (3) three consecutively stacked bases align approximately

linearly. We include the above features into our model. We compute the distance distributions of stacked bases from available RNA structures. We find that distribution depends on the types (purine or pyrimidine), and we identify the stacking cutoff distances: 4.65 Å between purines, 4.60 Å between pyrimidines, and 3.80 Å between purine and pyrimidine. To approximately model the linearity of the stacking interactions, we penalize two bases, which form stacking interactions to the same base, from coming closer than 6.5 Å. As a result, these three bases effectively form an obtuse angle. Next, we discuss the energy parameterization of the base-stacking interaction, base pair, and hydrophobic interactions.

5.2.7 Parametrization of the hydrogen-bond, base-stacking, and hydrophobic interactions

In order to determine the pairwise interaction parameters for the stacking and hydrophobic interactions for all pairs of the bases, we decompose the sequence-dependent free energy parameters for individual nearest-neighbor hydrogen-bond model (INN-HB) (MSZT99). We assume that the interaction of neighboring base pairs in INN-HB is the sum of the hydrogen-bond, base-stacking, and hydrophobic interactions. In a nearest neighboring base-pair configuration (Fig. 5.1), B_{i+1} and B_i (B_{j-1} and B_j) usually stack on top of each other. However, if both bases $Bi+1$ and B_j are purines, we find that they tend to stack instead. The bases B_i and B_{j-1} are usually farther than the cutoff distance of 6.5 Å. Given the experimentally tabulated energy between all possible neighboring base pairs (MSZT99), we are able to determine the values of E_{Stack}, E_{HB}, and $E_{Hydrophobic}$, which are consistent with the experimental measurements using singular value decomposition.

5.2.8 Loop entropy

The loop entropy plays a pivotal role in RNA folding kinetics and thermodynamics (TB99). Hence, an RNA folding prediction method should take the entropic effect into account, either implicitly (in all-atom MD simulations (SNR$^+$04)) or explicitly (Monte Carlo or dynamic programming methods (RE99; Mat06). However, due to the reduction of the degrees of freedom in our simplified RNA model, the entropy is often underestimated in our DMD simulations. For example, we often observe that the RNA molecule forms long loops readily and is kept trapped in a nonnative conformation for a long simulation time. To overcome such an artifact due to the coarse-graining process, we develop a simple approach in the DMD simulation to model the loop entropy explicitly. We use the experimentally tabulated free energies for different types of loops, including hairpin, bulge, and internal loops (MSZT99). The free energy of a loop depends on its size and type (hairpin, bulge, or internal loops). We compute the effective loop free energy in DMD simulations based on the set of base pairs formed in simulations. Upon the formation or breaking of each base pair, the total loop free energy changes. We estimate the loop free energy difference ΔG_{loop} for each base pair formation during the simulation and determine the probability to form such a base pair by coupling to a Monte Carlo procedure using a Metropolis algorithm with a probability, $p = e^{-\beta \Delta G_{loop}}$. If it is possible to form the base pair after the stochastic estimation, the particular base pair will form only if the kinetics energy is enough to overcome the possible potential difference before and after the base pair formation. Upon breaking of a base pair, the stochastic procedure is not invoked, since it is always entropically favorable to break the base pair. The breaking of the base pair is only governed by the conservation of momentum, energy, and angular momentum before and after the base pair breakage.

5.2.9 Replica-exchange DMD simulations

We use DMD (DBSS98) simulations to investigate the dynamics of RNA folding. Efficient exploration of the potential energy landscape of molecular systems is the central theme of most molecular modeling applications. Sampling efficiency at a given temperature is governed by the ruggedness and the slope toward the energy minimum in the landscape. Although passage out of local minima is accelerated at higher temperatures, the free energy landscape is altered due to larger entropic contributions. To efficiently overcome energy barriers while maintaining conformational sampling corresponding to a relevant free energy surface, we utilize the replica exchange sampling scheme. In replica exchange computing, multiple simulations or replicas of the same system are performed in parallel at different temperatures. Individual simulations are coupled through Monte Carlo-based exchanges of simulation temperatures between replicas at periodic time intervals. Temperatures are exchanged between two replicas, i and j, maintained at temperatures T_i and T_j and with energies E_i and E_j according to the canonical Metropolis criterion with the exchange probability p = 1 if $\Delta = (1/k_BT_i - 1/k_BT_j)(E_j - E_i) \leq 0$, and $p = e^{-\Delta}$, if $\Delta > 0$. We perform the replica exchange method to rapidly sample the conformational space available to RNA. For simplicity, we use the set of eight temperatures in all the replica exchange simulations: 0.200, 0.208, 0.214, 0.220, 0.225, 0.230, 0.235, and 0.240. The temperature is in the abstract units of kcal/(molk_B). Note that we approximate the pairwise potential energy between the coarse-grained beads with the experimentally determined free energy of nearest neighboring base pairs, instead of the actual enthalpy. As a result, the temperature does not directly correspond to the physical temperatures. In DMD, constant temperature simulation is achieved by the Andersen thermostat (Andersen 1980). Folding simulation of a 36-nt-long RNA sequence (median size of RNA chains in the sample) for 2×10^6 DMD time units took ≈ 5 h of wall-clock time utilizing eight 3.6-GHz Intel Xeon compute nodes, communicating over the Message Passing Interface library (http://www-unix.mcs.anl.gov/mpi).

5.2.10 Q-value of a putative RNA structure

We use the fraction of the total number of native base pairs, the Q-value, as one criterion to evaluate the accuracy of a putative RNA structure predicted from simulations. As used in protein folding studies (SSK94), the Q-value quantifies the extent of native-likeness of a putative structure with respect to the native structure. To compute the Q-value of a putative RNA structure, the native structure or at least the native secondary structure must be known. If a Q-value equals 1, the putative structure correctly predicts the native base pairs and features all native secondary structures. If the Q-value is close to zero, the corresponding structure does not resemble the native state.

5.2.11 Weighted histogram analysis method

The weighted histogram analysis method (KBS$^+$92) was used to analyze the thermodynamics of RNA folding. The MMTSB toolset (FKB04) was used to perform WHAM on replica exchange trajectories. Since our simulations are started from a fully extended conformation, we exclude the first 5×10^5 time units of the simulation trajectories and use the last 1.5×10^6 time units of simulation trajectory for performing the WHAM analysis.

5.3 Results

5.3.1 Large-scale benchmark test of DMD-based *ab initio* RNA structure prediction on 153 RNA sequences

We test the predictive power of the DMD-based RNA folding approach by selecting a set of intermediate-length RNA sequences, whose experimentally derived structures are available at the Nucleic Acid Database (NDB, http://ndbserver.rutgers.edu), and compare our predictions with experimentally derived structures and folding dynamics.

We restrict our study to RNA molecules having a length greater than 10 nucleotides (nt) and shorter than 100 nt. Short RNA molecules lack well-formed tertiary structures and were excluded from this study. Notably, this set of 153 molecules spans a range of tertiary structural motifs: cloverleaf-like structures, L-shaped tRNAs, hairpins, and pseudoknots.

For each RNA molecule, we first generate a linear conformation using the nucleotide sequence. Starting from this extended conformation, we perform replica exchange simulations at different temperatures (see Materials and Methods). The three-dimensional conformation corresponding to the lowest free energy is predicted as the putative structure of the RNA molecule, assuming that the corresponding native structure is unknown. The extent of native structure formation in simulations is measured by computing the Q-values (akin to protein folding experiments (SSK94), see Materials and Methods), defined as the fraction of native base pairs present in a given RNA conformation. We compute Q-values for the lowest free energy states (i.e., predicted putative structures) and also the maximum Q-values sampled during the course of simulations (Fig. 5.2). For a majority of the simulated RNA sequences, the lowest free-energy structures from simulations have predicted Q-values close to unity, suggesting the correct formation of native base pairs in simulations. The average Q-value for all 153 RNA molecules under study is 94%. For comparison with available secondary structure prediction methods, we also compute the Q-values using Mfold (Fig. 5.2), and the average Q-value is 91%. The DMD-based RNA folding approach shows improvement over the Mfold method in predicting the native base pairs, especially for pseudoknots.

Out of 153 RNA molecules studied, there are three cases (NDB codes: 1P5O, 1P5M, and 2AP5) where the predicted and maximum Q-values as well as the Q-value from the Mfold prediction are small. Additionally, there are a few cases where the predicted Q-values are not unity while the maximum Q-values are unity (Fig. 5.2). This suggests that our simulations are able to sample the native state, but the force field cannot

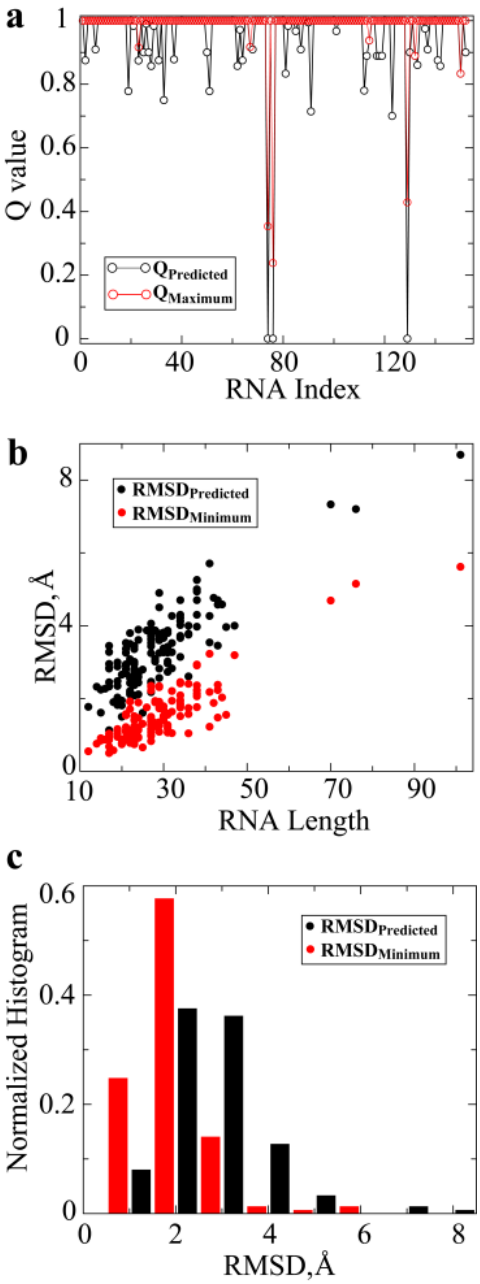

Figure 5.2: Ab initio RNA folding using DMD. (A) Fraction of native base pairs (Q-values) present in the predicted RNA 3D structure. The maximum Q-values during the course of simulations are also shown, which depict the conformational sampling efficiency of the DMD algorithm to reach the native states. We also show the Mfold predicted Q-values. (B) Scatter plots of RMSD for the final folded conformation with respect to the experimentally derived native structure as a function of RNA size. Large RNA molecules have increased fluctuations due to larger conformational freedom and consequently have greater RMSD from the native conformation. (C) Normalized histogram of predicted and least RMSD to the native RNA structure.

capture it. Therefore, further optimization of the force field parameters is necessary.

The objective of this work is *ab initio* tertiary structure prediction. Toward this goal, we evaluate our predicted tertiary structures by computing their root mean square deviation (RMSD) from corresponding native structures, excluding the three RNA molecules where the secondary structures are not correctly formed (Fig. 5.2). The RMSD value is computed based on the backbone phosphate atoms. We notice that the predicted lowest free energy structure usually does not have the lowest RMSD with respect to the corresponding crystal structure (Fig. 5.2), possibly due to inaccuracy of the force field and the coarse-grained nature of the simplified RNA model. Despite these approximations, the method features striking predictive power. We observe that for the RNA molecules with nucleotide length < 50, the predicted RMSD are < 6Å. Longer RNA molecules exhibit larger RMSD due to the highly flexible nature of RNA molecules. Among the 153 sequences simulated, 84% of the predicted tertiary structures have an RMSD of < 4Å with respect to the experimentally derived native RNA structure. Many functionally important RNA molecules have short sequences, e.g., pre-miRNA is typically 70-100 nt long, suggesting a potential for DMD-based RNA folding for de novo structure prediction of functional RNA molecules.

5.3.2 Folding dynamics in DMD simulations

We analyze the folding thermodynamics and kinetics for several nontrivial RNA motifs, the pseudoknot and tRNA. We also study the folding thermodynamics of B-RNA (Escherichia coli 23S rRNA, G1051-C1109) (LD94), 72 RNA (E. coli α-operon mRNA fragment G16-A72) and its mutants: 72-C RNA (G16-A72, G51→C) and 72-14 RNA (G16-A72, AA44→CC, UU54→GG) (GD94), and compare our simulations with corresponding experimental measurements.

5.3.3 Pseudoknot folding

The RNA pseudoknot structure has non-nested base pairing and minimally consists of base-pairing between a loop region and a downstream RNA segment. Pseudoknots serve diverse biological functions, including formation of protein recognition sites mediating replication and translational initiation, self-cleaving ribozyme catalysis, and inducing frameshifts in ribosomes (SB05). We study pseudoknot folding dynamics by selecting a 44-nt-long representative pseudoknot whose structure is available at high resolution (NDB code: 1A60) (Fig. 5.3). This pseudoknot represents the T-arm and acceptor stem of the turnip yellow mosaic virus (TYMV) and has structural similarity with TYMV genomic tRNA (KvdW$^+$98). The model pseudoknot is stabilized by the hairpin loop formed at the 5' end of RNA, and by the interactions with the loops of the pseudoknot in the 3' end.

We calculate the folding thermodynamics using the weighted histogram analysis method (WHAM) (see Materials and Methods). The specific heat (Fig. 5.3) has one peak centered at temperature $T^* = 0.245$ and a shoulder near $T^* = 0.21$ (temperature expressed in reduced units, see Materials and Methods), suggesting the presence of intermediate states in the folding pathway (Fig. 5.3). The thermodynamic folding intermediate species is characterized by computing the two-dimensional potential of mean force (2D-PMF) as a function of total number of base pairs (N) and the number of native base pairs (NN). The 2D-PMF plots at temperatures corresponding to the two peaks in the specific heat (Fig. 5.3) show two intermediate states with distinct free energy basins: The first intermediate state corresponds to the folded 5' hairpin, while the second intermediate corresponds to the formation of one of the helix stems for the 3' pseudoknot. For example, the 2D-PMF plot at $T^* = 0.21$ (Fig. 5.3) shows that the shoulder in the specific heat plot corresponds to the formation of the second intermediate state. The basins corresponding to the two intermediate states have a weak barrier, resulting in a lower height in the specific heat plot. Contact frequencies at the folding intermediates

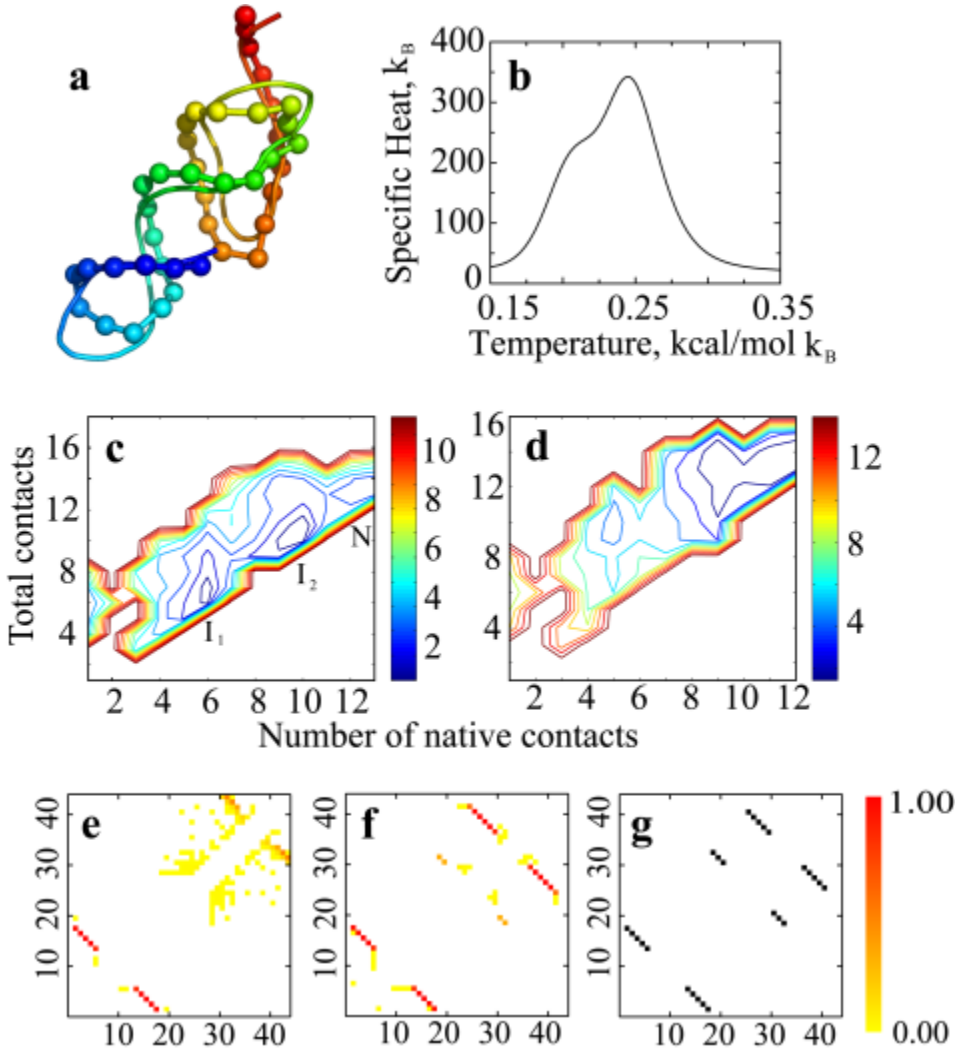

Figure 5.3: Ab initio folding kinetics and energetics of a model pseudoknot RNA. (A) Superposition of experimental pseudoknot structure (NDB code: 1A60, ribbon) against DMD prediction (ribbon backbone trace with backbone spheres). Backbone ribbons are colored blue (N terminus) to red (C terminus). (B) Graph of specific heat of the pseudoknot molecule as a function of simulation temperature. (C) Two-dimensional potential of mean force 2D-PMF for pseudoknot folding at T* = 0.245 (corresponds to the major peak in the specific heat). (I1, I2) The two intermediate states, (N) native state. (D) The 2D-PMF plot at T* = 0.21. (E) Internucleotide base-pairing contact frequencies at the first folding intermediate (I1) corresponding to the state where the hairpin is folded. (F) Internucleotide base-pairing contact frequencies at the second intermediate state (I2) corresponding to the formation of the major groove helix stem of the pseudoknots. (G) Contact map of the native state (N) as observed in the experimental structure (NDB code: 1A60).

and the native state contact map (Fig. 5.3) demonstrate the progress in the pseudoknot folding pathway.

5.3.4 tRNA folding

The transfer RNA (tRNA) molecules serve as information transducers, linking the amino acid sequence of a protein and the information in DNA, thereby, decoding the information in DNA. Crystallographic studies of tRNA molecules reveal a distinct L-shaped 3D structure (Fig. 5.4). Here, we study the folding of a yeast phenylalanine tRNA (NDB code: 1evv). For the tRNA molecule, the predicted Q-value is ≈ 0.87. We find that the RMSD of the putative structure is ≈ 7.20Å with respect to the crystal structure, while the lowest RMSD in the simulation is ≈ 5.2Å. The predicted structure misses the tertiary contacts between the TΨC-loop and D-loop (Fig. 5.4); such long-range contacts are stabilized by metal ion coordination as shown in high-resolution X-ray crystallography structures (Fig. 5.4). Since our model does not include nucleotide metal ion coordination effects, such tertiary contacts mediated by metal coordination are not expected to form during the DMD simulations. However, this methodology is still able to recapitulate all other tertiary contacts, including the long-range helix between the 5' and 3' ends and co-stacking between the terminal helix and D-helix, and between the TΨC-helix and anticodon helix (Fig. 5.4). The specific heat of tRNA exhibits a single peak at $T^* = 0.22$ (Fig. 5.4). However, a single peak in the specific heat does not guarantee the absence of folding intermediates (DCD$^+$04). We first compute the 2D-PMF as the function of the total number of contacts and the number of native contacts at $T^* = 0.22$ (Fig. 5.4). We observe two major basins: one corresponding to the unfolded/misfolded states ($NN = 0$ and $N \geq 0$), and the other corresponding to a state that has $NN \approx 6$. There are minor basins corresponding to states with NN ranging from 10 to 18 and the native state with $NN \approx 22$.

We examine the folding trajectories in simulations (Fig. 5.4) and observe that the

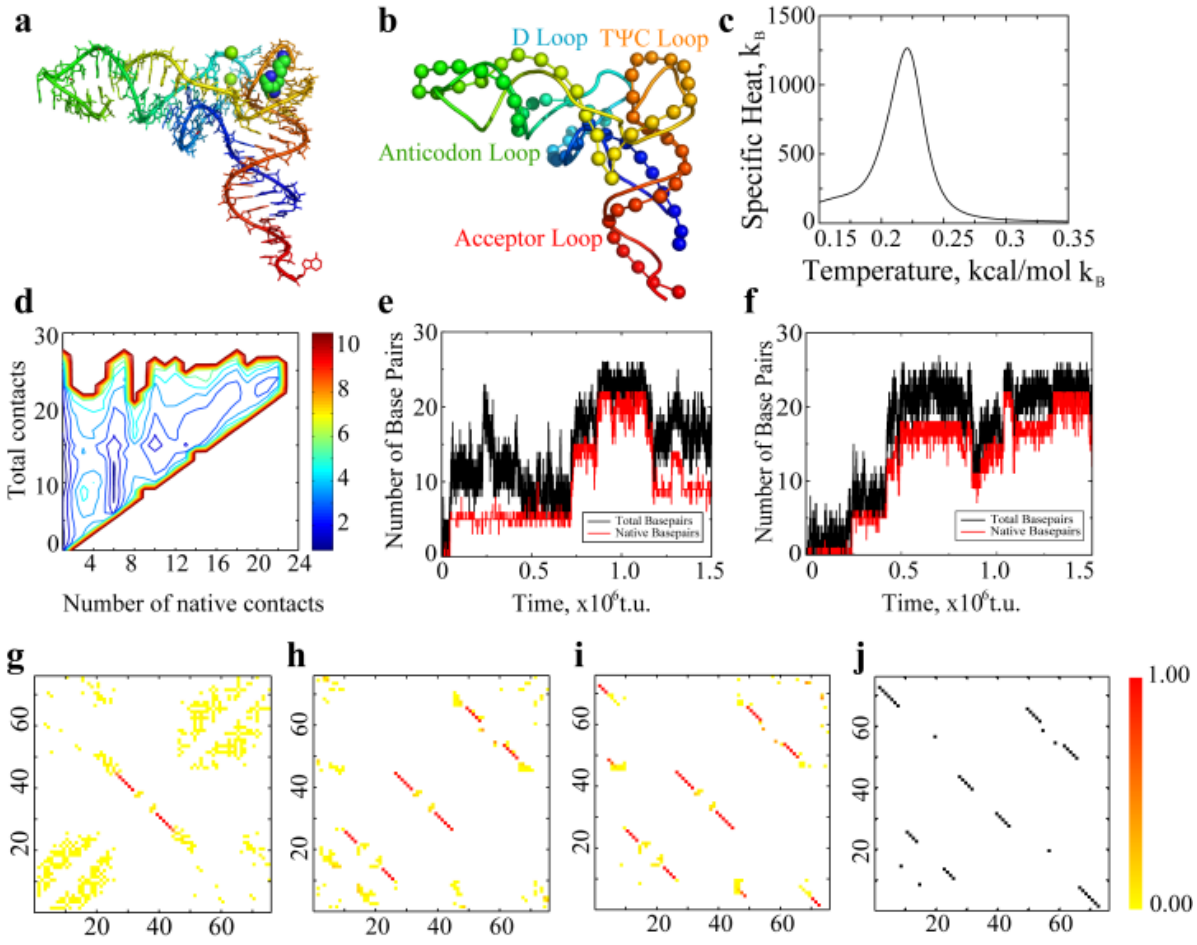

Figure 5.4: Ab initio folding kinetics and energetics of a model tRNA. (A) Mg2+ binding site (sphere) in the tRNA. (B) Superposition of experimental tRNA structure (NDB code: 1EVV, ribbon) against DMD prediction (ribbon backbone trace with backbone spheres). Backbone ribbons are colored blue (N terminus) to red (C terminus). D loop, anticodon loop, and acceptor loop are indicated with color representing their position in the tRNA secondary structure. (C) The specific heat of the tRNA molecule as the function of simulation temperature. (D) The 2D-PMF as the function of the total number of contacts and the number of native contacts at $T^* = 0.22$. (I1, I2, I3) Folding intermediates, (N) native conformation. (E,F) Folding events in the trajectories of tRNA replica exchange simulation. Two folding events in corresponding different replicas are observed out of eight replicas. (G-I) Internucleotide base-pairing contact frequencies at the threefolding intermediate states, I1, I2, and I3, respectively. (J) Contact map of the native conformation (N) as observed in the experimental structure (NDB code: 1EVV)

tRNA folding process is not cooperative and follows multiple folding pathways. The two folding trajectories (Fig. 5.4) consist of distinct folding intermediates populated along the successful folding pathway. While the rest of the folding pathways are different in the two folding events, common to the two folding events is the initial formation of the anticodon helix, suggesting that these intermediate states (NN ranging from 10 to 18; Fig. 5.4) have similar free energies. Sorin and coworkers investigated the folding mechanism of tRNA using all-atom molecular dynamics simulations with $G\bar{o}$ model ($\mathrm{SNR}^{+}04$). These investigators observed that the tRNA folds via multiple folding pathways with distinct intermediates populated upon folding. This observation is consistent with our studies. The advantage of our methodology is that we do not impose the native structure bias in the simulations.

5.3.5 Folding of ribosomal and messenger RNA fragments

We compare our predictions with experimental data by studying the thermodynamics of four RNA sequences: B-RNA (E. coli 23S rRNA, G1051-C1109) (LD94), 72 RNA (E. coli α-operon mRNA fragment G16-A72), and the 72-C RNA (G16-A72, G51→C), 72-14 RNA (G16-A72, AA44→CC, UU54→GG) mutants (GD94). The 72 RNA fragment contains a coding RNA sequence, suggesting functional implication of folding thermodynamics associated with translational regulation. Gluick and Draper (GD94) measured the melting curves of wild-type and mutant 72 RNA. Mutations at key 72-RNA nucleotides resulting in the 72-C RNA, 72-14 RNA sequences were engineered to probe significant events in 72-RNA folding thermodynamics (GD94). We compute the temperature dependence of specific heat of wild-type and mutant 72 RNA sequences from simulations (Fig. 5.5). The predicted specific heat curves show a single dominant peak for each of the three 72 RNA sequences. We observe a shoulder at the higher temperature regime of 72-14 RNA, suggesting a convolution of multiple small transitions in 72-14 RNA folding. Notably, in 72-14 RNA, the peak of specific heat, corresponding to

the experimentally measured melting temperature Tm, is shifted to the higher temperature regime, suggesting that the mutation AA44→CC, UU54→GG stabilizes the RNA. The predicted changes of Tm for 72-14 RNA and 72-C RNA with respect to wild-type 72-RNA are in agreement with the experimental measurements (GD94).

B-RNA represents the highly conserved 59-nt fragment (G1051-C1109) of E. coli 23S rRNA, serving as a recognition site for two structurally different ligands: ribosomal protein L11 and thiostrpton (a class of thiazole-containing antibiotics) (LD94). Laing and Draper (LD94) have experimentally measured the melting curves of B-RNA in 100 mM KCl, 0.1 mM MgCl2. For B-RNA, we find that the specific heat profile has a broader peak ($T^* = 0.22$) than that of 72-RNA and its mutants (Fig. 5.5). We also observe that the magnitudes of specific heat of B-RNA at different temperatures are significantly smaller, as compared with that of the 72 RNA variants. The peak of specific heat for B-RNA is shifted toward the low-temperature regime, relative to 72-RNA, suggesting that B-RNA has lower stability than 72-RNA (Fig. 5.5). These observations are consistent with calorimetric experiments of (LD94) and (GD94). Also, the predicted B-RNA structure is in agreement with experimental observations (Fig. 5.5). The corresponding experimental structure is taken from the 23S rRNA structure (NDB code: 1C2W), which is reconstructed from cryo-electron microscopy. We find that our predicted structure with the lowest free energy state agrees with the cryo-electron microscopic structure with a backbone RMSD of 6.2 Å.

There is also a broad shoulder in the B-RNA specific heat at the low-temperature regime. The flattened as well as skewed melting curve suggests a possible convolution of multiple folding transitions between intermediate states as observed in experiments (LD94). The 2D-PMF of B-RNA at $T^* = 0.22$ shows twofolding intermediates and one non-native state (Fig. 5.5). Internucleotide contact frequencies in the near-native state are in agreement with the folded RNA state (Fig. 5.5). Such accord between the predicted folding thermodynamics and experimental observations suggests that this

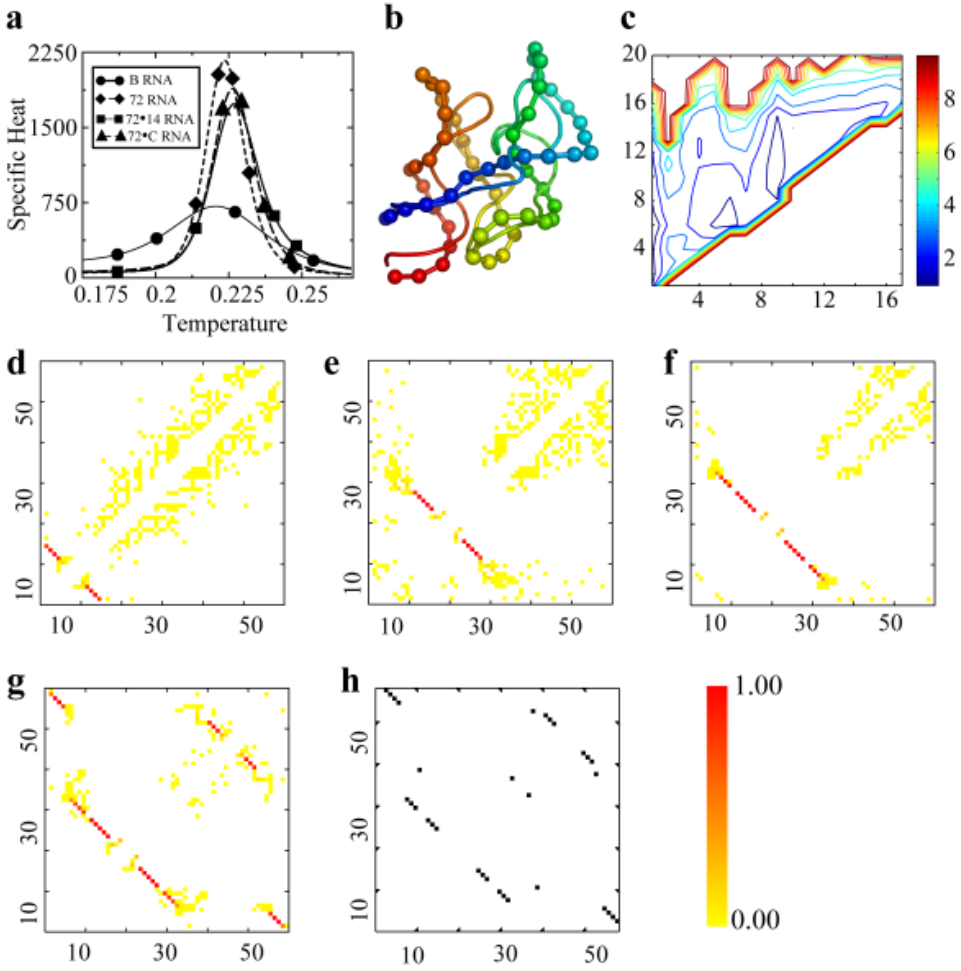

Figure 5.5: Thermodynamics of B-RNA and 72 RNA variants. (A) Specific heat: (circles) B-RNA, (diamonds) 72-RNA, (squares) 72-C RNA, (triangles) 72-14 RNA (shown in DMD units). (B) Superposition of experimental B-RNA structure (ribbon) against DMD prediction (ribbon with backbone spheres). Backbone ribbons are colored blue (N terminus) to red (C terminus). (C) 2D-PMF of B-RNA as the function of the number of total base pairs and native base pairs. We find that there are three major basins in the 2D-PMF corresponding to intermediate states I1, I2, and I3. (I4) Near-native intermediate conformation, (N) native conformation. (D) Internucleotide contact frequencies at the intermediate state with about zero native contacts (i.e., non-native state I1). (E) Internucleotide contact frequencies at the B-RNA folding intermediate state with about five native contacts, (non-native state I2). (F) Internucleotide contact frequencies at the B-RNA folding intermediate state I3 with about nine native contacts. (G) Internucleotide contact frequencies at the B-RNA folding intermediate state I4 at near-native conformation. (H) Contact map in the native state (N) observed in the experimental structure (NDB code: 1C2W).

approach is suitable for probing thermodynamics of RNA folding.

5.4 Discussion

DMD-based RNA folding is rapid and potentially applicable for a number of molecular biotechnology and molecular biology-related applications. Rapid and accurate prediction of RNA tertiary structure is the core of the RNA folding problem. For small RNA molecules, *ab initio* predictions developed in this work have yielded significantly accurate structures. The available conformational space increases exponentially with increasing length of the simulated RNA. For example, we observed large structural flexibility for longer RNAs in DMD simulations. The complexity of adequately sampling conformational space through DMD simulations also increases significantly for large RNA molecules. We suggest that the hierarchical organization of RNA secondary and tertiary structures may be exploited to predict the structure of complex RNA molecules. Additionally, experimentally derived constraints, such as base pairs from SHAPE chemistry (WMW06), proximity information from hydroxyl radical experiments, and size measurements from small-angle X-ray scattering (SAXS), can help the structure determination of large RNA molecules. We can use biased interaction potential to guide simulations and generate RNA structures consistent with experimental measurements.

Two alternative scenarios for the time-course of RNA folding are possible: (1) the sequential hierarchical folding, where the secondary structure forms first, then tertiary contacts finally shape a specific tertiary structure (TB99); and (2) the mutually dependent interplay of RNA secondary and tertiary interactions, where substantial rearrangement of folding intermediates successively takes place (SZW$^+$99). We posit that using simplified models for folding RNA is apt for investigating RNA folding mechanisms in de novo RNA fragments, as no assumptions regarding the folding mechanisms are made a priori. For the folding of the pseudoknot, the folding intermediate I1 forms

a weak non-native stem (contacts between nucleotides 32 and 42; see Fig. 5.3), while intermediate I2 does not have this but does have native stems (see Fig. 5.3). The correct folding requires the disruption of the non-native stems. Similarly, the folding trajectories of tRNA (Fig. 5.4) suggest that the folding of the RNA always accompanies the formation of non-native base pairs, with the total number of base pairs larger than the number of native base pairs. Therefore, our simulations suggest that RNA folds in a non-hierarchical manner, with nonnative conformations accumulated during the folding as observed in experiments (WMW05a).

RNA folding has been investigated experimentally using single-molecule fluorescence spectroscopy (Zhu05). These experiments conclude that RNA folding proceeds via a highly frustrated energy landscape, and adequate sampling of the RNA conformational ensemble is necessary for predicting RNA folding kinetics. The agreement of the thermodynamics between simulation predictions and experiments for B-RNA, 72-RNA, and its mutants encourages the efficacy of this method to qualitatively study folding thermodynamics of RNA molecules.

The coarse-graining process might alter the conformational entropy of molecules. To circumvent this coarse-graining artifact, the entropic contribution of loop formation is effectively modeled by estimating the loop free energies in simulations. The predicted structures correspond to the lowest free energy state, which is the result of an intricate interplay between enthalpy and entropy. We compute the effective loop free energies during simulations and introduced a stochastic approach to evaluate the formation of each base pair, corresponding to changes in loop lengths or formation/disruption of loops (see Materials and Methods). We find that this procedure is crucial for the correct prediction of the RNA structures: without taking the loop entropy into account, the simulation maximizes the number of base pairs but does not penalize the formation of additional loops, resulting in non-native RNA structures (data not shown).

One of the salient features of our approach is the rapid conformational sampling

efficiency of DMD. We have previously reported estimates of experimental time scales accessible by DMD simulations (DBD05). Typically, DMD simulations performed on a single processor can span time scales of the order of microseconds. Because of parallelization of replica exchange methodology, much larger time scales are accessible with short simulations. We perform the replica exchange method to rapidly sample the conformational space available to RNA. Folding simulation of a 36-nt-long RNA sequence for 2×10^6 DMD time units took $\approx$5 h of wall-clock time utilizing eight 3.6-GHz Intel Xeon compute nodes, communicating over MPI. Within the 2×10^6 time units of simulations, multiple folding transition events were observed. Since the DMD codes are highly optimized, we found that the computational time scales linearly with respect to the system size. Das and Baker (DB07) have reported the prediction of tertiary structures of RNA molecules with lengths of $\geq$30 nt. This approach utilizes assembly of short RNA fragments using Monte Carlo sampling with a knowledge-based energy function to predict putative RNA conformations. The DMD-based RNA folding approach is able to predict folding for longer RNA molecules having better agreement with the corresponding native structures. Generating 50,000 fragments with 45 sec per fragment would require 625 CPU-hours of computation, as opposed to 33 CPU-hours for a 30-nt-long RNA. Our method is fully automated, since a unique tertiary structure is predicted, corresponding to the least free energy conformation. In addition, replica exchange DMD simulations also offer probing the mechanistic features, (e.g., folding kinetics and thermodynamics) of the RNA folding process. Due to the computational efficiency of the DMD-based RNA folding prediction, we are able to test a larger set of RNA molecules than is accessible using the fragment-based approach. Finally, the web-based DMD simulation tool iFold (http://ifold.dokhlab.org) (SDN$^+$06) may be extended for predicting the folded structure and probing the folding dynamics of de novo RNAs.

Chapter 6

Applications of computer automation to biomolecular simulations

6.1 Introduction

One of the most challenging issues with studies of biological systems is the time and length scales that are relevant to biology. For example, the course of some chemical reactions occur at the time scales of femtoseconds (1×10^{-15} s), while protein aggregation occurs at the time scales of hours and even years (1×10^{4} s). Hence, the range of biologically-relevant time scales spans 20 orders of magnitude. Similarly, the range of length scales that are of interest to biology spans over six orders of magnitude. None of the experimental, theoretical, and computational approaches can alone probe these time and length scales as a whole. Dynamic and structural features of large biomolecules are often invisible to current experimental techniques due to their inherent resolution limitations in length and time scales. Computational approaches offer a unique opportunity to uncover the atomic structure and biological properties of experimentally challenging molecules and molecular complexes.

Direct computational approaches employing all-atom molecular dynamics (MD) simulations provide detailed information on the local dynamics of molecules. However, owing to the complexity of protein conformational space, all-atom MD simulations have

severe limitations on the time and length scales that can be studied. An alternative approach is the simplification of protein models. In the simplified protein models, amino acids are coarse-grained to the level of effective beads (DBSS98) The interaction potential between these beads can be derived from protein structure, in vivo experiments or biophysical analyses. A more realistic simulation approach for simplified protein models is discrete molecular dynamics (DMD). This approach permits the rapid and accurate sampling of the conformational space of biomolecules and their complexes (DD05). One of the remarkable illustrations of the speed and accuracy of DMD approach is its ability to recapitulate the experimental studies of unfolded protein states and unravel their properties (DJD05). Success of DMD approach in studies of proteins dynamics makes it a very valuable tool for the community of computational molecular biologists. The goal of this work is to bring DMD to the multidisciplinary community of bioinformatics researchers through the web (http://iFold.dokhlab.org). The goal of iFold is not protein structure prediction, but rather utilizing the native structure for deciphering protein dynamics. The simplicity of the iFold user-interface allows this server to be used also by experimentalists for probing possible molecular states.

6.2 iFold - a platform for interactive folding simulations of proteins

We built a novel web-based platform for performing discrete molecular dynamics simulations of proteins. In silico protein folding involves searching for minimal frustration in the vast conformational landscape. Conventional approaches for simulating protein folding insufficiently address the problem of simulations in relevant time and length scales necessary for a mechanistic understanding of underlying biomolecular phenomena. Discrete molecular dynamics (DMD) offers an opportunity to bridge the size and timescale gaps and uncover the structural and biological properties of experimentally undetectable

protein dynamics. The iFold server supports large-scale simulations of protein folding, thermal denaturation, thermodynamic scan, simulated annealing and p_{fold} analysis using DMD and coarse-grained protein model with structure-based $G\bar{o}$-interactions between amino acids.

6.2.1 iFold automation methodology

Hardware resources for large-scale DMD simulation jobs run by the iFold server are based on a 300-node Beowulf Linux cluster provided by the University of North Carolina. The underlying tool in iFold is of DMD (DBSS98). The front end of iFold server consists of two subparts: (1) Client-side: The presentation layer that the user interacts with, using a web-browser; and (2) Server-side: The business logic part that processes the information the user inputs and aggregates information that needs to be sent back to the user (e.g. queue contents). The client-side is constructed in HTML and JavaScript and the server-side is built using PHP (http://www.php.net). The glue between the server-side and the client-side is the Smarty (http://smarty.php.net/) templating engine. Smarty allows creating HTML templates with cavities for PHP variables. At runtime, these cavities are filled in by PHP scripts, allowing an easy segregation between the presentation layer and the business logic. The server-side process interacts with the iFold scheduler - a Java application that verifies the user-specified inputs and submits simulations to iFold compute nodes using TCP-IP connections based on Java Sockets API.

Once a DMD simulation task is submitted to the iFold scheduler, it appends the simulation job to a pending jobs queue in which simulations are executed on a first-come-first-serve basis. As soon as an iFold compute node is available, the simulation's input parameters and desired outputs are dispatched to the compute node, over the Internet, using a Java socket connection. Upon successful completion of a DMD simulation, the compute node parses the list of desired outputs (simulation trajectory; p_{fold} value; graphs of energy versus temperature, energy versus time, or gyration radius versus time)

and executes standard scripts for performing these analyses on DMD simulation results. The user is notified about the simulation summary via email and he/she may login to iFold web server to download the desired simulation results.

6.2.2 Design of the iFold server

The design of iFold consists of a PHP front end and a Java back end (which connects to the protein-folding scripts) that communicate through sockets connections. Key functions provided by the system are user registration, job submission and execution, and administrative support. The front end handles all of the administrative and job preparation functions, while the back end is the workhorse that performs and manages the actual simulations. This division of function is important in order to support multiple back end processors and the commensurate workload. The front end is a sophisticated PHP application using advanced technologies such as Smarty templates and AJAX techniques. The backend is a long-running application-independent daemon that receives requests from the front end scheduler and tracks the current state of execution. With these technologies, the entire application is easily adapted to support additional applications.

6.2.3 User registration

User registration is a three-step process:

1. The user fills out the registration form on the main iFold server page, explaining the need for desired computational privileges - regular user/advanced user/administrator - and submits the registration form.

2. An email is automatically dispatched to the user's specified email address with a uniquely generated link that the user clicks on to verify the email address.

3. After validation of the email address, iFold administrators are notified (also via email) to approve the user registration request. The registration and administrative approval are important safeguards to assure that the user is given an appropriate level of access. Once the iFold administrators approve the user and assign appropriate compute privileges, another email is sent to the user notifying him that he may login to the server and submit folding simulation jobs.

6.2.4 Job submission and execution

Once the user submits the protein folding request, the simulation task is added to the queue of pending simulations, which is implemented on a MySQL database. Administrators are given full management access to this queue and can reprioritize or delete jobs. On the web server, a java-based task scheduler polls the database for queued simulation tasks and checks for daemon availability.

Once a task reaches the top of the queue, the scheduler will process it as soon as there is an available back end processor. The scheduler initiates the execution of the task by sending all information associated with the task (what outputs it needs and what parameters it has) across a socket connection to a daemon application running on a compute node - a machine that contains the protein folding scripts. The daemon application receives the message regarding the task and constructs a shell command to run the appropriate script to handle the requested task. The script runs and generates the requested outputs, which could be graphs of radius of gyration vs. simulation time, energy of simulated protein vs. simulation time, or energy of simulation vs. temperature or movies of simulation trajectories.

These outputs are placed in a temporary directory on the compute node. When the daemon thread finds that there are new outputs, it sends them across a socket connection back to the scheduler on the web server and then deletes them from the compute node. The simulation scheduler receives the outputs and updates the database

to reflect that the simulation task has successfully completed, stores the simulation outputs in a directory for the user on the web server machine, and sends an email to the user notifying him that the task is complete - along with any messages about its status. The user logs back in to the iFold web site, which displays the updated information from the database, and the user can download the outputs from the iFold server.

6.2.5 Administrative tasks

Administrators manage users, the backend processes and create/edit help pages through a rich text editor (http://www.kevinroth.com/rte/). The Manage Users section allows an administrator to quickly edit/delete/approve a user and the Manage iFold section allows administrators to pause and restart the daemon processes and check their status.

6.3 Prototypical iFold simulation results

The key functions supported by the iFold front end are as follows:

1. *Task Submission*: The iFold task submission process is driven by an XML file that holds all classes of simulation tasks available on the iFold server and the corresponding simulation parameters in a hierarchical manner. When the user loads the task submission page, the server-side PHP scripts parse the XML file, validating the parameters and filling the smarty templates for the client-side processing.

2. *Registration Process*: To ensure security of the iFold server, human intervention is necessary for completion of the registration process. When the user registers on the main page, an email is automatically dispatched to his/her specified email address with a unique URL that contains a mathematically generated, encrypted key. When the user clicks the link, he/she is presented with a success page and another email is sent to an iFold administrator for approving the user's request for iFold.

3. *Queue Management*: The activity page allows users to view their activity and the outputs generated by iFold. Users can see and delete only their own tasks; they can see if there are other tasks in the queue, but no information about the tasks. Administrators, on the other hand, can see all information on the queue and are able to delete tasks if necessary.

The iFold server has following three modes of operation:

- The *guided user mode*: This mode is designed to provide a convenient user interface to biologists unfamiliar with simulation techniques. In this mode, simulations are performed using apposite default values for simulation parameters. Thus, by choosing a simulation task and specifying the structure of the protein, the users may run DMD simulations to collect relevant data such as melting temperature of their protein.

- The *advanced user mode*: This mode is designed for researchers familiar with various simulation techniques and gives the user freedom to specify valid ranges of simulation parameters and to download the simulation outputs in all formats supported by iFold.

- The *administrative user mode*: In this mode, the user is provided special privileges to have administrative access over other users of iFold. The regular operation of iFold server is fully automated; however, web-based support of multiple administrators for iFold is useful for providing expedited access for new users and monitoring new hardware requirements.

The iFold server is hosted at http://iFold.dokhlab.org and supports the following simulations. The input parameters for each of these simulations, their default values and the corresponding observables are described in the Methods section.

1. *Protein folding simulation*: The user enters the initial and final temperatures (T_{init}, T_{final}) at which the folding simulation of the protein is performed; starting from

a linear conformation of the protein at temperature Tinit, the temperature of the system is reduced at a constant, user-specified rate, until the system reaches the final temperature T_{final}.

2. *Thermal denaturation*: Starting with the native protein structure at low temperature, the system's temperature is raised at a constant rate until the protein starts to unfold.

3. *Thermodynamic scan*: Multiple constant temperature DMD simulations of the protein are performed over a range of temperatures to ascribe thermodynamic properties (such as heat, capacity and melting temperature) of the protein.

4. *Simulated annealing*: The annealing process is iterated a number of times for effectively sampling the conformation space-rapidly raising temperature from the last stable conformation and relaxing protein at a slow rate. The lowest energy conformation approximates the folded protein.

5. *Folding probability analysis*: p_{fold} measures the probability of a decoy to fold. It is a quantitative measure of progress in the folding pathway of the given conformation.

6.4 Simulation tasks supported by iFold

6.4.1 Protein folding simulation

Folding simulations are performed by starting from a linear conformation of the biomolecule at a high initial temperature and reducing the temperature at the specified rate until the system reaches the final temperature. Folding simulations are among the most useful methods for rapidly characterizing the biophysical properties of proteins. The user uploads the protein's structure file in Protein Databank (PDB) format or specifies the four letter PDB code of the corresponding protein and then selects the following simulation

parameters:

- Initial temperature - desired temperature from where the simulation is initiated.
- Final temperature - final temperature at which the simulation is stopped.
- Heat exchange coefficient - rate of heat transfer in the system.
- Potential set of folding simulation outputs - trajectory of simulation, graphs and/or data for radius of gyration vs. simulation time, energy vs. simulation time, or energy vs. simulation temperature.

The folding simulation parameters include initial temperature T_i, final temperature T_f, heat exchange coefficient C, and total simulation time t_{max}. By default $T_i = 1.2$; $T_f = 0.4$, C = 0.0001, and $t_{max} = 100000$. The advanced user can assign arbitrary values to these parameters.

6.4.2 Protein unfolding simulation

Unfolding simulations correspond to simulating the thermal denaturation of the chosen protein. In unfolding simulations, we start with the native structure of the protein, simulated at low temperature, and slowly raise the temperature of the system until the protein starts to unfold. Unfolding simulations are essentially an inverse operation of folding simulation, except that instead of starting from the linear conformation, we start from the native conformation. The user uploads the protein structure file in Protein Databank (PDB) format or specifies the four letter PDB code of the corresponding protein and then selects the following simulation parameters:

- Initial temperature - desired temperature from where the simulation is initiated.
- Final temperature - final temperature at which the simulation is stopped.
- Heat exchange coefficient - rate of heat transfer in the system.

- Potential set of unfolding simulation outputs - trajectory of simulation, graphs and/or data for radius of gyration vs. simulation time, energy vs. simulation time, or energy vs. simulation temperature.

The unfolding simulation parameters include initial temperature T_i, final temperature T_f, heat exchange coefficient C, and total simulation time t_{max}. By default $T_i = 1.2$; $T_f = 0.4$, C = 0.0001, and $t_{max} = 100000$. The advanced user can assign arbitrary values to these parameters.

6.4.3 Protein thermodynamic scan

In thermodynamic scan, we perform constant temperature DMD simulations of the given biomolecule over a range of temperatures to get the thermodynamic properties of the simulated protein. The user uploads the biomolecule's structure file in Protein Databank (PDB) format or specifies the four letter PDB code of the corresponding protein/DNA/RNA and specifies the following set of parameters required for thermodynamic scan simulations:

- Range of temperature used for thermodynamic scans. Since the transition temperature is in the range of T = 0.7-0.9, a default value of temperature T = [0.6, 0.7, 0.8, 0.9, 1.0, 1.1] (temperature in of ϵ/k_B units of DMD scale) is kept for guided user mode.

- The desired set of thermodynamic scan outputs - graphs and/or data for variation of heat capacity of the biomolecule vs. simulation temperature, average frequencies of inter-residue or inter-nucleotide contacts made over the given temperature range, energy vs. simulation time, radius of gyration vs. simulation time, or energy vs. simulation temperature.

For each DMD simulation, we set the temperature as constant ($T_i = T_f$), default heat exchange coefficient C = 0.0001, and total simulation time t_max= 50000 t. u. The

advanced user can assign arbitrary values for these parameters.

6.4.4 Simulated annealing

In this mode we start from a high temperature conformation where the protein is unfolded and slowly reduce the temperature to a very low value, whereby the protein folds to a stable conformation. This process is repeated a certain number of times, suddenly raising the temperature from the last stable conformation and allowing the protein to cool at a very slow rate. Multiple runs ensure that the protein reaches the most stable conformation - the native state. The user uploads the protein's structure file in Protein Databank (PDB) format or specifies the four letter PDB code of the corresponding protein and specifies the following set of parameters required for simulated annealing simulations:

- Number of runs for simulated annealing - default value of 5 runs is kept for guided user mode.
- The desired subset of unfolding simulation outputs - graphs and/or data for variation of energy of the biomolecule vs. simulation temperature, average frequencies of inter-residue or inter-nucleotide contacts made over the given temperature range, energy vs. simulation time, radius of gyration vs. simulation time, or energy vs. simulation temperature.

6.4.5 Folding probability analysis

Folding probability (p_{fold}) refers to the probability likelihood that a given decoy conformation of the protein will fold before unfolding. It is a quantitative measure of progress in protein folding of the given conformation. We have developed a method for performing p_{fold} scan using DMD simulations. A default value of 10 iterations at transition temperature $T = 0.8$ is kept for guided user mode. The user uploads the protein's

structure file in Protein Databank (PDB) format or specifies the four letter PDB code of the corresponding protein and specifies the following set of parameters required for p_{fold} scan:

- The structure of the decoy conformation of the given protein.
- The desired set of p_{fold} analysis outputs - the folding probability value (p_{fold}) for the given decoy conformation.

6.5 iFoldRNA - three-dimensional RNA structure prediction and folding

Three-dimensional RNA structure prediction and folding is of significant interest in the biological research community. Here, we present iFoldRNA, a novel web-based methodology for RNA structure prediction with near atomic resolution accuracy and analysis of RNA folding thermodynamics. iFoldRNA rapidly explores RNA conformations using discrete molecular dynamics simulations of input RNA sequences. Starting from simplified linear-chain conformations, RNA molecules ($\leq$50 nt) fold to native-like structures within half an hour of simulation, facilitating rapid RNA structure prediction. All-atom reconstruction of energetically stable conformations generates iFoldRNA predicted RNA structures. The predicted RNA structures are within 2-5 Å root mean squre deviations (RMSDs) from corresponding experimentally derived structures. RNA folding parameters including specific heat, contact maps, simulation trajectories, gyration radii, RMSDs from native state, fraction of native-like contacts are accessible from iFoldRNA. We expect iFoldRNA will serve as a useful resource for RNA structure prediction and folding thermodynamic analyses.

The central dogma of molecular biology presented RNA as the fundamental ingredient in genetic translational machinery. However, recent discoveries of RNAi, ribozymes

and aptamers have extended the scope of RNA function beyond the central dogma. We now understand that RNA molecules serve diverse structural, catalytic and regulatory function in eukaryotic cells. The tertiary structure of RNA molecules plays a crucial role in determining RNA function. However, accurate prediction of three-dimensional (3D) structure and folding kinetics of RNA presents a significant challenge in molecular biotechnology. The necessity of large quantities of pure RNA samples and technical limitations hinder the applications of X-ray crystallography, and nuclear magnetic resonance for high-throughput structure elucidation. These challenges have lead to a newfound interest in computational prediction of RNA tertiary structure (SYKB07) and investigating thermodynamics and mechanism of RNA folding.

A majority of computational methods for probing RNA structure and folding dynamics are limited to secondary structure elucidation, while stochastic models have been used to study RNA folding kinetics. Although these computational approaches have demonstrated a significant utility in predicting the RNA secondary structure, they are largely inadequate for predicting 3D RNA structures. Recently, fragment assembly Monte Carlo (DB07), nucleotide cyclic motifs (PM08) and discrete molecular dynamics (DMD) simulations (DSC$^+$08) have been proposed for 3D RNA structure prediction. The iFoldRNA (http://iFoldRNA.dokhlab.org), is a web-resource for rapid and accurate predictions of 3D RNA structures and probing folding thermodynamics. iFoldRNA performs folding simulations using the DMD engine (DSC$^+$08; DBSS98) and Medusa force field (DD06) to simulate RNA folding dynamics.

6.5.1 iFoldRNA automation methodology

A simplified three-bead per nucleotide model of RNA and replica-exchange DMD simulation protocol with eight replicas is used to sample the RNA conformational space (DSC$^+$08). An estimate of the free energies of RNA loop regions is explicitly included in the force-field to model the entropic contributions from RNA loop formation. The

simulation is followed by a reconstruction protocol to generate atomic resolution structures. 3D structures corresponding to the lowest free energy states in DMD scale are ascribed as the near-native conformations.

A 520-processor Topsail Linux cluster from the University of North Carolina is used for performing replica-exchange DMD simulations of RNA folding (DSC$^+$08). The back-end of iFoldRNA distributes simulation tasks from the iFoldRNA website to compute nodes of the Topsail cluster using a queue scheduler and a Java-based network communication (SDN$^+$06). Once a DMD simulation completes, the compute node generates the putative native-like structures having least relative free energy in DMD scale and user-specified simulation thermodynamic outputs. These outputs are dispatched back to the scheduler and subsequently the user is notified of simulation results via email.

6.5.2 Prototypical iFoldRNA simulation results

Multiple native-like RNA topologies and the corresponding relative free energy values are accessible from the iFoldRNA server. Our recent work has demonstrated the efficacy of the DMD conformational sampling engine in rapid simulations of RNA folding dynamics (DSC$^+$08). The iFoldRNA resource enables world-wide access to rapid tertiary structure prediction and folding thermodynamics of RNA molecules using the DMD engine. Folding parameters including inter-nucleotide contact maps, simulation trajectories, gyration radii, root mean square deviations (RMSDs) from native state, and fraction of native-like contacts (Q-value) are accessible from the iFoldRNA server. Secondary structures generated by iFoldRNA are consistent with Mfold and ViennaRNA predictions.

Low RMSDs (2-3 Å) are observed in 3D superpositions of iFoldRNA predictions against experimental structures, demonstrating the accuracy of iFoldRNA in structure prediction (Fig. 6.1). Typical iFoldRNA folding simulations and analyses are performed within an hour as compared to months to years spent on conventional molecular dynam-

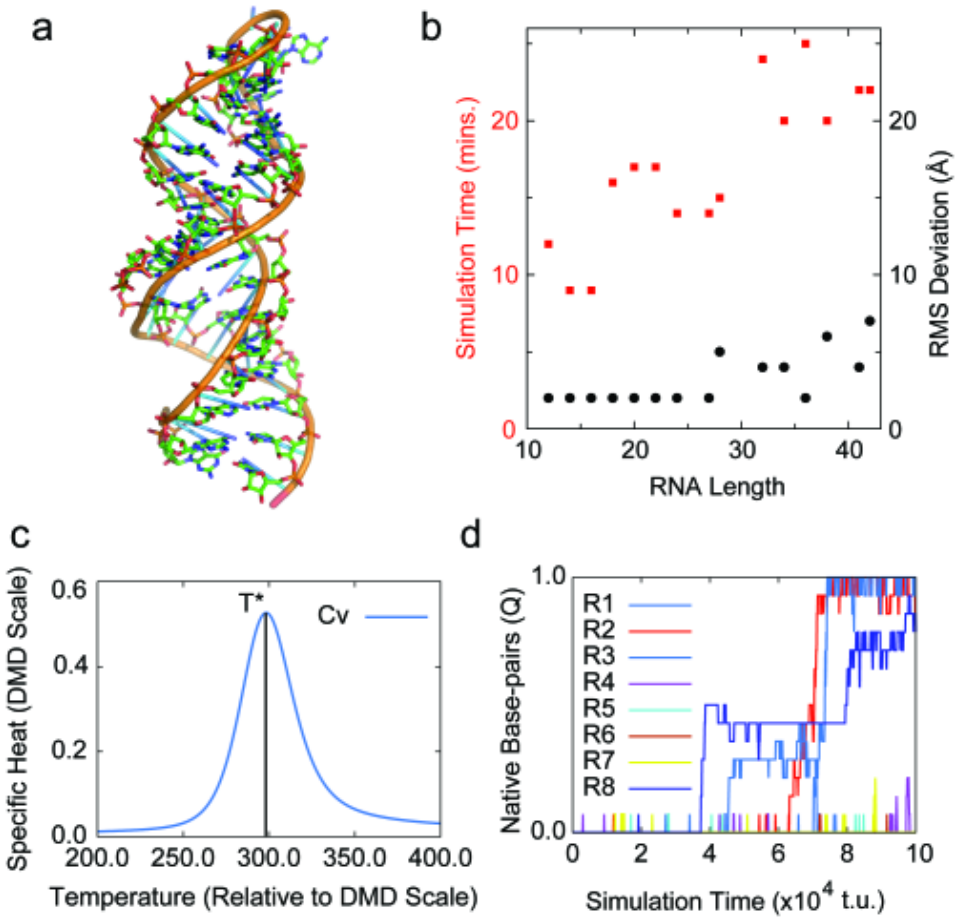

Figure 6.1: iFoldRNA tertiary structure prediction and folding thermodynamics.

ics simulations to adequately explore the conformational space. Fast conformational sampling ability of DMD enables rapid structure prediction of putative RNA sequences using iFoldRNA. We have also developed a post-simulation analysis tool, iFoldRNA-Analysis available at the iFoldRNA website for user-specified analyses of RNA folding using the weighted histogram analysis method (http://www.mmtsb.org). Sample simulation outputs obtained from iFoldRNA and iFoldRNA-Analysis are elucidated in Fig. 6.1. Folding transition temperatures obtained from specific heat graph (Fig. 6.1) and fractions of native base pairs (Fig. 6.1) can be directly compared across different RNA sequence.

6.6 Discussion

The architecture of iFold server is modular in nature and utilizes a Linux cluster as its compute resources. Adding more compute nodes to the system is feasible, thereby allowing the iFold server to scale up to an order of a million simulation tasks submitted. iFold's convenient user-interface is expected to make simulations accessible to molecular biologists for probing possible protein conformations, especially when conventional experimental techniques become unfeasible. Protein conformations featuring p_{fold} values close to 0.5 constitute the transition state ensemble. Combining results of DMD simulations from iFold with traditional molecular dynamics and quantum mechanics simulations will entail studies of proteins over vast time and length scales. Thus, the iFold server will enable effective sampling of biomolecular conformations, studying protein thermodynamics, kinetics and experimentally aided modeling.

Large RNA molecules having $\geq$50 nt (e.g. ribosomal RNA, NDB: 2il9, 142 nt) require significantly longer time scales to sample the exponentially increasing conformational space. This limits the accuracy of the iFoldRNA structure prediction to intermediate-length RNA molecules ($\leq$50 nt). In future, experimental constraints, e.g. using SHAPE (WMW05b) may be integrated with iFoldRNA to overcome such size limitations. We anticipate that the iFoldRNA server will gather significant attention in the research community interested in predicting 3D structures and probing folding mechanisms of RNA molecules. The iFoldRNA server is freely accessible at http://iFoldRNA.dokhlab.org for academic and non-profit users.

Chapter 7

Molecular dynamics simulations of cisplatinated and oxaliplatinated DNA

7.1 Introduction

Cisplatin (CP; cis-diamminedichloroplatinum(II)) and carboplatin (CBDCA,cis-diammine-1,1-cyclobutanedicarboxylatoplatinum(II)) are widely used for treatment of testicular cancer, ovarian cancer, head and neck tumors and a variety of other solid tumors. However, many tumors are intrinsically resistant or develop acquired resistance to these chemotherapeutic agents, and tumors that are resistant to one of these two platinum compounds are usually cross-resistant to the other. The mutagenicity of CP *in vivo* (Gre92) is also of concern because secondary malignancies have been associated with CP chemotherapy (TCS$^+$97). Considerable effort has been made to develop third generation platinum anticancer agents that would not share these limitations. Oxaliplatin (OX; trans-R,R-1,2-diaminocyclohexaneoxalatoplatinum(II)) is one such compound, and has recently been approved for the treatment of colorectal cancer and tumors that are resistant to CP and CBDCA. While OX does have some mutagenicity, (SCD$^+$05) it appears to be less mutagenic than CP (BKB$^+$04). CP and CBDCA form Pt-DNA adducts that contain the cis-diammine carrier ligands, while OX forms Pt-DNA adducts that contain the trans-RR-1,2-diaminocyclohexane carrier ligand. For simplicity, we will re-

fer to these as CP-DNA and OX-DNA adducts throughout this paper. Other than the differences in carrier ligand, the adducts formed by CP, CBDCA and OX appear to be identical in terms of the type of adduct formed (60-65% intrastrand GG, 25-30% intrastrand AG, 5-10% intrastrand GNG and 1-3% interstrand) and the site of adduct formation.(PHSC90; JEK89; WCN$^+$98)

Because of their abundance, the intrastrand GG adducts are thought to be major determinants of the cytotoxic response to platinum anticancer agents. The basis for the differences in tumor range and mutagenicity of OX compared to CP and CBDCA is not known, but is thought to be determined by the ability of proteins involved in damage recognition, damage repair and/or damage tolerance to discriminate between CP and OX adducts. For example, hMSH2 and MutS bind with greater affinity to CP-GG adducts than to OX-GG adducts (FNA$^+$96; ZMF$^+$02) and, as might be expected from this difference in binding affinity, defects in mismatch repair result in resistance to CP and CBDCA, but not to OX (FNA$^+$96; AKHG$^+$96; FZN$^+$97; VVU$^+$98; BHG$^+$97). Similarly, a number of damage recognition proteins and transcription factors, especially those with high mobility group (HMG) domains, have been shown to discriminate between CP- and OX-GG adducts (WCSL01; ZBJE98). The mechanism(s) by which the binding of these proteins to Pt-DNA adducts influences the cytotoxic response is not known, but has been postulated to involve shielding of the adducts from DNA repair and tolerance mechanisms (HZR$^+$94; ML96; HLVL97; VLP$^+$99), activation of signaling pathways leading to cell cycle arrest or apoptosis, and/or hijacking of transcription factors needed for DNA replication or cell division (ZBJE98; TZJE94). The binding specificity has been determined for only a few of these proteins, but where it has been studied these proteins bind to CP-GG adducts with higher affinity than to OX-GG adducts (WCSL01; ZBJE98; CFV$^+$98). Finally, translesion DNA polymerases such as hpol β and hpol η have been shown to bypass OX-GG adducts with higher efficiency than CP-GG adducts (VLP$^+$99; VMHC00; VC00), which might contribute to the differences

in CP and OX mutagenicity.

The CP- and OX-GG adducts form in the major groove and bend the DNA in the direction of the major groove. The proteins that discriminate between CP- and OX-GG adducts either bind to bent DNA or bend the DNA in the direction of the major groove after binding (LPE$^+$00; OBHY00; ORH$^+$99; SPW$^+$97). Because these proteins primarily interact with the minor groove, we have hypothesized that the ability of the proteins to discriminate between CP- and OX-GG adducts probably results from subtle differences in conformation or conformational dynamics in the DNA containing the two adducts rather than from physical interaction of the proteins with the carrier ligands of the adducts in the major groove. A number of structures have been reported for CP-GG and OX-GG adducts. The overall conformation of DNA containing these adducts appears to be similar, but exact comparisons have been difficult to make because the structures have been determined by different techniques (crystallography versus NMR), in different sequence contexts and with oligonucleotides of different length. The NMR structures obtained to date have varied with respect to the number and resolution of NMR constraints obtained and the molecular mechanics simulations used to convert the NMR constraints to final structures (GL98; MSK$^+$01; WPH$^+$04). X-ray crystallographic structures have been reported for the CP-GG and OX-GG adducts in the same sequence context, (TFL96; SWL01) but these structures may have been constrained by crystal packing restraints. We have recently obtained high resolution NMR solution structures of the OX-GG,30 CP-GG adducts33 and undamaged DNA33 in the same mutagenic AGGC sequence context (the underlined bases indicate the position of the Pt-GG adduct). The NMR studies have identified several conformational differences between the solution structures of the two Pt-GG adducts. Here we have used molecular dynamics (MD) simulations to extend this analysis to differences in conformational dynamics between CP-GG adducts, OX-GG adducts and undamaged DNA in the AGGC sequence context. Our data are consistent with earlier reports of greater distortion on the 5' side of Pt-GG

adducts (MSK$^+$01; WPH$^+$04; BYLE93; BAT91; BDF87; BVH$^+$03; MKIH00; WJPP01). Our data are also consistent with previous reports that the cis-diammine (CP) ligand preferentially forms hydrogen bonds on the 5' side of the Pt-GG adduct, while the trans-RR-diaminocyclohexane (OX) ligand preferentially forms hydrogen bonds on the 3' side of the adduct. Finally, our data show that these differences in hydrogen bond formation are strongly correlated with differences in conformational dynamics, specifically the fraction of time spent in different DNA conformations, for CP- and OX-DNA adducts. These differences were particularly evident for propeller twist, buckle, slide and shift in the vicinity of the Pt-GG adduct. We postulate that these differences in conformational dynamics could allow differential recognition of CP- and OX-GG adducts by critical DNA-binding proteins that influence the cytotoxic response to these adducts. For example, we postulate that CP-GG adducts spend a greater percentage of time in conformations favorable for binding of mismatch repair and HMG-domain DNA-binding proteins, while OX adducts spend a greater percentage of time in conformations(s) favoable for bypass by hpol β and hpol η. Experiments are currently underway to test these hypotheses. Experiments are also underway to compare the effect of sequence context on the conformational differences between these two adducts.

7.2 Methods

7.2.1 Starting structures

All five starting structures contained a 12-mer DNA duplex in the same sequence context of 5'-d(CCTCAGGCCTCC)-3' for the strand containing the platinum adduct. For the DNA portion of the structures, the NMR solution structures of the OX-DNA adduct (WPH$^+$04) (Protein Data Bank accession code 1PGC) and the CP-DNA adduct (WBK$^+$07) (Protein Data Bank accession code 2NPW) solved in the Chaney laboratory at the University of North Carolina at Chapel Hill were used as two of the starting

structures. For the CP- and OX-DNA adducts, the DNA portion of the NMR solution structure of DNA complexed with hSRY (Protein Data Bank accession code 1J46) was also used as a starting structure for the DNA backbone. The original DNA sequence from 1J46 was mutated to the same DNA sequence used for our NMR solution structures. Since the DNA backbone of the hSRY-DNA complex was more bent than the DNA in the NMR solution structures of the Pt-DNA adducts by themselves, it provided a test of whether our MD simulations were capable of driving significantly distinct starting structures to non-distinguishable structures when simulations reached equilibrium. For the platinum adducts on the DNA backbone derived from the hSRY-DNA complex, the oxaliplatin adduct was obtained from our NMR solution structure of OX-GG ($WBK^{+}07$) (1PGC) and the cisplatin adduct was obtained from the NMR solution structure of the CP-GG adduct reported by Marzilli et al. ($MSK^{+}01$) (1KSB). The Pt-GG adducts were superimposed on the DNA from the hSRY-DNA complex, with removal of the two G bases of the adduct producing the starting structures of the CP- and OX-DNA adducts. For undamaged DNA, ideal B-DNA was used as the starting structure. The molecular modeling package InsightII was used to perform all manual structural preparation and building, including the undamaged DNA and the DNA sequence change for the CP- and OX-DNA adducts.

7.2.2 Force field parametrization

The atomic partial charges for CP-GG and OX-GG adducts are not defined in the standard AMBER force field library. However, they are required and crucial in MD simulations. In order to determine the atomic partial charges of CP-GG and OX-GG adducts, the 9-methyl-guanine derivatives cis- [Pt(NH3)2(9-Me-Guo)2]2+ (CP-meG2) and [Pt(trans-RR-1,2-diaminocyclohexane)(9-Me-Guo)2]2+ (OX-meG2) were used to simplify the calculation. These derivatives were manually built from our NMR solution structures using Insight II. The atomic partial charges were determined using the

Mulliken method implemented within Gaussian03 based on either the structure geometry optimized by Gaussian03 or the NMR structure modified to the 9-methyl-guanine derivative. The density functional method B3LYP implemented within Gaussian03 was utilized; the LanL2DZ basis set was used for the platinum atom and 6-31Gd basis set was used for the rest of the atoms. The Mulliken method was found to be insensitive to the geometry of the structure and the resulting atomic partial charges were consistent in both geometry optimized and non-optimized structures.

The partial charge of the platinum atom was close to the previously published values, while the partial charges of four nitrogen atoms surrounding the platinum atom were significantly different from the published values. The partial charges of the rest of the atoms were within the theoretical range and comparable with their counterparts in the standard AMBER force field. The Mulliken charges based on the structure of the original 9-methyl-guanine derivative modified from our NMR solution structure were used without further geometry optimization as the new atomic partial charges for Pt-GG adducts (Table S7) and incorporated into our force field parameters. The atomic partial charges of chemically equivalent atoms (such as the pseudo-equatorial hydrogen atoms of the ammine group and the corresponding atoms in the two guanine bases) were averaged and the small charge discrepancy due to the structural difference between 9-methyl-guanine and deoxyguanine was distributed to the sugar according to the standard charge transfer technique used by Yao et al. (YPM94). Besides the atomic partial charges, other force field parameters of the Pt-GG adducts were referenced from AMBER parm99 force field parameters or from previous work by Yao et al. (YPM94) and Scheeff et al. (SBH99).

7.2.3 MD simulations and trajectory analysis

Altogether 25 10 ns unrestrained and fully solvated MD simulations were carried out by using SANDER module of AMBER v8.0. There were five replicas for each of the

five starting structures described above. The five replicas differed only in the MD initial velocity assigned when the system was heated for the last time. The atomic coordinates of the structures were saved every 1 ps. Both the 5' and 3'-terminal base-pairs were excluded from analysis because in several simulations the terminal bases were not base-paired and in some extreme cases they were stacked with each other. In contrast with previous MD simulations of Pt-DNA adducts, we did not use artificial constraints to hold the terminal base-pairs together. Therefore only the central ten base-pairs were considered in trajectory analysis. The PTRAJ module of AMBER v8.0 was employed for trajectory analysis.

7.2.4 Principal component analysis

The ptraj program from the AMBER v8 simulation software was used to perform PCA. The final 6 ns of equilibrated MD simulations of each of the 20 trajectories of CP-DNA and OX-DNA were subject to PCA analysis. The first three principal components described $> 90\%$ of the essential modes of dynamics for the platinum-adducted DNA complexes. The dynamics of the first three principal components were visualized using the PyMOL molecular graphics tool.

7.2.5 Hydrogen bond occupancy

A distance of less than 3.5 Å and an angle of greater than 120° between the potential hydrogen bond donor and acceptor were used as the criteria for a hydrogen bond formation. The occupancy of one hydrogen bond was defined as the percentage of frames in which the hydrogen bond existed. The hydrogen bond occupancy of one base-pair was defined as the average occupancy of hydrogen bonds existing within this base-pair. For example, the hydrogen bond occupancy of normal G·C base-pair was the average occupancy of the three standard Watson-Crick hydrogen bonds. However, there were hydrogen bonds formed between non-complementary bases in damaged DNA adducts because of

the distortion and misalignment of bases. When calculating the hydrogen bond occupancy of base-pairs, the occupancy of such non-complementary hydrogen bonds of at least 5% was necessary for consideration and the occupancy was split evenly between two base-pairs.

7.2.6 Inter-proton distance constraint comparison

A set of inter-proton distance constraints have been derived from NMR data and used to compute the NMR solution structures for the OX-DNA and CP-DNA ($WPH^{+}04$) and ($WBK^{+}07$) by CNS (crystallography and NMR system) program. In the CNS calculations, a distance violation is defined as an inter-proton distance deviating from the range of corresponding distance constraints by more than 0.5 Å. The same criterion was applied here when we tried to evaluate how well our MD simulation structures agreed with the NMR data. In addition, we classified these constraints into several categories based on the locations of two protons. The two protons may be from the same nucleoside, from two nucleosides within the same strand (intrastrand) or from two nucleosides on different strands (interstrand). For each proton, it may be a base proton or a sugar pucker proton. The H1' proton was treated as a base proton because its position is largely independent of sugar pucker. All the other protons of the sugar were considered as sugar pucker protons because their positions were strongly influenced by the sugar pucker conformation. Only the distance constrains of the central four base pairs 5'-d(A5G6G7C8)-3' were considered in this study.

7.2.7 DNA helical parameter analysis of trajectories

Five independent simulations with randomized initial velocities were performed for each of the following starting conformations: CP-SRY (the CP-GG adduct, 1KSB, superimposed on the DNA structure of the hSRY-DNA complex, 1J46), OX-SRY (the OX-GG adduct, 1PGC, superimposed on the DNA structure of the hSRY-DNA complex, 1J46),

CP-NMR (the NMR structure of the CP-DNA 2NPW), OX-NMR (the NMR structure of the OX-DNA adduct, 1PGC) and B-DNA (constructed with Insight II). The ptraj tool from AMBER-8short parallel was used to extract the equilibrated conformations between 4 ns and 10 ns of simulation time, recording snapshots at every 1 ps time-interval of the five independent runs of AMBER simulation trajectories. Using ptraj these trajectory snapshots were saved in the Protein Data Bank (PDB) format, resulting in a total of $(10,000 - 4000) \times 5$ snapshots, i.e. 30,000 snapshots for each of CP-SRY, OX-SRY, CP-NMR, OX-NMR and B-DNA simulations.

Each nucleotide type was converted from AMBER format to PDB format, and the resulting snapshots were subjected to CURVES analysis. The following CURVES parameters were extracted: global inter base-pair parameters: shift, slide, rise, tilt, roll and twist; global base-base parameters: shear, stretch, stagger, buckle, propeller and opening. Histograms were then constructed for percent occupancy versus discrete units of each DNA helical parameter. Initially, for both CP-DNA and OX-DNA adducts histograms were constructed for the CURVES parameters obtained with each starting structure (NMR and SRY) individually to determine whether the starting structure affected the distribution of CURVES parameters. Except for slight differences observed for shift at the A5-C6 base-pair step, the CURVES parameters were not influenced by starting structure (data not shown). For CP-DNA and OX-DNA simulations, the CURVES trajectories for both the SRY and NMR starting structures were combined to get better statistics of fluctuations in the CURVES parameters (60,000 snapshots each). Percentage occupancy distributions of the DNA helical parameters were calculated by normalizing the frequency distributions to 100

The Kolmogorov-Smirnov test (HPAT88) was performed for each helical parameter to calculate the P value for differences between CP-DNA versus OX-DNA adducts (Table S5) and between B-DNA and both CP-DNA and OX-DNA adducts (Table S4). The Kologorov-Smirnov test determines how significantly two distributions differ from

each other, without making any assumption regarding the distribution of data (non-parametric and distribution-free). For ease of comparison P values were reported on a negative logarithmic scale, -log(P) such that larger values imply greater statistical difference (Supplementary Data, Tables S4 and S5). In order to determine which of these -log(P) values were significant, the Kolomogorov-Smirnov test was also performed for the same helical parameters comparing five individual simulations of the same structure (i.e. five for CP-DNA, five for OX-DNA and five for B-DNA). The highest -log(P) value obtained from comparisons of the same structure was considered to set the threshold of significance. Only -log(P) values for the comparisons between CP-DNA and OX-DNA or between B-DNA and both CP-DNA and OX-DNA that were greater than the -log(P) value for comparisons of the same structures were considered significant.

7.2.8 Correlation of patterns of hydrogen bond formation with DNA helical parameters

A trajectory-wide binary profile of all combinations of the presence and absence of the following hydrogen bonds was generated using the hydrogen bond occupancy procedure described above: CP-DNA: 5' side: 5' Pt-amine hydrogen to 5N7; 3' side: 3' Pt-amine hydrogen to 7O6; and OX-DNA: 5' side: 5' Pt-amine equatorial hydrogen to 5N7; 5' Pt-amine axial hydrogen to 5N7; 3' side: 3' Pt-amine equatorial hydrogen to 7O6. DNA helical parameters of the corresponding trajectory snapshots were recorded using the helical parameter analysis procedure described above. Histograms were then constructed for percent occupancy versus discrete units of each DNA helical parameter for each of these hydrogen-bond combinations. The Kolmogorov-Smirnov test (HPAT88) was performed for each helical parameter to calculate P values for the differences between 5' hydrogen bond only and 3' hydrogen bond only for both the CP-DNA (Table S6) and the OX-DNA adducts (Table S7). The significance of these P values was determined as described above.

7.3 Results

7.3.1 The MD simulations were independent of starting structure

In total, twenty-five 10 ns unrestrained fully solvated molecular dynamics simulations were performed starting from five different DNA conformations. In order to critically evaluate whether the starting structures influenced the MD simulations, two very different starting structures were used for both the CP-DNA and OX-DNA simulations. One of the starting structures for the CP- and OX-DNA simulations used the NMR structure of DNA complexed with hSRY (Protein Data Bank accession code 1J46, bend angle = 54 °) for the DNA backbone. The other starting structures were based on the NMR structures of the CP-DNA (Protein Data Bank accession code 2NMW; bend angle = 22 °) and OX-DNA (Protein Data Bank accession code 1PGC; bend angle = 31 °) adducts alone. Idealized B-DNA was used as the starting structure for the B-DNA simulation. Five simulations each were performed with the CP-SRY, CP-NMR, OX-SRY, OX-NMR and B-DNA starting structures as described in Methods. The all-atom mass-weighted root-mean-square deviations (RMSDs) referenced to the corresponding NMR structures for the CP-DNA and OX-DNA structures and the starting B-DNA structure for undamaged DNA calculated over all 25 trajectories were used to determine how long it took the simulations to reach equilibrium, and, in turn, the range of trajectories to be used for the subsequent analyses. Plots of RMSD over time are shown for CP-DNA (Fig. 7.1), OX-DNA (Fig. 7.1) and B-DNA. The average RMSD plots of all ten CP-DNA simulations were within close proximity of one another within the final 6 ns. Thus, the simulations of the CP-DNA adducts converged to similar equilibrium structures even when the starting structures differed significantly in the initial degree of DNA bending and/or MD initial velocity. Similar behavior was evident for the OX-DNA adducts. Based on these plots of average RMSD over time, we concluded that the sim-

ulations reached equilibrium within 4 ns. Therefore our comparisons were made based on the final 6 ns of the trajectories. The average RMSD compared to the NMR solution structures (WPH$^+$04) and (WBK$^+$07) was 2.08($\pm$0.43) Å over all ten simulations for OX-DNA, 2.73($\pm$0.53) Å over all ten simulations for CP-DNA and 2.80($\pm$0.30) Å over all five simulations for B-DNA.

7.3.2 Comparison of the MD simulations with previously reported structures

For the purposes of comparison with previous structures, centroid structures (the structure from the simulation with the lowest RMSD compared to the average structure) were determined for both the CP-DNA and OX-DNA simulations. Fig. 7.2 shows an overlay of the CP-DNA and OX-DNA centroid structures and sausage diagrams of each structure individually showing the overall variation in the structures during the simulations. The centroid structures of the CP-DNA and OX-DNA adducts were very similar (the all-atom RMSD for the DNA portion of the structures was 1.3 Å). The sausage diagrams showed that the greatest variations in both the CP-DNA and OX-DNA structures occurred at the ends of the DNA molecules, but there was relatively little fluctuation of the DNA backbone during the simulations.

The centroid structures of the CP-DNA and OX-DNA simulations were also compared with the previous crystal (TFL96; SWL01) and NMR (WPH$^+$04; WBK$^+$07) structures of the same adducts (Fig. 7.3). In both cases, the simulations agreed slightly better with the NMR structures than the crystal structures. For example, the all atom RMSD comparisons of the CP-DNA centroid structure with the corresponding crystal (TFL96) and NMR (WBK$^+$07) structures were 4.2 Å and 3.1 Å, respectively. Similarly, the all-atom RMSD comparisons of the OX-DNA centroid structure with the corresponding crystal (SWL01) and NMR (WPH$^+$04) structures were 4.0 Å and 3.1 Å, respectively. For the central four base-pairs, the RMSD values were 3.5 Å, 2.8 Å, 3.6 Å and 2.9 Å

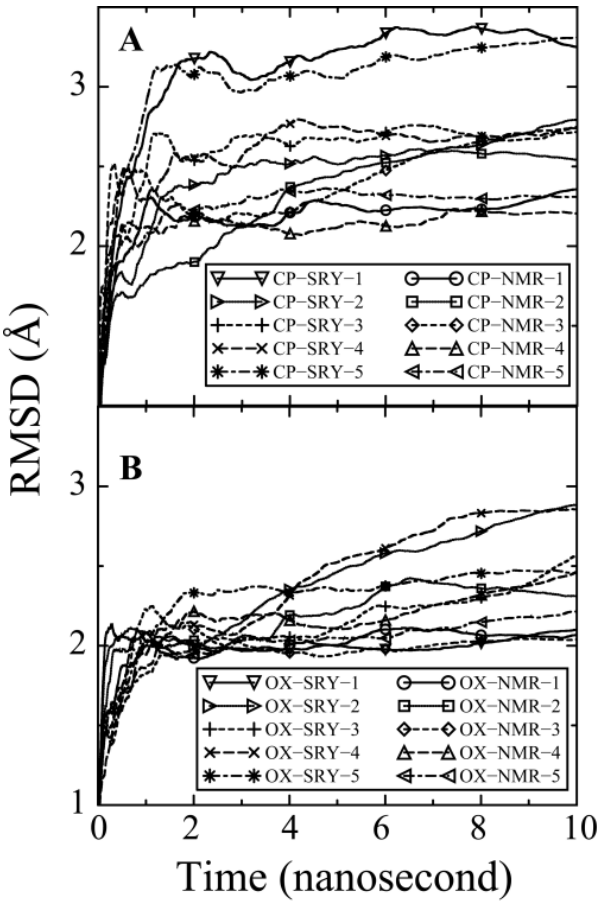

Figure 7.1: Average RMSD values for the MD simulations over time. The RMSD values for each of the 10 simulations compared to the corresponding NMR solution structure for the CP-DNA (a) and OX-DNA adducts (b) are shown for the full 10 ns of each simulation. RMSD at time t represents the average of RMSD from time zero to time t. CP-NMR and OX-NMR represent simulations starting from NMR solution structures of the Pt-DNA 12-mer duplexes. CP-SRY and OX-SRY stand for simulations starting from the more distorted DNASRY structures. The five simulation trajectories starting from CP-NMR and OX-NMR structures with different initial MD velocities are represented as continuous lines (circle symbol), dotted lines (square symbol), broken lines (diamond symbol), long broken lines (upward triangle symbol) and dot-dashed lines (left triangle symbol). The five simulation trajectories starting from the more distorted CP-SRY and OX-SRY structures with different initial MD velocities are represented as continuous lines (downward triangle symbol), dotted lines (right triangle symbol), broken lines (plus symbol), long broken lines (cross symbol) and dot-dashed lines (asterisk symbol).

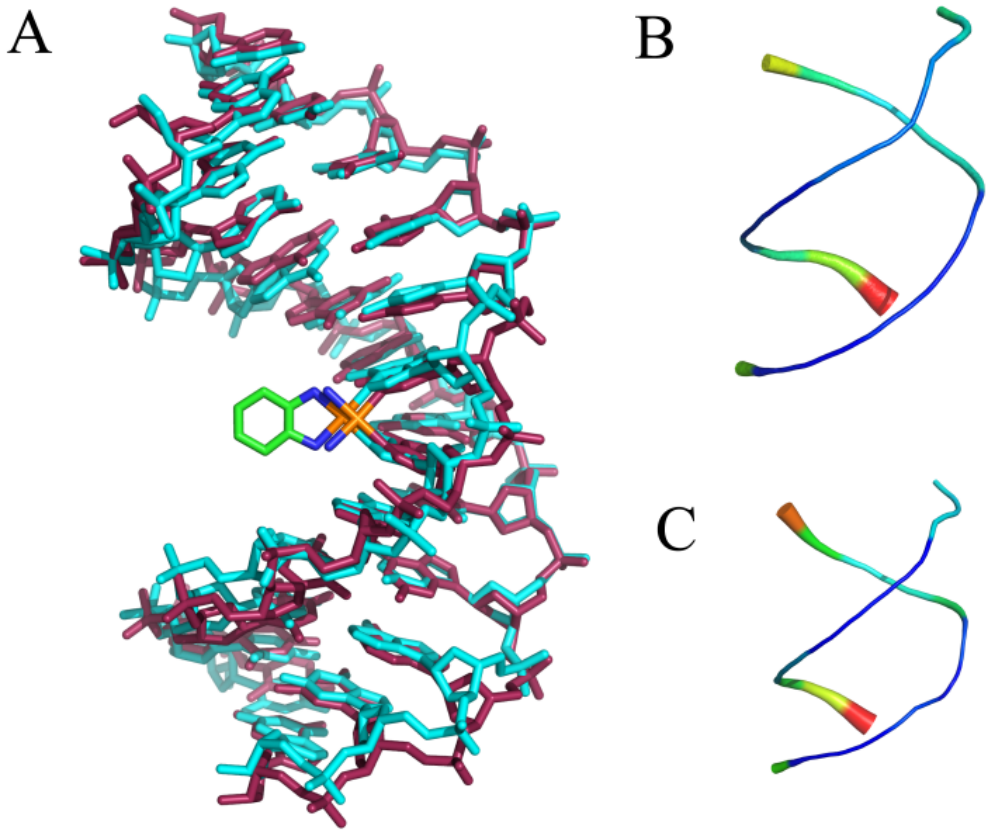

Figure 7.2: Average CP-DNA and OX-DNA structures obtained using the AMBER ptraj tool. In each case, the overall average structure was computed over the equilibrated final 6 ns trajectories combined for the five independent simulations starting with the CP-SRY, CP-NMR and OX-SRY, OX-NMR, respectively. Pair-wise Kabsch alignment was performed for the SRY and NMR starting structures. Because the RMSD values indicated that the starting structure had little influence on the average structure (RMSD = 0.83 for CP-SRY versus CP-NMR and 1.03 for OX-SRY versus OX-NMR), the two average structures were further averaged to obtain a single average structure for the last 6 ns of all ten simulations of the CP-DNA and OX-DNA adducts. Kabsch alignment was then performed against CP-DNA and OX-DNA average structures against every 1 ps snapshot of simulation trajectory. The CP-DNA and OX-DNA MD simulation snapshots having lowest RMSD with corresponding average structures were selected as the centroid structures. (a) Overlay of CP-DNA and OX-DNA centroid structures. Cyan, OX-DNA; purple, CP-DNA. (b) and (c) Sausage diagrams of CP-DNA and OX-DNA centroid structures, respectively. Root-mean-square atomic fluctuations of the CP-DNA and OX-DNA phosphate backbone were obtained by averaging the backbone fluctuations over the 30,000 snapshots of equilibrated final 6 ns of all ten simulation trajectories. The atomicfluct routine in the AMBER ptraj tool was used to compute these fluctuations. The backbone thickness and color (in a scale varying from blue to red) corresponds to the relative fluctuations of CP-DNA (b) and OX-DNA (c) backbone phosphate groups.

for the CP-crystal versus CP-centroid, CP-NMR versus CP-centroid, OX-crystal versus OX-centroid and OX-NMR versus OX-centroid, respectively.

In order to further assess how well the MD simulation structures agreed with our previously reported NMR structures (WPH$^+$04; WBK$^+$07), the extent to which the MD simulations reproduced NMR-derived inter-proton distances was also evaluated. There were 171 and 160 inter-proton distance constraints derived from NMR data within the central four base-pairs 5'-d(A5G6G7C8)-3' of CP-DNA and OX-DNA adducts, respectively (WPH$^+$04; WBK$^+$07). The statistics of the inter-proton distance violations (defined as an inter-proton distance deviating from the range of corresponding distance constraints by more than 0.5 Å; see Methods) are listed in Table 1 and the details of these distance violations are listed in Supplementary Data, Table S1. Overall, there were two (¡2%) and five (¡3%) violations for OX-DNA and CP-DNA, respectively. Most of the violations in CP-DNA simulation structures and one of the two violations in OX-DNA simulation structures involved A5 and G6 nucleotides, which were on the 5' side of the adduct and may represent the intrinsic dynamics of this portion of the molecule. Furthermore, one of the OX-DNA violations and three of the CP-DNA violations involved sugar protons, and may represent differences in sugar pucker, rather than differences in the position of the purine or pyrimidine bases. The very small number of violations compared to the total number of distance constraints indicated that our MD simulations largely reproduced the NMR data.

The geometry of the Pt-DNA adducts was also compared with previously reported structures for the CP- and OX-DNA adducts. The platinum displacement from the plane of the central guanine bases and the platinum out of guanine plane bending angles were much closer to the previously reported values for the CP- and OX-DNA adducts than was the recent MD simulation of a CP-DNA adduct reported by Elizondro-Riojas and Kozelka (ERK01), possibly reflecting the refinement of partial charges that we derived and utilized for these simulations. Our MD simulations do not reflect the differences

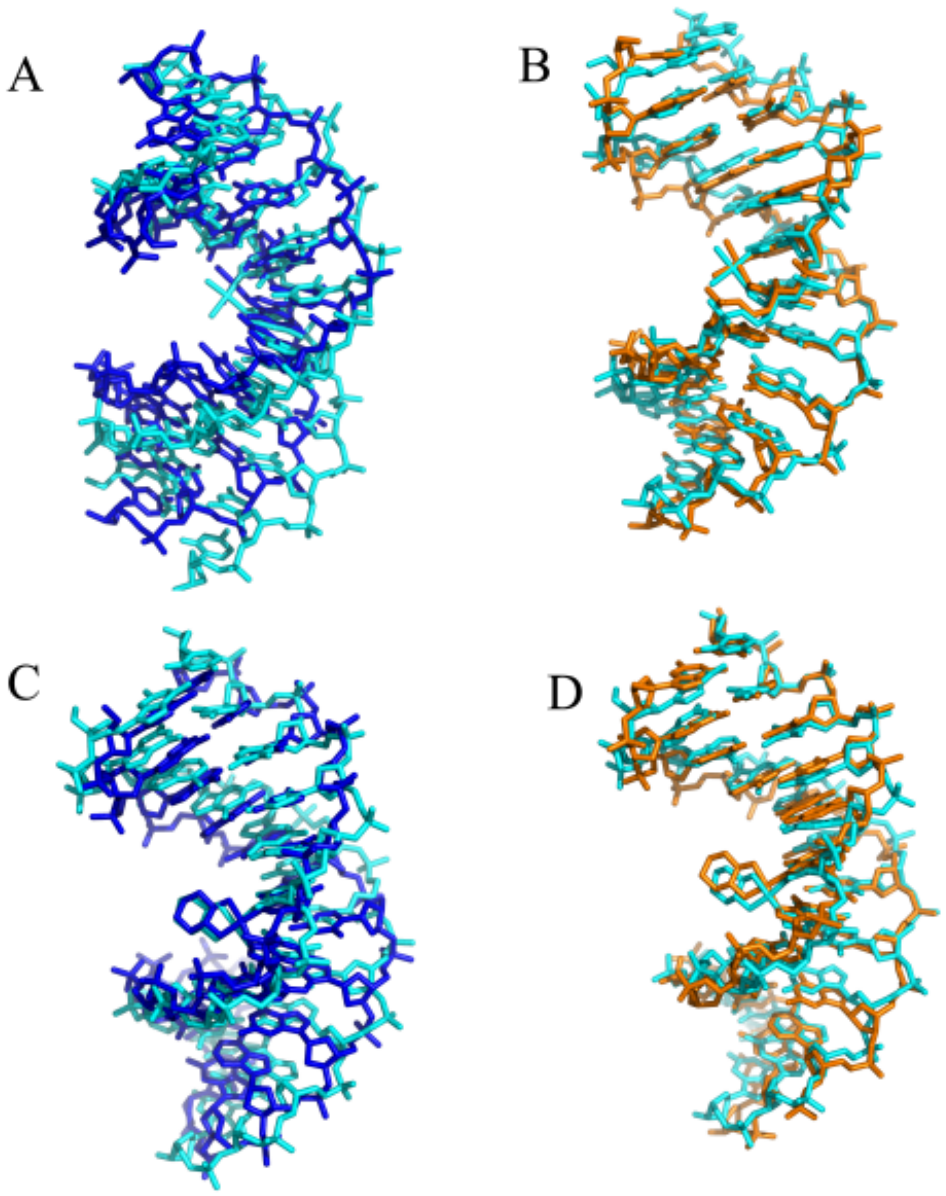

Figure 7.3: Comparison of CP-DNA and OX-DNA centroid structures with previously determined crystal and NMR structures. Pair-wise Kabsch alignment of the centroid structure (cyan) obtained from the CP-DNA and OX-DNA simulations versus the corresponding crystal structures (red) and NMR structures (orange). The NMR solution structures of CP-DNA and OX-DNA (Protein Data Bank entries 2NPW and 1PG9, respectively) in the d(CCTCAGGCCTCC)-3' sequence context were used to compare against corresponding centroid structures. The crystal structures corresponding to CP-DNA31 and OX-DNA32 were derived from the available structures in the 5'-d(CCTCTGGTCTCC)-3' sequence context (Protein Data Bank entries 1AIO and 1IHH, respectively). Representative crystal structures in the 5'-d(CCTCAGGCCTCC)-3' sequence context were derived by mutating the nucleotide sequence using Tripos Sybyl. Pair-wise Kabsch alignment of the CP-DNA and OX-DNA centroid structures was performed using PyMOL's pair-fit tool.

between CP- and OX-DNA adducts with respect to platinum displacement from the plane of the central guanine bases and the platinum out of guanine plane bending angles that were seen in our previous NMR solution structures of the adducts (WPH$^+$04; WBK$^+$07). However, it is important to note that the position of the platinum is not directly determined by NMR constraints in the solution structures and that the CNS software with the partial charges reported earlier by Yao et al. (YPM94) and Scheef et al. (SBH99) was used to derive the lowest energy solution structures. Thus, this discrepancy between our MD simulations and the solution structures in the geometry of the platinum adduct may simply reflect differences in structure refinement, such as partial charges or differences between the AMBER and the CNS softwares used for calculations of the NMR structures. AMBER and CNS use different force fields and the AMBER simulation was performed under unrestrained and fully solvated condition, whereas CNS calculation used the simulated annealing protocol in vacuum and was restrained by the inter-proton distance constraints.

7.3.3 Principal component analysis of major conformational dynamics

Principal component analysis (PCA) can be used to segregate large-scale correlated motions from random thermal fluctuations, thereby probing the essential dynamics of the system. PCA is an orthogonal linear transformation that maps the data to a new coordinate system so that the system can be deconvoluted along these coordinates. The order of coordinates is rank-ordered according to their contribution to the motion of the system. Principle component analysis was used to analyze the trajectories from each set of simulations for major conformational motions. The first three principal components described more than 90% of the essential modes of dynamics for the platinum-adducted DNA complexes. As expected for DNA simulations, the first three components of conformational motions roughly corresponded to a superposition of bending, twisting and

winding motions, in that order. By superimposing the major motions for CP- and OX-DNA adducts, it is evident that the overall conformational flexibility of DNA containing CP-GG and OX-GG adducts is very similar.

7.3.4 Hydrogen bonds

To assess the stability of the DNA duplex, the occupancy of all possible hydrogen bonds (calculated as the percentage of time during the simulation that the hydrogen bonds existed), was measured for CP-DNA and OX-DNA adducts and B-DNA of the same sequence. The data for Watson-Crick hydrogen bond occupancy are shown in Fig. 7.4. When compared to B-DNA, both CP-DNA and OX-DNA adducts show a significant decrease in standard Watson-Crick hydrogen bond occupancy for the A5·T20 and G6·C19 base-pairs on the 5' side of the adduct, whereas the base pairs on the 3' side of the adduct are almost completely intact.

Two hydrogen bonds with significantly high occupancy between the platinum carrier ligand and the DNA were identified and are shown in Fig. 7.5. One is between the 3' amine hydrogen of the platinum and the oxygen atom O6 of the 3' G7 and the other is between the 5' amine hydrogen of the platinum and the nitrogen atom N7 of A5. For the CP-DNA adduct the amine hydrogen atoms are equivalent so only four combinations of hydrogen bond formation are possible. The occupancy of each of these hydrogen bond combinations is summarized in Table 2A. Differences in global DNA conformation associated with 5' only, 5' and 3', and 3' only hydrogen bonds were minimal. However, significant conformational differences were observed in the central four base-pair region. The CP-DNA adduct spends 40.2% of its time in conformations that allow formation of the 5' hydrogen bond only and 13.3% of its time in conformations that allow formation of the 3' hydrogen bond only. However, the CP-DNA adduct also spends a significant amount of time (34.0%) in conformations that allow formation of both the 5' and 3' hydrogen bonds. The total occupancy of the 5' and 3' hydrogen bonds for the CP-DNA

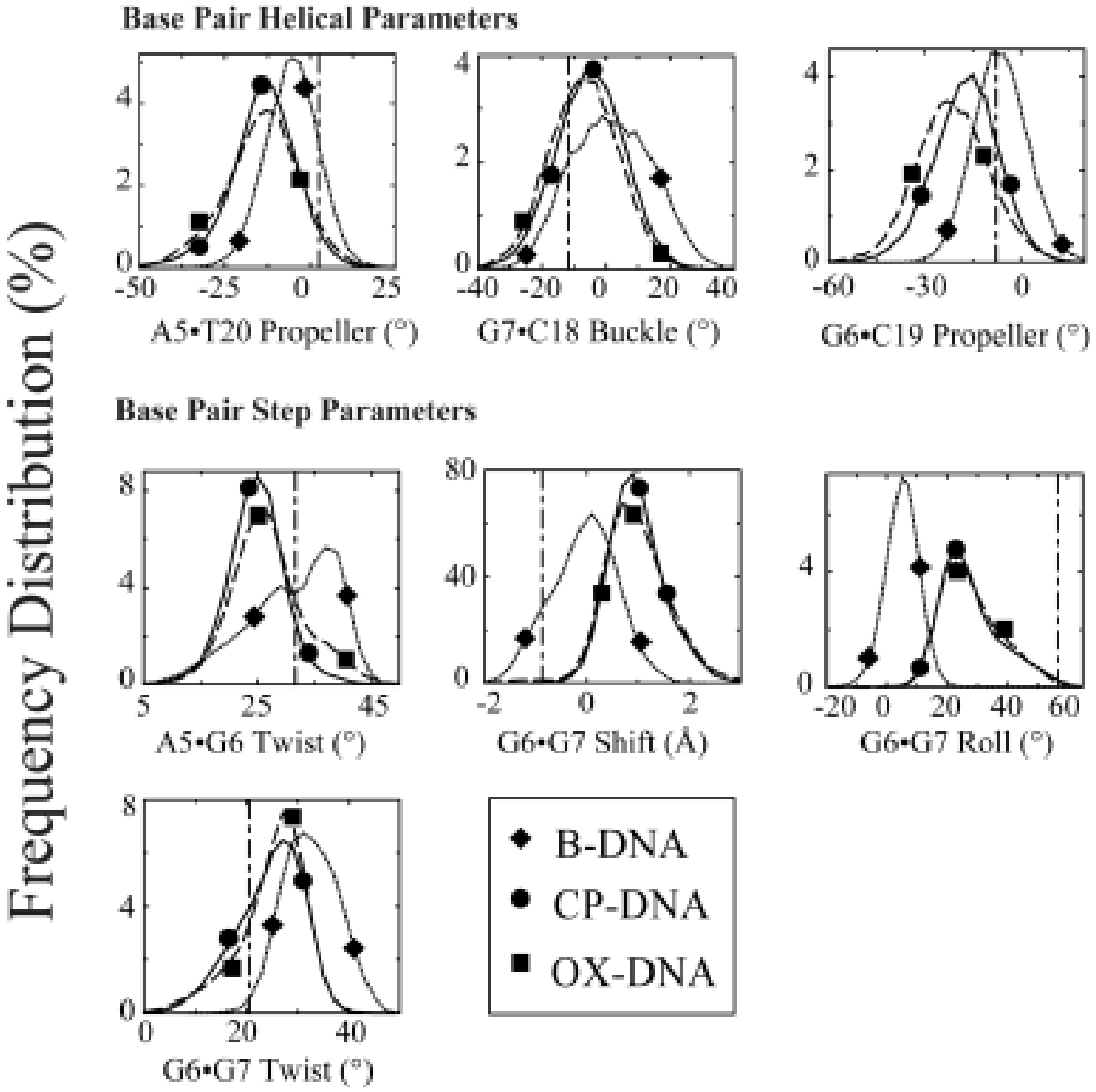

Figure 7.4: Hydrogen bond occupancy of the central four base-pairs. The hydrogen bond occupancy of the central four base-pairs was calculated as described in Methods. The standard Watson-Crick hydrogen bond occupancy of base-pairs for B-DNA (circle), CP-DNA (square) and OX-DNA (triangle) are shown.

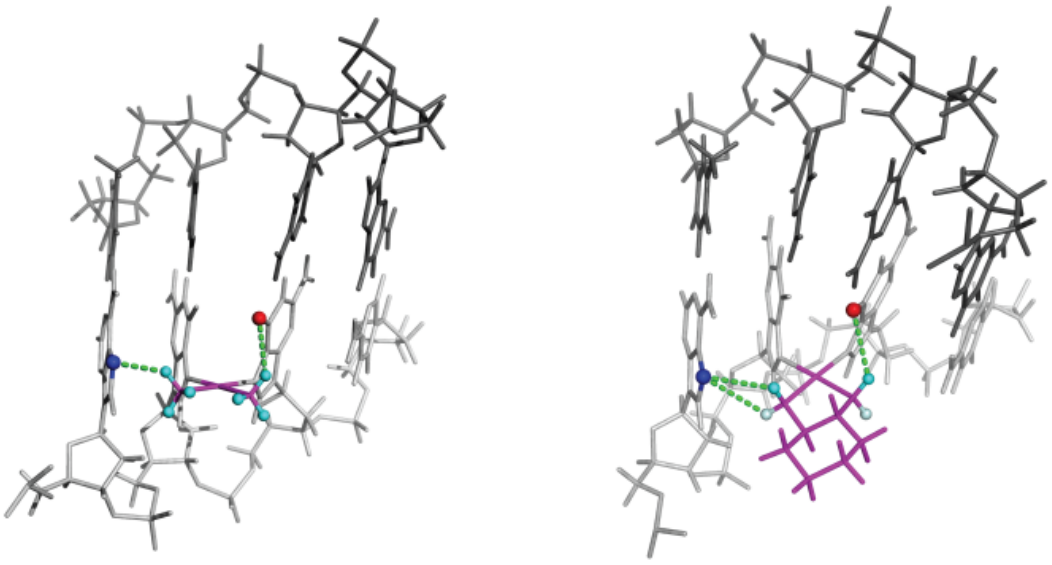

Figure 7.5: Hydrogen bonds between platinum carrier ligands and DNA. Structures illustrating observed hydrogen bond formation between Pt-amine hydrogen atoms and the surrounding DNA bases are shown for both the CP-DNA adduct (left) and the OX-DNA adduct (right). The DNA backbone is shown in light gray for the strand containing the Pt-DNA adduct and darker gray for the complementary strand. The Pt adduct including the carrier ligand is shown in maroon. For the CP adduct, all amine hydrogen atoms are in equilibrium and are shown in cyan. For the OX adduct the amine hydrogen atoms are not in equilibrium. The equatorial hydrogen is shown in cyan and the axial hydrogen is shown in light blue. The broken lines illustrate the potential hydrogen bonds between the platinum amine groups and the DNA bases.

adduct is 74.2% and 47.2%, respectively. Thus, for the CP-DNA adduct the occupancy is greater for the hydrogen bond on the 5' side of the adduct, suggesting that the CP-DNA adduct is preferentially oriented towards the 5' side.

For the OX-DNA adduct the situation is somewhat more complicated because the axial and equatorial amine hydrogen atoms are not equivalent (Fig. 7.5). Thus, on the 5' side of the adduct, the hydrogen bond between the axial hydrogen and N7 of A5 is not equivalent to the hydrogen bond between the equatorial hydrogen and N7 of A5. However, on the 3' side of the adduct only the equatorial hydrogen is in sufficient proximity to the O6 of G7 to form a hydrogen bond. Thus, for the OX-DNA adducts there are six possible combinations of hydrogen bond formation. The occupancy of

each is summarized in Table 2B. As seen for the CP-DNA adduct, differences in global DNA conformations associated with these hydrogen patterns were minimal. The OX-DNA adduct spends 34.2% of its time in conformations that allow formation of the 3' hydrogen bond only and 13.7% of its time in conformations that allow formation of 5' hydrogen bonds only. However, it also spends a significant amount of time (40.9%) in conformations that allow simultaneous formation of both a hydrogen bond between the equatorial hydrogen on the 5' side and N7 of A5 and a hydrogen bond between the equatorial hydrogen on the 3' side and O6 of G7. The total occupancy of the 5' and 3' hydrogen bonds for the OX-DNA adduct is 58.1% and 78.7%, respectively. Thus, for the OX-DNA adduct the occupancy is greater for the hydrogen bond on the 3' side of the adduct, suggesting that the OX-DNA adduct is preferentially oriented towards the 3' side.

7.3.5 DNA conformational dynamics

While the overall centroid structures of the CP-DNA and OX-DNA simulations were very similar, we observed significant differences between the two simulations in terms of DNA conformational dynamics. To determine the effect of CP- and OX-GG adducts on the conformational dynamics of DNA, the frequency distributions (fraction of the time spent in each conformation) from the trajectories of the CP-DNA, OX-DNA and undamaged DNA simulations were calculated using the program CURVES v5.3. From previous comparisons of CP- and OX-DNA adducts, it appeared that they were most likely to differ in terms of overall bend angle and the DNA helical parameters of the central four base pair region (TFL96; SWL01; WPH$^+$04; WBK$^+$07). The overall bend angle was calculated from the CURVES output using MadBend (http://monod.biomath.nyu.edu). No significant differences in the frequency distribution of bend angles was observed for the CP- and OX-DNA adduct simulations (data not shown). The frequency distributions of DNA helical parameters for the central four base pairs were taken directly from

the CURVES output and were analyzed for statistical significance by the Kolmogorov-Smirnov test (HPAT88). The Kologorov-Smirnov test determines how significantly two distributions differ from each other, without making any assumption regarding the distribution of data (non-parametric and distribution-free). Cases in which the frequency distributions of DNA helical parameters were significantly different from undamaged DNA for both CP- and OX-DNA adducts are indicated in bold in Supplementary Data, Table S3. The distribution of frequency values for those helical parameters between both types of Pt-DNA adducts and undamaged B-DNA are shown in Fig. 7.6. When comparing both Pt-DNA adducts to B-DNA, there were several striking differences identified, including buckle and propeller twist for the A5T20 base-pair, shear and propeller twist for the G6C19 base-pair, buckle for the G7C18 base-pair; roll at the G6-G7 base-pair step, and slide, tilt and roll at the G7-C8 base-pair step. These differences indicated that the conformational dynamics profile of B-DNA was altered by the platinum adducts and are consistent with the previously reported conformational distortions imposed on B-DNA by Pt-GG adducts (GL98; MSK$^+$01; TFL96; SWL01; YvBR$^+$95). At the G6-G7 base-pair step, the profiles of frequency distribution of roll were almost identical for CP-DNA and OX-DNA adducts, which was consistent with the GG dihedral angles for these two adducts.

The significance of differences in the frequency distributions of DNA helical parameters between CP- and OX-DNA adducts was also analyzed by the Kolmogorov-Smirnov test (HPAT88) (Table S4). The distributions of those helical parameters showing the greatest difference between CP- and OX-DNA adducts are shown in Fig. 7.7, along with the distribution pattern of undamaged B-DNA for comparison. When comparing CP-DNA adducts to OX-DNA adducts, there were some noticeable differences. For base-base helical parameters, differences were observed for propeller twist for the G6C19 and G7C18 base-pairs and buckle for the C8G17 base-pair. For base-pair step helical parameters, differences were observed for slide at the A5-G6, G6-G7 and G7-C8 base-

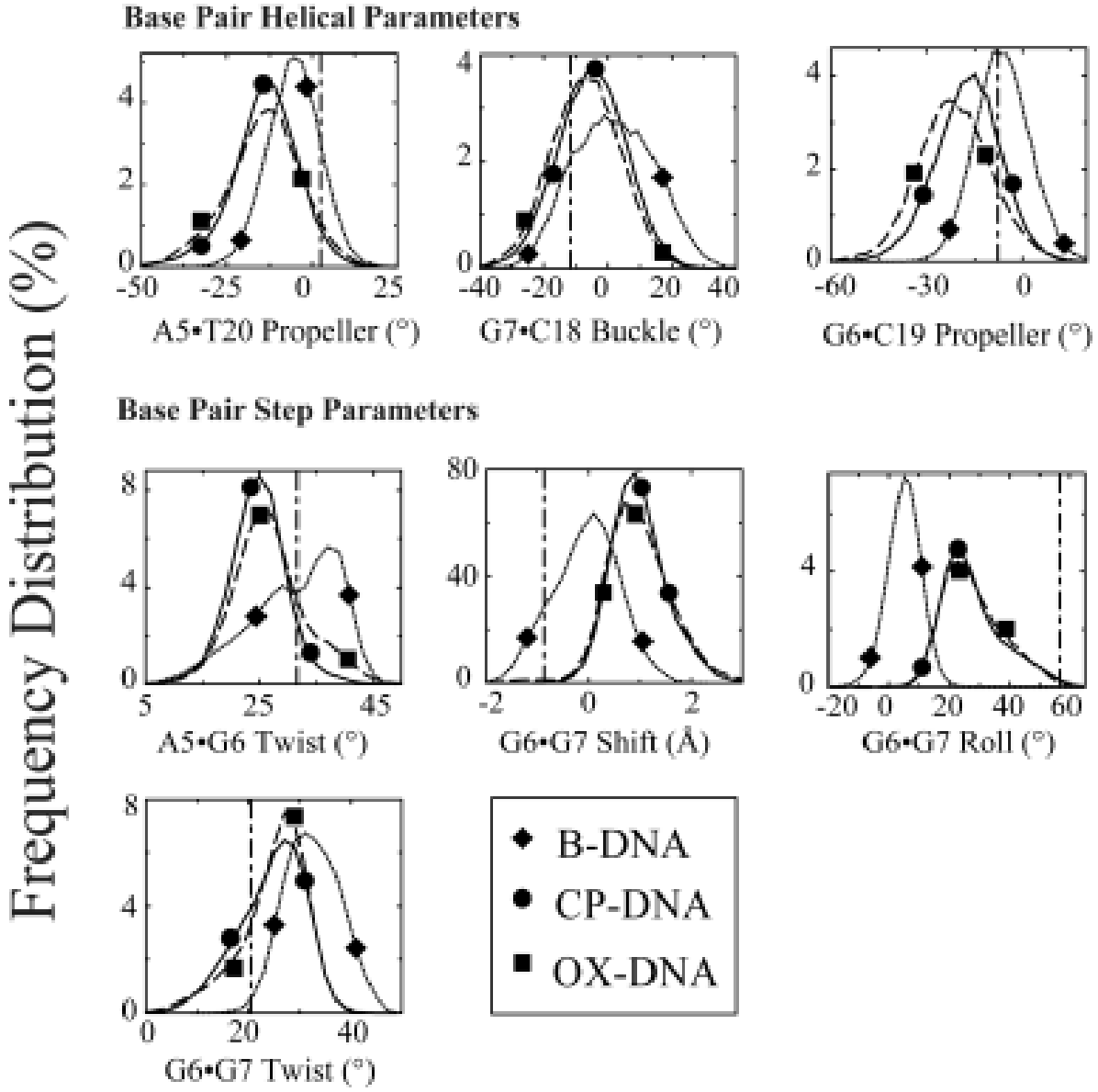

Figure 7.6: Frequency distributions of representative DNA duplex helical parameters for the central four base-pairs: differences between Pt-DNA adducts and undamaged DNA. Frequency distributions of each DNA duplex helical parameter for the central four base-pairs were calculated from the simulation trajectories by CURVES v5.3 as described in Methods. The frequency distribution histograms were calculated from the structures obtained at every picosecond over the final 6 ns of each equilibrated MD simulation. Thus, the histograms for CP-DNA and OX-DNA adducts were derived from 60,000 structures, while the histograms for undamaged DNA were derived from 30,000 structures. Selected frequency distributions that show differences between Pt-DNA adducts and undamaged DNA are shown. B-DNA (dotted lines with diamond symbols); CP-DNA (continuous lines with circle symbols); OX-DNA (long broken lines with square symbols). (The symbols do not represent the data ponts; rather they are shown to help distinguish the curves.)

pair steps and shift at the G7-C8 base-pair step. All of the above differences suggest that CP-DNA and OX-DNA adducts differ in conformational dynamics.

7.3.6 Correlation between platinum amine hydrogen bond formation and DNA conformational dynamics

As described above the percentage occupancy for hydrogen bonds between the Pt-amine hydrogens and nearby bases was greater on the 5' side for the CP-DNA adduct and the 3' side for the OX-DNA adduct. The data described above indicated that the CP- and OX-DNA adducts primarily influenced the conformational dynamics of the central four base-pairs. Thus, to determine whether the conformational dynamics of CP- and OX-DNA adducts might be influenced by the formation of these hydrogen bonds, the trajectory data for DNA helical parameters for the central four base pairs were separated according to the patterns of hydrogen bond formation (Fig. 7.8 and Fig. 7.9). The resulting frequency distributions were analyzed for significance by the Kolmogorov-Smirnov test (HPAT88) (Tables S5 and S6) and cases in which the frequency distribution of DNA helical parameters associated with 5' only hydrogen bonds were significantly different than the DNA helical parameters associated with 3' only hydrogen bonds are indicated in bold in Tables S5 and S6. The large number of examples in which the frequency distribution of DNA helical parameters is associated with the pattern of hydrogen bond formation (5' only versus 3' only) suggests that the conformational dynamics of many of the DNA helical parameters in the central four base-pair region are strongly correlated with the pattern of hydrogen bond formation.

Selected examples for the CP-DNA adduct are shown in Fig. 7.8. For most of the DNA helical parameters shown there is a clear difference in the distribution of DNA helical parameters when the hydrogen bond exists on the 5' side compared to when the hydrogen bond exists on the 3' side. In the case of the CP-DNA adduct, the two major hydrogen bond patterns are hydrogen bond on 5' side only (40.2%) and hydrogen

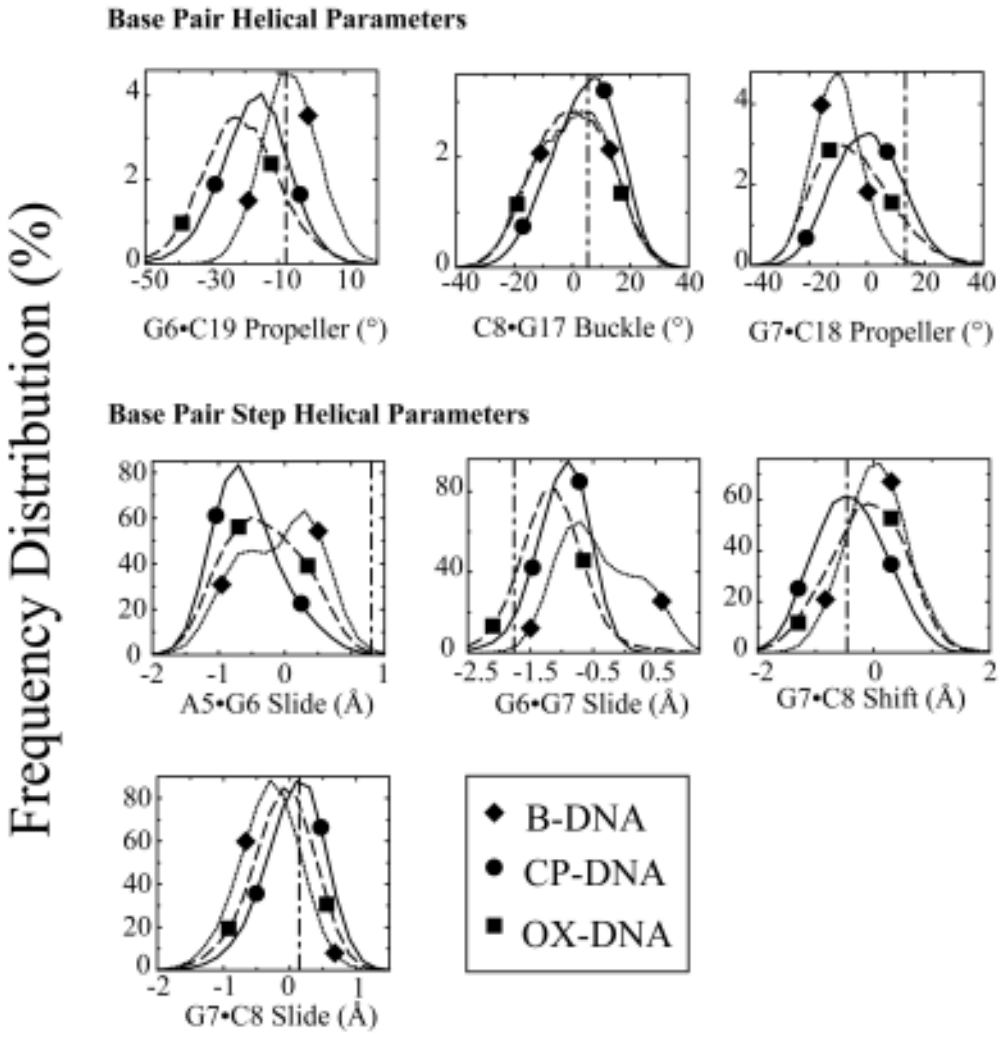

Figure 7.7: Frequency distributions of representative DNA duplex helical parameters for the central four base-pairs: differences between CP-DNA and OX-DNA adducts. Frequency distributions of each DNA duplex helical parameter for the central four base-pairs were calculated from the simulation trajectories by CURVES v5.3 as described in Methods. The frequency distribution histograms were calculated from the structures obtained at every picosecond over the final 6 ns of each equilibrated MD simulation. Thus, the histograms for CP-DNA and OX-DNA adducts were derived from 60,000 structures, while the histograms for undamaged DNA were derived from 30,000 structures. Selected frequency distributions that show differences between CP-DNA and OX-DNA adducts are shown. B-DNA (dotted lines with diamond symbols), CP-DNA (continuous lines with circle symbols); OX-DNA (long broken lines with square symbols). (The symbols do not represent the data points; rather they are shown to help distinguish the curves.) The vertical dash-dot lines show the value of the corresponding DNA helical parameter on the 3 side of the adduct in the crystal structure of the HMGCP-DNA adduct.

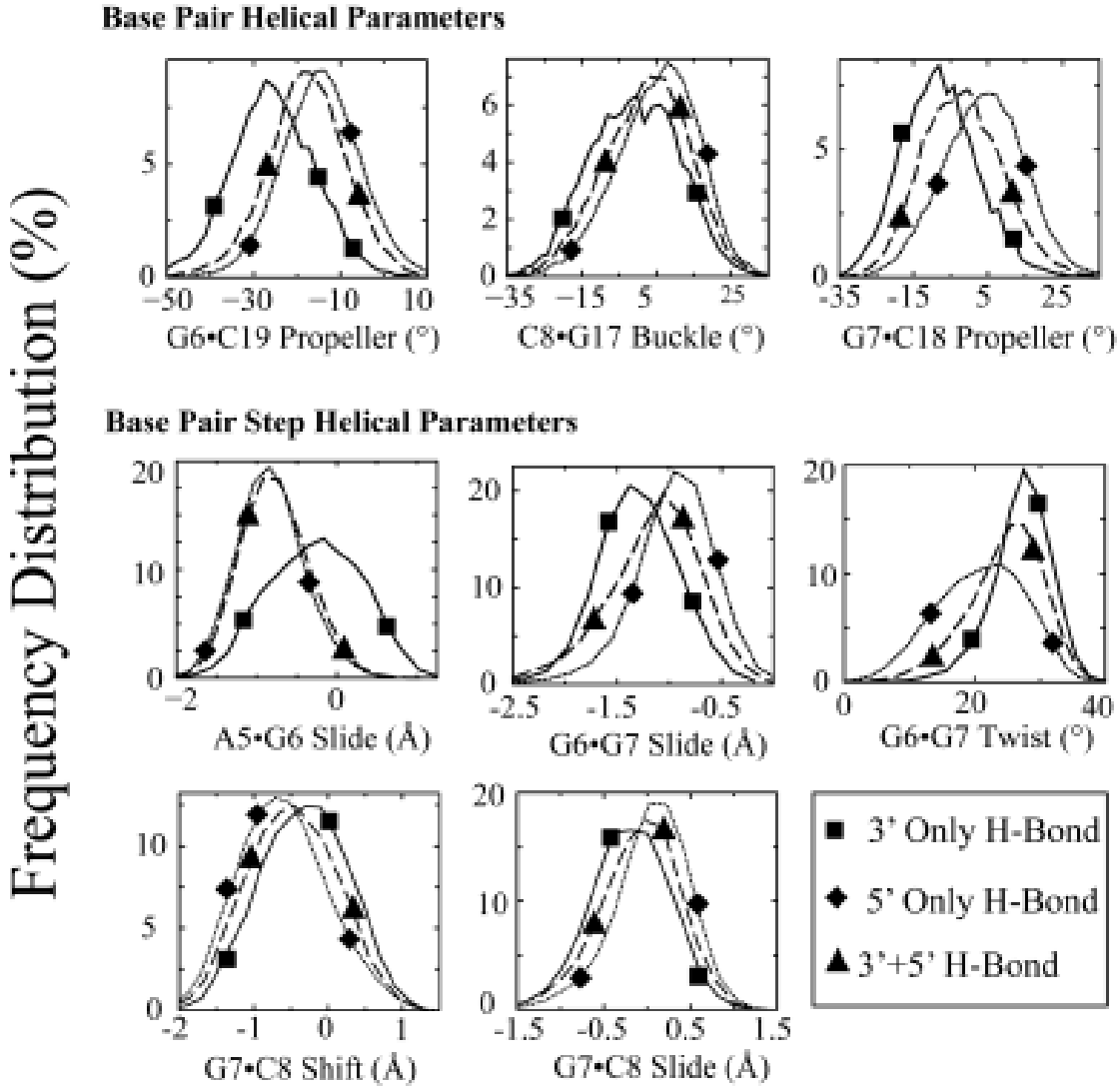

Figure 7.8: Effect of hydrogen bonding patterns on frequency distributions of selected DNA duplex helical parameters for the central four base-pairs of CP-DNA adducts. The effect of hydrogen bond patterns between the Pt-amine hydrogen atoms and adjacent DNA bases on the frequency distributions of each DNA duplex helical parameter for the central four base-pairs of the CP-DNA adduct were determined as described in Methods. The frequency distribution histograms were obtained from all of the CP-DNA structures containing a particular hydrogen bond pattern (Table 2). Thus, the histograms for 5 only, 5 plus 3 and 3 only hydrogen bonds were obtained from 24,120, 20,400 and 7980 structures, respectively. Selected frequency distributions that show difference in the distribution of DNA helical parameters when the hydrogen bond exists on the 5 side compared to when the hydrogen bond exists on the 3 side of the CP-DNA adduct are shown. 3 Hydrogen bond only (continuous lines with square symbols); 5 hydrogen bond only (dotted lines with diamond symbols); hydrogen bonds on both 3 and 5 sides (long broken lines with upward triangle symbols). (The symbols do not represent the data points; rather they are shown to help distinguish the curves.) The frequency distributions for no hydrogen bonds on either the 3 side or 5 side are not shown for ease of viewing.

bonds on both the 5' and 3' side (34.0%). Furthermore, the frequency distributions of DNA helical parameters associated with these two hydrogen bond patterns are generally either almost identical (A5·G6 slide) or very similar (G6·C19 propeller twist, G7·C18 propeller twist, C8·G17 buckle, G6-G7 slide) G7-C8 shift, G7-C8 slide). These data suggest that DNA conformations in which the hydrogen bonds are seen on both the 5' and 3' side may represent conformational transitions between the 5' only and 3' only hydrogen bond conformations.

For the OX-DNA adduct (Fig. 7.9) there also appear to be clear differences in the distribution of DNA helical parameters when the hydrogen bond exists on the 5' side only (either for the axial hydrogen or the equatorial hydrogen) compared to when the hydrogen bond exists on the 3' side only. In the case of the OX-DNA adduct, the two major hydrogen bond patterns are a hydrogen bond with the equatorial hydrogen on 3' side only (34.3%) and hydrogen bonds with both the equatorial hydrogen on the 5' and the equatorial hydrogen on the 3' side (40.9%). As with the CP-DNA adduct, the frequency distributions of DNA helical parameters associated with these two hydrogen bond patterns are generally either almost identical (G6-G7 twist, G7-C8 shift, G7-C8 slide) or very similar (G6·C19 propeller twist, G7·C18 propeller twist, C8·G17 buckle, G6-G7 slide). These data again suggest that DNA conformations in which the hydrogen bonds are seen on both the 5' and 3' side may represent conformational transitions between the 5' only and 3' only hydrogen bond conformations.

Finally, when one compares the data in Fig. 7.8 and Fig. 7.9 with the data in Fig. 7.7 it becomes apparent that most of the conformational differences associated with hydrogen bond formation in the CP- and OX-DNA adducts are highly correlated with the conformational differences seen between CP- and OX-DNA adducts in Fig. 7.7. For example, when one looks at propeller twist at the G6·C19 base-pair, the two major distribution patterns associated with 5' hydrogen bond formation (5' only and both 5' and 3') of the CP-DNA adduct (representing 74.2% of the total hydrogen bond oc-

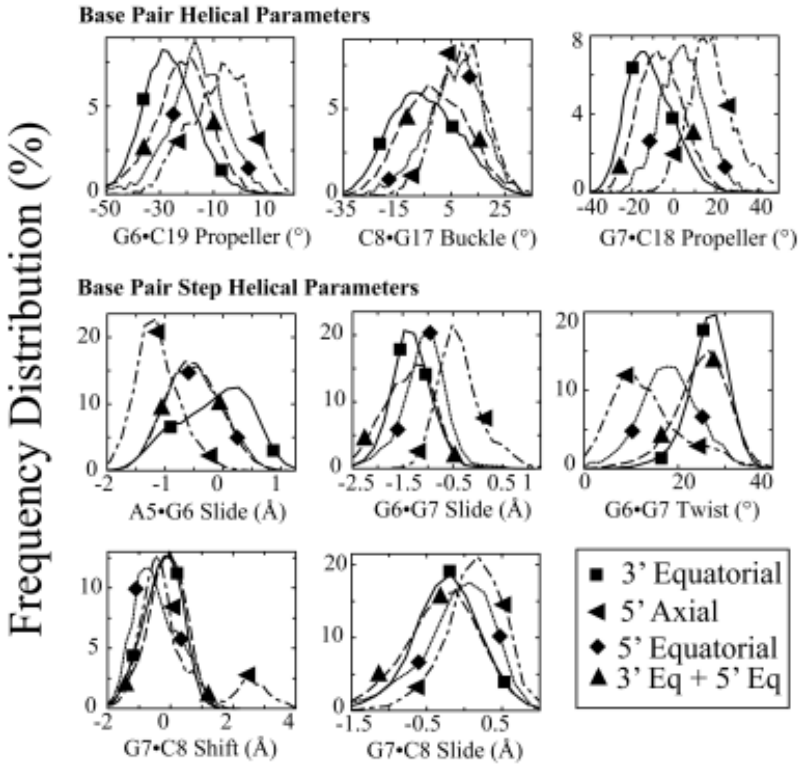

Figure 7.9: Effect of hydrogen bonding patterns on frequency distributions of selected DNA duplex helical parameters for the central four base-pairs of OX-DNA adducts. The effect of hydrogen bond patterns between the Pt-amine hydrogen atoms and on adjacent DNA bases on the frequency distributions of each DNA duplex helical parameter for the central four base-pairs of the OX-DNA adduct were determined as described in Methods. The frequency distribution histograms were obtained from all of the OX-DNA structures containing a particular hydrogen bond pattern (Table 2). Thus, the histograms for 5' axial only, 5' equatorial only, 5' axial plus 3' equatorial, and 3' equatorial only hydrogen bonds were obtained from 3600, 4620, 24540 and 20500 structures, respectively. Selected frequency distributions that show difference in the distribution of DNA helical parameters when the hydrogen bond exists on the 5' side compared to when the hydrogen bond exists on the 3' side of the OX-DNA adduct are shown. 3' Equatorial hydrogen bond only (continuous lines with square symbols); 5' axial hydrogen bond only (dot-dashed lines with circle symbols); 5' equatorial hydrogen bond only (dotted lines with diamond symbols); hydrogen bonds with both 3' equatorial and 5' equatorial hydrogens (long broken lines with upward triangle symbols). (The symbols do not represent the data points; rather they are shown to help distinguish the curves.) The frequency distributions for no hydrogen bonds on either the 3' side or 5' side and hydrogen bond formation for both the 3' equatorial and 5' axial hydrogen atoms are not shown for ease of viewing.

cupancy) are centered at -18° to -16° (Fig. 7.8), while the two major distribution patterns associated with 3' hydrogen bond formation (3' only and equatorial 5' plus 3') for the OX-DNA adduct (representing 75.2% of the total hydrogen bond occupancy) are centered at -29° and -21° (Fig. 7.9). This precisely accounts for the difference in frequency distribution patterns seen between CP- and OX-DNA adducts (-18° for CP and -25° for OX) for G6·C19 propeller twist in Fig. 7.7. Similar comparisons can be made for G7·C18 propeller twist, C8·G17 buckle, A5-G6 slide, G6-G7 slide, and G7-C8 slide.

However, these patterns are not universal. For G7-C8 shift, the differences in distribution patterns between CP- and OX-DNA adducts in Fig. 7.7 correlate with the major distribution patterns for CP- and OX-DNA adducts in Fig. 7.8 and Fig. 7.9, but the differences between 5' only and 3' only hydrogen bond conformational distributions are not significant for either CP- or OX-DNA adducts (Supplementary Data, Tables S5 and S6). At the opposite extreme, the 5' only and 3' only hydrogen bond conformational distributions are significantly different for G6-G7 twist (Supplementary Data, Tables S5 and S6)), but the major distribution patterns (5' only and both 5' and 3' for CP and 3' only and both 5' axial and 3' for OX) are very similar ($22 - 26^\circ$ for CP and 28° for OX). Thus, the distribution patterns for CP- and OX-DNA adducts for G6-G7 twist are not significantly different (Supplementary Data, Table S4).

7.4 Discussion

7.4.1 Accuracy of the MD simulations

Partial atomic charges for the platinum atom and the surrounding atoms of Pt-GG adducts are not available in the standard AMBER force field and must, therefore be developed ab initio. The previously available charges for CP-GG adducts were adopted by Yao et al.(YPM94) from the ab initio calculations of the [Pt(NH3)3(Ade)]2+ complex

by Kozelka et al. (KSB$^+$93) over 13 years ago. The same atomic charge parameters were used by Scheef et al. (SBH99) for simulations of the OX-GG adduct, with the sole exception that the charge on the amine nitrogen was modified to reflect its attachment to the cyclohexane ring. Although the partial charge parameters on the CP-GG adduct have been updated by Elizondo-Riojas and Kozelka (ERK01), those charge parameters were not actually reported. Because the partial charge parameters for the CP-GG adduct were not derived empirically for a Pt-GG adduct and the OX-GG partial charge parameters had never been determined empirically, we have developed new partial charge parameters for both CP-GG and OX-GG adducts. The new partial charges on the platinum atom were similar to the previous published values (ERK01; YPM94). However, the new partial charges on the surrounding nitrogen atoms were significantly different. Therefore, it was important to validate the accuracy of our new partial charges in MD simulations.

Our MD simulations converged to a common set of structures within the first 4 ns that were independent of the starting structure and MD initial velocities and these structures reproduced the crystal and NMR solution structures of the same adducts in DNA by several criteria. First, the simulation structures in the final 6 ns had an average RMSD of $\leq$ 3Å with respect to our NMR solution structures (Fig. 7.1). Second, the centroid structures had RMSD values of $\approx$ 4.1Å compared to the corresponding crystal structures (TFL96; SWL01) and around 3.1Å compared to the corresponding NMR structures (WPH$^+$04; WBK$^+$07). Third, for the central four base-pair region there were less than 3% violations of NMR-derived inter-proton distance constraints for the CP-GG and OX-GG adducts and the majority of these violations involved sugar pucker protons. Finally, both CP-DNA and OX-DNA adducts in our MD simulations showed significantly increased roll angle at the G6·G7 base-pair step and the GG dihedral angle with respect to the undamaged B-DNA, which is consistent with both the crystal and NMR solution structures of the Pt-DNA adducts (MSK$^+$01; WPH$^+$04; TFL96; SWL01;

WBK$^+$07). By all of these criteria, the MD simulations were excellent representations of both the crystal and NMR solution structures of the same adducts. In addition, these MD simulations provided a better estimation of the α angles and displacements of the platinum atom out of the 5' and 3' guanine planes than the previous MD simulation of the CP-GG adduct by Elizondo-Riojas and Kozelka (ERK01).

7.4.2 The DNA duplex is more distorted on the 5' side of the adduct than on the 3' side

We found a significantly greater decrease of occupancy of standard Watson-Crick hydrogen bonds for the 5' A5·T20 and 5' G6-C19 base-pairs than for the 3' G7-C18 base-pair for both CP-GG and OX-GG adducts, which suggests that the Pt-GG adduct is more distorted on the 5' side of the adduct than on the 3' side of the adduct. This is consistent with a number of biological and structural studies that have indicated that Pt-GG adducts are more distorted on the 5' side of the adduct than on the 3' side of the adduct. For example, several authors have reported that the majority of misinsertion mutations occur opposite the 5'G of CP-GG adducts (BYLE93; BAT91; BDF87), and we and others have shown that hpol η has significantly greater difficulty extending the DNA chain past the 5'G than the 3'G of either CP-GG or OX-GG adducts (BVH$^+$03; MKIH00; WJPP01). With respect to structural studies, Marzilli et al. (MSK$^+$01) have reported a faster exchange rate between the water and imino protons of the 5'G of CP-GG adducts and we have shown a faster exchange rate between water and imino protons of the 5'G of both the OX-GG (WPH$^+$04) and CP-GG adducts (WBK$^+$07). Marzilli et al. (MSK$^+$01) also reported an unusually large positive shift and slide at 5' X/G base pair-step, and Elizondo-Riojas and Kozelka (ERK01), in their MD simulations, have shown greater mobility of 5' G-C base-pair with respect to the base-pair 5' to the Pt-GG adduct. All of these previous studies have indicated that the 5' G-C base-pair was more flexible and/or more distorted than the 3' G-C base-pair.

7.4.3 Orientation of CP-DNA and OX-DNA adducts

In the crystal structure of CP-DNA adduct, a hydrogen bond was reported between the hydrogen atom of the 5' oriented ammine NH3 ligand and an oxygen atom of the phosphate group on the 5' side of the Pt-GG adduct. In contrast, in the crystal structure of OX-DNA adduct, a hydrogen bond was found between the 3' oriented amine NH2 group of the diaminocycloheaxane ligand and the oxygen atom O6 of the 3'G (SWL01). Thus it has been suggested that the cis-diammine carrier ligand of cisplatin was oriented more towards the 5' side of the adduct and the (trans-R,R)1,2-diaminocyclohexane carrier ligand of oxaliplatin was oriented more towards the 3' side of the adduct. However, neither hydrogen bond was observed in NMR solution structures of the CP- and OX-GG adducts (GL98; WPH$^+$04; WBK$^+$07). Here, we assessed the occupancy of these hydrogen bonds for both CP-DNA and OX-DNA adducts and found that the occupancy of the hydrogen bond between the 5'-oriented NH3 and the phosphate oxygen reported in the crystal structure of the CP-DNA adduct was less than 1% for both CP-DNA and OX-DNA adducts. In their work based on MD simulations of the CP-DNA adduct, Elizondo-Riojas and Kozelka (ERK01) also reported that the formation of this hydrogen bond is a very rare event, unless one considers the possibility of hydrogen bonds bridged by one or two water molecules. The fact that this hydrogen bond is not observed at any significant level in either NMR solution structures or MD simulations suggests that the formation of such a hydrogen bond is a rare event and that the crystal structure of the CP-DNA adduct happened to capture this rare conformation, possibly due to crystal packing.

However, we did observe significant occupancy of the hydrogen bond formed between the hydrogen atom of the 5' ammine NH3 ligand of CP or the 5' amine NH2 group of diaminocyclohexane ligand of OX and the nitrogen atom N7 of A5. The occupancy of this hydrogen bond was higher for the CP-DNA adduct (74.2%) than the OX-DNA adduct (58.1%). We also observed significant occupancy of the hydrogen bond formed

between the hydrogen atom of the 3' ammine NH3 ligand of CP or the 3' amine NH2 group of diaminocyclohexane ligand of OX and the oxygen atom O6 of the 3' G7, which had been previously reported in the crystal structure of the OX-DNA adduct, but not the CP-DNA adduct. The occupancy of this hydrogen bond was higher for the OX-DNA adduct (78.7%) than the CP-DNA adduct (47.2%). The differences in occupancy of hydrogen bonds between the ammine/amine hydrogen atoms of the CP/OX and the bases on the 5' and 3' side of the Pt-GG adducts suggest a difference in the orientation of Pt carrier ligands with respect to the DNA between CP- and OX-GG adducts. The cis-diammine carrier ligand of CP appears to be oriented more towards the 5' side of the adduct and the (trans-R,R)1,2-diaminocyclohexane carrier ligand of OX appears to be oriented more towards the 3' side of the adduct. Thus, our data are consistent with the observations reported by Lippard and colleagues (TFL96; SWL01) with respect to the preferential orientation of carrier ligands of CP-DNA and OX-DNA adducts. This preferential formation of hydrogen bonds on the 5' side of the CP-DNA adducts and on the 3' side of OX-DNA adducts is highly correlated with differences in conformation and conformational dynamics of DNA containing CP- and OX-GG adducts. We postulate that the hydrogen bond formation between the Pt-amine hydrogen atoms and adjacent bases drives some of the important conformational differences between CP- and OX-DNA adducts.

These experiments do not address the possible effect of sequence context on the orientation of CP- and OX-GG adducts. Our MD simulations were done in the AGGC sequence context, while the crystal structures reported by Lippard and colleagues (TFL96; SWL01) were in the TGGT sequence context, which suggests that the relative orientation of CP- and OX-GG adducts are similar in those sequence contexts. However, in the MD simulations of the CP-GG adduct in the CGGA sequence context by Elizondo-Riojas and Kozelka (ERK01), a hydrogen bond was observed on the 3' side of the adduct and between the 3' NH3 ligand of the CP-GG adduct and N7 of the A on the 3' side

of the adduct. This observation may help explain previous reports that an A on the 3' side of Pt-GG adducts has a significant affect on the binding affinity and binding specificity of DNA damage recognition proteins for CP-GG adducts. Experiments are currently underway to determine the effect of sequence context on the orientation and conformational dynamics of CP- and OX-GG adducts.

7.4.4 Differences in conformational dynamics between CP-DNA and OX-DNA adducts

From the comparison of frequency distribution of DNA duplex helical parameters, we observed several differences which appear to reflect the distinct conformational dynamics between CP-DNA and OX-DNA adducts. While the total range of DNA conformations explored by CP- and OX-GG adducts is very similar, the fraction of time spent in these conformation by CP- and OX-GG adducts differed with respect to several DNA helical parameters in the vicinity of the Pt-GG adducts. We postulate that these differences in conformational dynamics could allow differential recognition of CP- and OX-GG adducts by critical DNA-binding proteins that influence the cytotoxic response to these adducts. While the differences in conformational dynamics between CP- and OX-DNA adducts are relatively small, they are fully consistent with the 1.5- to twofold differences in recognition of the adducts by most of the DNA-binding proteins studied to date.

We postulate that CP-GG adducts spend a greater percentage of time in conformations favorable for binding of mismatch repair and HMG-domain DNA-binding proteins, while OX adducts spend a greater percentage of time in conformations(s) favorable for bypass by hpol β and hpol η. For example, the HMG domain of both HMG-A1 and HMG-B1 has been shown to bind with higher efficiency to CP-GG adducts than to OX-GG adducts14. and the crystal structure of the HMG-CP-DNA complex has been reported (ORH^{+}99). Since HMG binds to the CP-DNA adduct on the 3'side of the adduct only,26 the DNA helical parameters on the 3' side of the HMG-CP-DNA com-

plex are shown as vertical broken lines in Fig. 7.7 for comparison with the frequency distributions of helical parameters for the CP-GG and OX-GG adducts. For propeller twist of the G7·C18 base-pair, buckle of the C8·G17 base-pair and shift and slide at the G7-C8 base-pair step, the CP-GG adduct appears to spend a greater percentage of its time in a conformation that is similar to the conformation of the HMG-CP-GG complex than the OX-GG adduct. If any of these conformations are characteristic of the conformation of the protein-DNA complex in the initial recognition step, they might facilitate the recognition of the CP-GG adduct by the HMG domain. The conformational dynamics differences at the G7-C8 base-pair step might be particularly critical because both the affinity of the HMG domain for the CP-GG adduct (WCSL01; CMHL00; DL97) and the ability of the HMG domain to discriminate between CP-GG and OX-GG adducts14 is highly dependent on the base to the 3' side of the adduct.

There are a number of important limitations inherent in such comparisons. We are comparing a crystal structure of the HMG-CP-GG DNA complex with unrestrained simulations of the CP-GG and OX-GG DNA adducts; the final, stable protein DNA complex likely has a different conformation from the initial protein-DNA recognition complex; and the sequence context of the CP-GG adduct in the HMG-CP-GG complex was different from the one in the CP-GG and OX-GG simulations. To better characterize those conformational differences that are important for differential protein recognition, experiments are underway to simulate the CP-GG and OX-GG adducts in the same AGGC sequence context in complex with the HMG domains of one or more proteins that discriminate between the adducts.

Chapter 8

Conclusion

8.1 Concluding Remarks

This dissertation presents several physically-principled computational modeling and automation techniques to explore biomolecular dynamics at atomic scales. In chapter 2, we discussed multiscale modeling of nucleosome dynamics using discrete molecular dynamics. Using DMD simulations, we show that our simplistic model recapitulates the stability and simulates the dynamics of nucleosomes for experimentally relevant timescales. We find that in our simulations of mononucleosomes, histone tails form strong salt-bridge interactions with nucleosomal DNA, suggesting their direct role in forming higher-order chromatin structure. Based on constant-temperature discrete molecular dynamics simulations, we find that bending across the H3-H3 interface is a prominent mode of nucleosome dynamics. The dynamics of the NCP is dominated by histone tails with subsequent normal modes composed of large-scale interhistone motions. Analysis of frequencies of histone-DNA contacts formed in constant-temperature DMD simulations shows persistent contacts formed with C-terminal H2A and the nucleosomal dyad axis, thereby suggesting functional roles of the H2A C-terminal domain. We determine a coarse-grained phase space of the NCP under altering potentials of histone-DNA interactions. Our approach of amalgamating rapid conformation sampling techniques like DMD with

coarse-grained models of nucleosome is applicable for analyzing the effects of histone variants and the effects of DNA sequence on nucleosome positioning.

Discrete molecular dynamics simulation facilitate rapid exploration of protein conformational space. Automation of discrete molecular dynamics simulations of proteins is presented in chapter 6. Simulations of protein folding, protein unfolding, thermodynamic scan, simulated annealing and p_{fold} analysis are enabled using the iFold server (http://iFold.dokhlab.org). RNA molecules with novel functions have revived interest in the accurate prediction of RNA three-dimensional (3D) structure and folding dynamics. In chapter 5 we presented a robust computational approach for rapid folding of RNA molecules based on replica-exchange discrete molecular dynamics simulations. Automation of ab initio RNA structure prediction and folding thermodynamic analyses is presented in chapter 6. We developed a simplified RNA model for discrete molecular dynamics (DMD) simulations, incorporating base-pairing and base-stacking interactions. We demonstrate correct folding of over 150 structurally diverse RNA sequences. The majority of DMD-predicted 3D structures have ¡ 4 Å deviations from experimental structures. The secondary structures corresponding to the predicted 3D structures consist of 94% native base-pair interactions. Folding thermodynamics and kinetics of tRNA(Phe), pseudoknots, and mRNA fragments in DMD simulations are in agreement with previous experimental findings. Folding of RNA molecules features transient, non-native conformations, suggesting non-hierarchical RNA folding. Our method allows rapid conformational sampling of RNA folding, with computational time increasing linearly with RNA length. We envision this approach as a promising tool for RNA structural and functional analyses. The iFoldRNA server (http://iFoldRNA.dokhlab.org) automates rapid tertiary structure prediction and probing folding thermodynamics of RNA molecules. Replica-exchange discrete molecular dynamcis simulations are used to predict RNA structure using a simplified three-bead/nucleotide model.

In chapter 6, we presented our work on ab initio RNA structure prediction and fold-

ing thermodynamic analyses using the iFoldRNA server (http://iFoldRNA.dokhlab.org). iFoldRNA is a web-based methodology for tertiary structure prediction of RNA molecules with near atomic resolution accuracy. The iFoldRNA server also enables computational analyses of RNA folding thermodynamics. iFoldRNA rapidly explores RNA conformations using discrete molecular dynamics simulations of input RNA sequences. Starting from simplified linear-chain conformations, RNA molecules (< 50 nt) fold to native-like structures within half an hour of simulation, facilitating rapid RNA structure prediction. All-atom reconstruction of energetically stable conformations generates iFoldRNA predicted RNA structures. The predicted RNA structures have ≤ 5Å root mean squre deviations (RMSDs) from corresponding experimentally derived structures. Thermodynamic analyses of RNA molecules can be performed using the Folding Thermodynamics module of the iFoldRNA server. The RNA folding parameters including specific heat, contact maps, simulation trajectories, gyration radii, RMSDs from native state, fraction of native-like contacts are accessible from iFoldRNA.

In chapter 7, we presented molecular dynamics simulations of dodecamer DNA adducted with platinum-based drugs. The simulations converged to a common set of structures within the first 4 ns that were independent of the starting structure and MD initial velocities and these structures reproduced the crystal and NMR solution structures of the same adducts in DNA by several criteria. First, the simulation structures in the final 6 ns had an average RMSD of $\leq$ 3Å with respect to our NMR solution structures (Figure 1). Second, the centroid structures had RMSD values of $\approx$ 4.1Å compared to the corresponding crystal structures (TFL96; SWL01) and around 3.1$\mathring{A}$ compared to the corresponding NMR structures (WPH$^+$04; WBK$^+$07). Third, for the central four base-pair region there were less than 3% violations of NMR-derived interproton distance constraints for the CP-GG and OX-GG adducts and the majority of these violations involved sugar pucker protons. Finally, both CP-DNA and OX-DNA adducts in our MD simulations showed significantly increased roll angle at the G6·G7

base-pair step and the GG dihedral angle with respect to the undamaged B-DNA, which is consistent with both the crystal and NMR solution structures of the Pt-DNA adducts (MSK$^+$01; WPH$^+$04; TFL96; SWL01; WBK$^+$07). By all of these criteria, the MD simulations were excellent representations of both the crystal and NMR solution structures of the same adducts. In addition, these MD simulations provided a better estimation of the α angles and displacements of the platinum atom out of the 5' and 3' guanine planes than the previous MD simulation of the CP-GG adduct by Elizondo-Riojas and Kozelka (ERK01). We also report a significantly greater decrease of occupancy of standard Watson-Crick hydrogen bonds for the 5' A5·T20 and 5' G6·C19 base-pairs than for the 3' G7·C18 base-pair for both CP-GG and OX-GG adducts, which suggests that the Pt-GG adduct is more distorted on the 5' side of the adduct than on the 3' side of the adduct. This is consistent with a number of biological and structural studies that have indicated that Pt-GG adducts are more distorted on the 5' side of the adduct than on the 3' side of the adduct. We have shown that hpol η has significantly greater difficulty extending the DNA chain past the 5'G than the 3'G of either CP-GG or OX-GG adducts (BVH$^+$03; MKIH00; WJPP01).

We assessed the occupancy of these hydrogen bonds for both CP-DNA and OX-DNA adducts and found that the occupancy of the hydrogen bond between the 5'-oriented NH3 and the phosphate oxygen reported in the crystal structure of the CP-DNA adduct was less than 1% for both CP-DNA and OX-DNA adducts. We observed significant occupancy of the hydrogen bond formed between the hydrogen atom of the 5' ammine NH3 ligand of CP or the 5' amine NH2 group of diaminocyclohexane ligand of OX and the nitrogen atom N7 of A5. The occupancy of this hydrogen bond was higher for the CP-DNA adduct (74.2%) than the OX-DNA adduct (58.1%). We also observed significant occupancy of the hydrogen bond formed between the hydrogen atom of the 3' ammine NH3 ligand of CP or the 3' amine NH2 group of diaminocyclohexane ligand of OX and the oxygen atom O6 of the 3' G7, which had been previously reported

in the crystal structure of the OX-DNA adduct, but not the CP-DNA adduct. The occupancy of this hydrogen bond was higher for the OX-DNA adduct (78.7%) than the CP-DNA adduct (47.2%). The differences in occupancy of hydrogen bonds between the ammine/amine hydrogen atoms of the CP/OX and the bases on the 5' and 3' side of the Pt-GG adducts suggest a difference in the orientation of Pt carrier ligands with respect to the DNA between CP- and OX-GG adducts. The cis-diammine carrier ligand of CP appears to be oriented more towards the 5' side of the adduct and the (trans-R,R)1,2-diaminocyclohexane carrier ligand of OX appears to be oriented more towards the 3' side of the adduct. The preferential formation of hydrogen bonds on the 5' side of the CP-DNA adducts and on the 3' side of OX-DNA adducts is highly correlated with differences in conformation and conformational dynamics of DNA containing CP- and OX-GG adducts. We postulate that the hydrogen bond formation between the Pt-amine hydrogen atoms and adjacent bases drives some of the important conformational differences between CP- and OX-DNA adducts.

Our simulations do not address the possible effect of sequence context on the orientation of CP- and OX-GG adducts. From the comparison of frequency distribution of DNA duplex helical parameters, we observed several differences which appear to reflect the distinct conformational dynamics between CP-DNA and OX-DNA adducts. While the total range of DNA conformations explored by CP- and OX-GG adducts is very similar, the fraction of time spent in these conformation by CP- and OX-GG adducts differed with respect to several DNA helical parameters in the vicinity of the Pt-GG adducts. We postulate that these differences in conformational dynamics could allow differential recognition of CP- and OX-GG adducts by critical DNA-binding proteins that influence the cytotoxic response to these adducts. While the differences in conformational dynamics between CP- and OX-DNA adducts are relatively small, they are fully consistent with the 1.5- to twofold differences in recognition of the adducts by most of the DNA-binding proteins studied to date. We postulate that CP-GG adducts spend

a greater percentage of time in conformations favorable for binding of mismatch repair and HMG-domain DNA-binding proteins, while OX adducts spend a greater percentage of time in conformations(s) favorable for bypass by hpol β and hpol η.

Collectively, the research presented in this dissertation highlights applications of computational modeling and automation techniques to explore biomolecular interaction and dynamics at atomic levels and generate experimentally-testable hypotheses.

Bibliography

[ABD+99] C. Anselmi, G. Bocchinfuso, Santis P. De, M. Savino, and A. Scipioni. Dual role of dna intrinsic curvature and flexibility in determining nucleosome stability. *J.Mol.Biol.*, 286(5):1293–1301, March 1999. 35

[ADvH89] J. Ausio, F. Dong, and K.E. van Holde. Use of selectively trypsinized nucleosome core particles to analyze the role of the histone "tails" in the stabilization of the nucleosome. *J.Mol.Biol.*, 206(3):451–463, April 1989. 31

[AH02a] K. Ahmad and S. Henikoff. Histone h3 variants specify modes of chromatin assembly. *Proc.Natl.Acad.Sci.U.S.A*, 99 Suppl 4:16477–16484, December 2002. 36

[AH02b] K. Ahmad and S. Henikoff. The histone variant h3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol.Cell*, 9(6):1191–1200, June 2002. 36

[AKHG+96] S. Aebi, B. Kurdi-Haidar, R. Gordon, B. Cenni, H. Zheng, D. Fink, R. D. Christen, C. R. Boland, M. Koi, R. Fishel, and S. B. Howell. Loss of dna mismatch repair in acquired resistance to cisplatin. *Cancer Res*, 56(13):3087–3090, Jul 1996. 96

[ALB93] A. Amadei, A.B. Linssen, and H.J. Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–425, December 1993. 16, 23

[BAT91] G. J. Bubley, B. P. Ashburner, and B. A. Teicher. Spectrum of cis-diamminedichloroplatinum(ii)-induced mutations in a shuttle vector propagated in human cells. *Mol Carcinog*, 4(5):397–406, 1991. 98, 125

[Bau87] A. Baumgartner. *Applications of the Monte-Carlo simulations in Statistical Physics.* Springer, NY, 1987. 57

[BC82] K.S. Bloom and J. Carbon. Yeast centromere dna is in a unique and highly ordered structure in chromosomes and small circular minichromosomes. *Cell*, 29(2):305–317, June 1982. 47

[BDB+04] J.M. Borreguero, F. Ding, S.V. Buldyrev, H.E. Stanley, and N.V. Dokholyan. Multiple folding pathways of the sh3 domain. *Biophys.J.*, 87(1):521–533, July 2004. 33

[BDF87] D. Burnouf, M. Duane, and R. P. Fuchs. Spectrum of cisplatin-induced mutations in escherichia coli. *Proc Natl Acad Sci U S A*, 84(11):3758–3762, Jun 1987. 98, 125

[BE99] S. Busby and R.H. Ebright. Transcription activation by catabolite activator protein (cap). *J.Mol.Biol.*, 293(2):199–213, October 1999. 35

[Ber02] S.L. Berger. Histone modifications in transcriptional regulation. *Curr.Opin.Genet.Dev.*, 12(2):142–148, April 2002. 7, 31

[BHG+97] R. Brown, G. L. Hirst, W. M. Gallagher, A. J. McIlwrath, G. P. Margison, A. G. van der Zee, and D. A. Anthoney. hmlh1 expression and cellular responses of ovarian tumour cells to treatment with cytotoxic anticancer agents. *Oncogene*, 15(1):45–52, Jul 1997. 96

[Bis05] T.C. Bishop. Molecular dynamics simulations of a nucleosome and free dna. *J.Biomol.Struct.Dyn.*, 22(6):673–686, June 2005. 36

[BK99] Y. Blat and N. Kleckner. Cohesins bind to preferential sites along yeast chromosome iii, with differential regulation along arms versus the centric region. *Cell*, 98(2):249–259, July 1999. 46

[BKB+04] Ekaterina Bassett, Nicole M King, Miriam F Bryant, Suzanne Hector, Lakshmi Pendyala, Stephen G Chaney, and Marila Cordeiro-Stone. The role of dna polymerase eta in translesion synthesis past platinum-dna adducts in human fibroblasts. *Cancer Res*, 64(18):6469–6475, Sep 2004. 95

[BS01] D.A. Beard and T. Schlick. Computational modeling predicts the structure and dynamics of chromatin fiber. *Structure.*, 9(2):105–114, February 2001. 33, 36

[BSD06] K. Bloom, S. Sharma, and N.V. Dokholyan. The path of dna in the kinetochore. *Curr.Biol.*, 16(8):R276–R278, April 2006. 35

[BSK02] M.D. Blower, B.A. Sullivan, and G.H. Karpen. Conserved organization of centromeric chromatin in flies and humans. *Dev.Cell*, 2(3):319–330, March 2002. 52

[BVH+03] Ekaterina Bassett, Alexandra Vaisman, Jody M Havener, Chikahide Masutani, Fumio Hanaoka, and Stephen G Chaney. Efficiency of extension of mismatched primer termini across from cisplatin and oxaliplatin adducts by human dna polymerases beta and eta in vitro. *Biochemistry*, 42(48):14197–14206, Dec 2003. 98, 125, 133

[BYLE93] L. J. Bradley, K. J. Yarema, S. J. Lippard, and J. M. Essigmann. Mutagenicity and genotoxicity of the major dna adduct of the antitumor drug cis-diamminedichloroplatinum(ii). *Biochemistry*, 32(3):982–988, Jan 1993. 98, 125

[BZ02] T.C. Bishop and O.O. Zhmudsky. Mechanical model of the nucleosome and chromatin. *J.Biomol.Struct.Dyn.*, 19(5):877–887, April 2002. 36

[CASC00] P. Cheung, C.D. Allis, and P. Sassone-Corsi. Signaling to chromatin through histone modifications. *Cell*, 103(2):263–271, October 2000. 31

[CC06] S. Cao and S.J. Chen. Predicting rna pseudoknot folding thermodynamics. *Nucleic Acids Res.*, 34(9):2634–2652, 2006. 57

[CCD+05] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, December 2005. 9

[CDN+08] Yiwen Chen, Feng Ding, Huifen Nie, Adrian W Serohijos, Shantanu Sharma, Kyle C Wilcox, Shuangye Yin, and Nikolay V Dokholyan. Protein folding: then and now. *Arch Biochem Biophys*, 469(1):4–19, Jan 2008. 58

[CFV+98] F. Coin, P. Frit, B. Viollet, B. Salles, and J. M. Egly. Tata binding protein discriminates between different lesions on dna, resulting in a transcription decrease. *Mol Cell Biol*, 18(7):3907–3914, Jul 1998. 96

[CIS+04] D.J. Cousins, S.A. Islam, M.R. Sanderson, Y.G. Proykova, C. Crane-Robinson, and D.Z. Staynov. Redefinition of the cleavage sites of dnase i on the nucleosome core particle. *J.Mol.Biol.*, 335(5):1199–1211, January 2004. 34

[CK00] III Cheatham, T.E. and P.A. Kollman. Molecular dynamics simulation of nucleic acids. *Annu.Rev.Phys.Chem.*, 51:435–471, 2000. 8

[CMHL00] S. M. Cohen, Y. Mikata, Q. He, and S. J. Lippard. Hmg-domain protein recognition of cisplatin 1,2-intrastrand d(gpg) cross-links in purine-rich sequence contexts. *Biochemistry*, 39(38):11771–11776, Sep 2000. 129

[CW04] T.E. Cloutier and J. Widom. Spontaneous sharp bending of double-stranded dna. *Mol.Cell*, 14(3):355–362, May 2004. 35

[CY00] III Cheatham, T.E. and M.A. Young. Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers*, 56(4):232–256, 2000. 8

[CYK92] D.C. Chen, B.C. Yang, and T.T. Kuo. One-step transformation of yeast in stationary phase. *Curr.Genet.*, 21(1):83–84, January 1992. 43

[DB07] R. Das and D. Baker. Automated de novo prediction of native-like rna tertiary structures. *Proc.Natl.Acad.Sci.U.S.A*, pages –, August 2007. 57, 78, 91

[DBB+03a] F. Ding, J.M. Borreguero, S.V. Buldyrey, H.E. Stanley, and N.V. Dokholyan. Mechanism for the alpha-helix to beta-hairpin transition. *Proteins*, 53(2):220–228, November 2003. 10, 60

[DBB+03b] N.V. Dokholyan, J.M. Borreguero, S.V. Buldyrev, F. Ding, H.E. Stanley, and E.I. Shakhnovich. Identifying importance of amino acids for protein folding from crystal structures. *Methods Enzymol.*, 374:616–638, 2003. 8

[DBD05] F. Ding, S.V. Buldyrev, and N.V. Dokholyan. Folding trp-cage to nmr resolution native structure using a coarse-grained protein model. *Biophys.J.*, 88(1):147–155, January 2005. 12, 33, 78

[DBSS98] N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley, and E.I. Shakhnovich. Discrete molecular dynamics studies of the folding of a protein-like model. *Fold.Des*, 3(6):577–587, 1998. 8, 58, 64, 80, 81, 91

[DC97] Jr. Dunbrack, R.L. and F.E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–1681, August 1997. 17

[DCD+04] R.D. Dixon, Y. Chen, F. Ding, S.D. Khare, K.C. Prutzman, M.D. Schaller, S.L. Campbell, and N.V. Dokholyan. New insights into fak signaling and localization based on detection of a fat domain folding intermediate. *Structure.*, 12(12):2161–2171, December 2004. 71

[DD01] E.A. Doherty and J.A. Doudna. Ribozyme structures and mechanisms. *Annu.Rev Biophys.Biomol.Struct.*, 30:457–475, 2001. 56

[DD05] F. Ding and N.V. Dokholyan. Simple but predictive protein models. *Trends Biotechnol.*, 23(9):450–455, September 2005. 9, 15, 33, 57, 58, 80

[DD06] F. Ding and N.V. Dokholyan. Emergence of protein fold families through rational design. *PLoS.Comput.Biol.*, 2(7):e85–, July 2006. 39, 40, 41, 91

[DDB+02a] F. Ding, N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley, and E.I. Shakhnovich. Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys.J.*, 83(6):3525–3532, December 2002. 10, 32

[DDB+02b] Feng Ding, Nikolay V Dokholyan, Sergey V Buldyrev, H. Eugene Stanley, and Eugene I Shakhnovich. Molecular dynamics simulation of the sh3 domain aggregation suggests a generic amyloidogenesis mechanism. *J Mol Biol*, 324(4):851–857, Dec 2002. 32

[Dic98] R.E. Dickerson. Dna bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, 26(8):1906–1926, April 1998. 35

[DJD05] F. Ding, R.K. Jha, and N.V. Dokholyan. Scaling behavior and structure of denatured proteins. *Structure.*, 13(7):1047–1054, July 2005. 33, 80

[DL97] S. U. Dunham and S. J. Lippard. Dna sequence context and protein composition modulate hmg-domain protein recognition of cisplatin-modified dna. *Biochemistry*, 36(38):11428–11436, Sep 1997. 129

[DLD05] F. Ding, J.J. LaRocque, and N.V. Dokholyan. Direct observation of protein folding, aggregation, and a prion-like conformational conversion. *J.Biol.Chem.*, 280(48):40235–40240, December 2005. 8

[Dok06] N.V. Dokholyan. Studies of folding and misfolding using simplified models. *Curr.Opin.Struct.Biol.*, pages 79–85, January 2006. 8

[DPCD06] F. Ding, K.C. Prutzman, S.L. Campbell, and N.V. Dokholyan. Topological determinants of protein domain swapping. *Structure.*, 14(1):5–14, January 2006. 17

[DSC+08] F. Ding, S. Sharma, P. Chalasani, V.V. Demidov, N.E. Broude, and N.V. Dokholyan. Ab initio rna folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA.*, pages 1164–1173, May 2008. 91, 92

[DSL+02] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J.Mol.Biol.*, 319(5):1097–1113, June 2002. 6, 7, 8, 12, 31, 38

[Edd01] S.R. Eddy. Non-coding rna genes and the modern rna world. *Nat.Rev Genet.*, 2(12):919–929, December 2001. 56

[Edd04] S.R. Eddy. How do rna folding algorithms work? *Nat.Biotechnol.*, 22(11):1457–1458, November 2004. 57

[ERK01] M. A. Elizondo-Riojas and J. Kozelka. Unrestrained 5 ns molecular dynamics simulation of a cisplatin-dna 1,2-gg adduct provides a rationale for the nmr features and reveals increased conformational flexibility at the platinum binding site. *J Mol Biol*, 314(5):1227–1243, Dec 2001. 109, 124, 125, 126, 127, 133

[FKB04] M. Feig, J. Karanicolas, and III Brooks, C.L. Mmtsb tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J.Mol.Graph.Model.*, 22(5):377–395, May 2004. 65

[FNA+96] D. Fink, S. Nebel, S. Aebi, H. Zheng, B. Cenni, A. Nehm, R. D. Christen, and S. B. Howell. The role of dna mismatch repair in platinum drug resistance. *Cancer Res*, 56(21):4881–4886, Nov 1996. 96

[FRLT04] J.Y. Fan, D. Rangasamy, K. Luger, and D.J. Tremethick. H2a.z alters the nucleosome surface to promote hp1alpha-mediated chromatin fiber folding. *Mol.Cell*, 16(4):655–661, November 2004. 36

[FTD02] H. Fukunaga, J. Takimoto, and M. Doi. A coarse-graining procedure for flexible polymer chains with bonded and nonbonded interactions. *Journal of Chemical Physics*, 116(18):8183–8190, May 2002. 36

[FWA03] W. Fischle, Y. Wang, and C.D. Allis. Histone and chromatin cross-talk. *Curr.Opin.Cell Biol.*, 15(2):172–183, April 2003. 6

[FZN+97] D. Fink, H. Zheng, S. Nebel, P. S. Norris, S. Aebi, T. P. Lin, A. Nehm, R. D. Christen, M. Haas, C. L. MacLeod, and S. B. Howell. In vitro and in vivo resistance to cisplatin in cells that have lost dna mismatch repair. *Cancer Res*, 57(10):1841–1845, May 1997. 96

[GD94] T.C. Gluick and D.E. Draper. Thermodynamics of folding a pseudoknotted mrna fragment. *J.Mol.Biol.*, 241(2):246–262, August 1994. 68, 73, 74

[GL98] A. Gelasco and S. J. Lippard. Nmr solution structure of a dna dodecamer duplex containing a cis-diammineplatinum(ii) d(gpg) intrastrand cross-link, the major adduct of the anticancer drug cisplatin. *Biochemistry*, 37(26):9230–9239, Jun 1998. 97, 116, 126

[GMAM05] Ekaterina L Grishchuk, Maxim I Molodtsov, Fazly I Ataullakhanov, and J. Richard McIntosh. Force production by disassembling microtubules. *Nature*, 438(7066):384–388, Nov 2005. 52

[GP02] L.M. Gloss and B.J. Placek. The effect of salts on the stability of the h2a-h2b histone dimer. *Biochemistry*, 41(50):14951–14959, December 2002. 27

[GPS+05] M.K. Gardner, C.G. Pearson, B.L. Sprague, T.R. Zarzar, K. Bloom, E.D. Salmon, and D.J. Odde. Tension-dependent regulation of microtubule dynamics at kinetochores can explain metaphase congression in yeast. *Mol.Biol.Cell*, 16(8):3764–3775, August 2005. 54

[Gre92] M. H. Greene. Is cisplatin a human carcinogen? *J Natl Cancer Inst*, 84(5):306–312, Mar 1992. 95

[Gru97] M. Grunstein. Histone acetylation in chromatin structure and transcription. *Nature*, 389(6649):349–352, September 1997. 8

[HAM01] S. Henikoff, K. Ahmad, and H.S. Malik. The centromere paradox: stable inheritance with rapidly evolving dna. *Science*, 293(5532):1098–1102, August 2001. 54

[HB01] B. Hendrich and W. Bickmore. Human diseases with underlying defects in chromatin structure and modification. *Hum.Mol.Genet.*, 10(20):2233–2242, October 2001. 38

[HHTB00] J.M. Harp, B.L. Hanson, D.E. Timm, and G.J. Bunick. Asymmetries in the nucleosome core particle at 2.5 a resolution. *Acta Crystallogr.D.Biol.Crystallogr.*, 56 Pt 12:1513–1534, December 2000. 6, 7

[HLVL97] J. S. Hoffmann, D. Locker, G. Villani, and M. Leng. Hmg1 protein inhibits the translesion synthesis of the major dna cisplatin adduct by cell extracts. *J Mol Biol*, 270(4):539–543, Jul 1997. 96

[HMAM01] M. Howe, K.L. McDonald, D.G. Albertson, and B.J. Meyer. Him-10 is required for kinetochore structure and function on caenorhabditis elegans holocentric chromosomes. *J.Cell Biol.*, 153(6):1227–1238, June 2001. 52

[HMK05] C.E. Huang, M. Milutinovich, and D. Koshland. Rings, bracelet or snaps: fashionable alternatives for smc complexes. *Philos.Trans.R Soc.Lond B Biol.Sci.*, 360(1455):537–542, March 2005. 48

[Hof03] I.L. Hofacker. Vienna rna secondary structure server. *Nucleic Acids Res.*, 31(13):3429–3431, July 2003. 57

[HPAT88] William H.Press, Brian P.Flannery, Saul A.Teukolsky, and William T.Vetterling. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, New York, NY, USA, 1988. 16, 103, 104, 116, 118

[HS06] A. Huttenhofer and P. Schattner. The principles of guiding by rna: chimeric rna-protein enzymes. *Nat.Rev.Genet.*, 7(6):475–482, June 2006. 56

[HTSR01] J. A. Holbrook, O. V. Tsodikov, R. M. Saecker, and M. T. Record. Specific and non-specific interactions of integration host factor with dna: thermodynamic evidence for disruption of multiple ihf surface salt-bridges coupled to dna binding. *J Mol Biol*, 310(2):379–401, Jul 2001. 52

[HZR+94] J. C. Huang, D. B. Zamble, J. T. Reardon, S. J. Lippard, and A. Sancar. Hmg-domain proteins specifically inhibit the repair of the major dna adduct of the anticancer drug cisplatin by human excision nuclease. *Proc Natl Acad Sci U S A*, 91(22):10394–10398, Oct 1994. 96

[JA01] T. Jenuwein and C.D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, August 2001. 6, 31, 36

[JEK89] M. M. Jennerwein, A. Eastman, and A. Khokhar. Characterization of adducts produced in dna by isomeric 1,2-diaminocyclohexaneplatinum(ii) complexes. *Chem Biol Interact*, 70(1-2):39–49, 1989. 96

[KA99] M.H. Kuo and C.D. Allis. In vivo cross-linking and immunoprecipitation for studying dynamic protein:dna associations in a chromatin environment. *Methods*, 19(3):425–433, November 1999. 34

[KBB+94] K.M. Kramer, J.A. Brock, K. Bloom, J.K. Moore, and J.E. Haber. Two different types of double-strand breaks in saccharomyces cerevisiae are repaired by similar rad52-independent, nonhomologous recombination events. *Mol.Cell Biol.*, 14(2):1293–1301, February 1994. 54

[KBO00] V. Katritch, C. Bustamante, and W.K. Olson. Pulling chromatin fibers: computer simulations of direct physical micromanipulations. *J.Mol.Biol.*, 295(1):29–40, January 2000. 36

[KBS+92] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman, and J.M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992. 65

[KDD03] S.D. Khare, F. Ding, and N.V. Dokholyan. Folding of cu, zn superoxide dismutase and familial amyotrophic lateral sclerosis. *J.Mol.Biol.*, 334(3):515–525, November 2003. 33

[KFH00] K. C. Keith and M. Fitzgerald-Hayes. Cse4 genetically interacts with the saccharomyces cerevisiae centromere dna elements cde i and cde ii but not cde iii. implications for the path of the centromere dna around a cse4p variant nucleosome. *Genetics*, 156(3):973–981, Nov 2000. 47

[KL99] R.D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–294, August 1999. 8

[Kor74] R.D. Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(139):868–871, May 1974. 6

[KP90] M. Karplus and G.A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, October 1990. 8

[KS83] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983. 17

[KSB+93] Jiří Kozelka, Roger Savinelli, Gaston Berthier, Jean-Pierre Flament, and Richard Lavery. Force field for platinum binding to adenine. *J. Comput. Chem.*, 14(1):45–53, 1993. 124

[KvdW+98] M.H. Kolk, Graaf M. van der, S.S. Wijmenga, C.W. Pleij, H.A. Heus, and C.W. Hilbers. Nmr structure of a classical pseudoknot: interplay of single- and double-stranded rna. *Science*, 280(5362):434–438, April 1998. 69

[LD94] L.G. Laing and D.E. Draper. Thermodynamics of rna folding in a conserved ribosomal rna domain. *J.Mol.Biol.*, 237(5):560–576, April 1994. 68, 73, 74

[LGK78] I.M. Lifshitz, A.Y. Grosberg, and A.R. Khokhlov. Some problems of statistical physics of polymer-chains with volume interaction. *Reviews of Modern Physics*, 50(3):683–713, 1978. 31

[LLBW05] G. Li, M. Levitus, C. Bustamante, and J. Widom. Rapid spontaneous accessibility of nucleosomal dna. *Nat.Struct.Mol.Biol.*, 12(1):46–53, January 2005. 33, 35, 36

[LLT01] N.M. Luscombe, R.A. Laskowski, and J.M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-dna interactions at an atomic level. *Nucleic Acids Res.*, 29(13):2860–2874, July 2001. 25

[LMR+97] K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, and T.J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, September 1997. 6, 7, 15, 34

[LPE+00] M. H. Lamers, A. Perrakis, J. H. Enzlin, H. H. Winterwerp, N. de Wind, and T. K. Sixma. The crystal structure of dna mismatch repair protein muts binding to a g x t mismatch. *Nature*, 407(6805):711–717, Oct 2000. 97

[LR98] K. Luger and T.J. Richmond. The histone tails of the nucleosome. *Curr.Opin.Genet.Dev.*, 8(2):140–146, April 1998. 8, 31

[LT99] N. Lee and D. Thirumalai. Stretching dna: Role of electrostatic interactions. *European Physical Journal B*, 12(4):599–605, 1999. 35

[LT02] N.M. Luscombe and J.M. Thornton. Protein-dna interactions: amino acid conservation and the effects of mutations on binding specificity. *J.Mol.Biol.*, 320(5):991–1009, July 2002. 25

[LT04] N. Lee and D. Thirumalai. Pulling-speed-dependent force-extension profiles for semiflexible chains. *Biophys J*, 86(5):2641–2649, May 2004. 35

[LW75] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, February 1975. 8

[LW04] G. Li and J. Widom. Nucleosomes facilitate their own invasion. *Nat.Struct.Mol.Biol.*, 11(8):763–769, August 2004. 35

[LYR+94] S.H. Leuba, G. Yang, C. Robert, B. Samori, Holde K. van, J. Zlatanova, and C. Bustamante. Three-dimensional structure of extended chromatin fibers as revealed by tapping-mode scanning force microscopy. *Proc.Natl.Acad.Sci.U.S.A*, 91(24):11621–11625, November 1994. 8

[MARR03] L.J. Murray, III Arendall, W.B., D.C. Richardson, and J.S. Richardson. Rna backbone is rotameric. *Proc.Natl.Acad.Sci.U.S.A*, 100(24):13904–13909, November 2003. 60

[Mat06] D.H. Mathews. Revolutions in rna secondary structure prediction. *J.Mol.Biol.*, 359(3):526–532, June 2006. 56, 57, 63

[MDSH05] J.J. Miranda, Wulf P. De, P.K. Sorger, and S.C. Harrison. The yeast dash complex forms closed rings on microtubules. *Nat.Struct.Mol.Biol.*, 12(2):138–143, February 2005. 50

[MF80] J.D. McGhee and G. Felsenfeld. Nucleosome structure. *Annu.Rev.Biochem.*, 49:1115–1156, 1980. 39

[MGA+98] V. Mutskov, D. Gerber, D. Angelov, J. Ausio, J. Workman, and S. Dimitrov. Persistent interactions of core histone tails with nucleosomal dna following acetylation and transcription factor binding. *Mol.Cell Biol.*, 18(11):6293–6304, November 1998. 34

[MGC93] F. Major, D. Gautheret, and R. Cedergren. Reproducing the three-dimensional structure of a trna molecule from structural constraints. *Proc.Natl.Acad.Sci.U.S.A*, 90(20):9408–9412, October 1993. 57

[MH03] H.S. Malik and S. Henikoff. Phylogenomics of the nucleosome. *Nat.Struct.Biol.*, 10(11):882–891, November 2003. 36

[MKIH00] C. Masutani, R. Kusumoto, S. Iwai, and F. Hanaoka. Mechanisms of accurate translesion synthesis by human dna polymerase eta. *EMBO J*, 19(12):3100–3109, Jun 2000. 98, 125, 133

[ML96] M. M. McA'Nulty and S. J. Lippard. The hmg-domain protein ixr1 blocks excision repair of cisplatin-dna adducts in yeast. *Mutat Res*, 362(1):75–86, Jan 1996. 96

[MLDL03] S. Mangenot, A. Leforestier, D. Durand, and F. Livolant. Phase diagram of nucleosome core particles. *J.Mol.Biol.*, 333(5):907–916, November 2003. 27, 29

[MLV+02] S. Mangenot, A. Leforestier, P. Vachette, D. Durand, and F. Livolant. Salt-induced conformation and interaction changes of nucleosome core particles. *Biophys.J.*, 82(1 Pt 1):345–356, January 2002. 27, 29, 32

[MP02] F. Muller-Plathe. Coarse-graining in polymer simulation: from the atomistic to the mesoscopic scale and back. *Chemphyschem.*, 3(9):755–769, September 2002. 19

[MS95] J.F. Marko and E.D. Siggia. Stretching dna. *Macromolecules*, 28(26):8759–8770, 1995. 35

[MS02] E.A. Mukamel and E.I. Shakhnovich. Phase diagram for unzipping dna with long-range interactions. *Phys.Rev.E.Stat.Nonlin.Soft.Matter Phys.*, 66(3 Pt 1):032901–, September 2002. 35

[MSK+01] L. G. Marzilli, J. S. Saad, Z. Kuklenyik, K. A. Keating, and Y. Xu. Relationship of solution and protein-bound structures of dna duplexes with the major intrastrand cross-link lesions formed on cisplatin binding to dna. *J Am Chem Soc*, 123(12):2764–2770, Mar 2001. 97, 98, 99, 116, 124, 125, 133

[MSZT99] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *J.Mol.Biol.*, 288(5):911–940, May 1999. 62, 63

[MTG+91] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of rna. *Science*, 253(5025):1255–1260, September 1991. 57

[MTS03] A.D. McAinsh, J.D. Tytell, and P.K. Sorger. Structure, function, and regulation of budding yeast kinetochores. *Annu.Rev Cell Dev.Biol.*, 19:519–539, 2003. 46

[MYG+98] P.B. Meluh, P. Yang, L. Glowczewski, D. Koshland, and M.M. Smith. Cse4p is a component of the core centromere of saccharomyces cerevisiae. *Cell*, 94(5):607–613, September 1998. 48

[Nic83] R. B. Nicklas. Measurements of the force produced by the mitotic spindle in anaphase. *J Cell Biol*, 97(2):542–548, Aug 1983. 52

[NLSK04] S.O. Nielsen, C.F. Lopez, G. Srinivas, and M.L. Klein. Coarse grain models and the computer simulation of soft materials. *Journal of Physics-Condensed Matter*, 16(15):R481–R512, 2004. 19, 33

[NSD+08] Shima Nakanishi, Brian W Sanderson, Kym M Delventhal, William D Bradford, Karen Staehling-Hampton, and Ali Shilatifard. A comprehensive library of histone mutants identifies nucleosomal residues required for h3k4 methylation. *Nat Struct Mol Biol*, 15(8):881–888, Aug 2008. 44

[OBHY00] G. Obmolova, C. Ban, P. Hsieh, and W. Yang. Crystal structures of mismatch repair protein muts and its complex with a substrate dna. *Nature*, 407(6805):703–710, Oct 2000. 97

[OGL+98] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin. Dna sequence-dependent deformability deduced from protein-dna crystal complexes. *Proc.Natl.Acad.Sci.U.S.A*, 95(19):11163–11168, September 1998. 35

[ORH+99] U. M. Ohndorf, M. A. Rould, Q. He, C. O. Pabo, and S. J. Lippard. Basis for recognition of cisplatin-modified dna by high-mobility-group proteins. *Nature*, 399(6737):708–712, Jun 1999. 97, 128

[OSR+96] V.V. Ogryzko, R.L. Schiltz, V. Russanova, B.H. Howard, and Y. Nakatani. The transcriptional coactivators p300 and cbp are histone acetyltransferases. *Cell*, 87(5):953–959, November 1996. 31

[OW87] I. Oohara and A. Wada. Spectroscopic studies on histone-dna interactions. ii. three transitions in nucleosomes resolved by salt-titration. *J.Mol.Biol.*, 196(2):399–411, July 1987. 27

[PA05] A.L. Pidoux and R.C. Allshire. The role of heterochromatin in centromere function. *Philos.Trans.R Soc.Lond B Biol.Sci.*, 360(1455):569–579, March 2005. 48

[PC03] J.W. Ponder and D.A. Case. Force fields for protein simulations. *Adv.Protein Chem.*, 66:27–85, 2003. 9

[PCC+95] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, September 1995. 8, 9

[PDU+04] S. Peng, F. Ding, B. Urbanc, S.V. Buldyrev, L. Cruz, H.E. Stanley, and N.V. Dokholyan. Discrete molecular dynamics simulations of peptide aggregation. *Phys.Rev.E.Stat.Nonlin.Soft.Matter Phys.*, 69(4 Pt 1):041908–, April 2004. 10, 12, 15, 33

[PHSC90] J. D. Page, I. Husain, A. Sancar, and S. G. Chaney. Effect of the diaminocyclohexane carrier ligand on platinum adduct formation, repair, and lethality. *Biochemistry*, 29(4):1016–1024, Jan 1990. 96

[PM08] M. Parisien and F. Major. The mc-fold and mc-sym pipeline infers rna structure from sequence data. *Nature*, 452(7183):51–55, March 2008. 91

[PMd97] J. Perez-Martin and V de, Lorenzo. Clues and consequences of dna bending in transcription. *Annu.Rev.Microbiol.*, 51:593–628, 1997. 35

[PMSB01] C.G. Pearson, P.S. Maddox, E.D. Salmon, and K. Bloom. Budding yeast chromosome structure and dynamics during mitosis. *J.Cell Biol.*, 152(6):1255–1266, March 2001. 46

[PTH+99] L.I. Pietrasanta, D. Thrower, W. Hsieh, S. Rao, O. Stemmann, J. Lechner, J. Carbon, and H. Hansma. Probing the saccharomyces cerevisiae centromeric dna (cen dna)-binding factor 3 (cbf3) kinetochore complex by using atomic force microscopy. *Proc.Natl.Acad.Sci.U.S.A*, 96(7):3757–3762, March 1999. 48, 50

[RD03] T.J. Richmond and C.A. Davey. The structure of dna in the nucleosome core. *Nature*, 423(6936):145–150, May 2003. 15

[RE99] E. Rivas and S.R. Eddy. A dynamic programming algorithm for rna structure prediction including pseudoknots. *J.Mol.Biol.*, 285(5):2053–2068, February 1999. 63

[RSKC05] J. Roach, S. Sharma, M. Kapustina, and Jr. Carter, C.W. Structure alignment via delaunay tetrahedralization. *Proteins*, 60(1):66–81, July 2005. 8

[SA00] B.D. Strahl and C.D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, January 2000. 6, 7, 36, 39

[SAMS97] M.S. Santisteban, G. Arents, E.N. Moudrianakis, and M.M. Smith. Histone octamer function in vivo: mutations in the dimer-tetramer interfaces disrupt both gene activation and repression. *EMBO J.*, 16(9):2493–2506, May 1997. 33, 34

[SB05] D.W. Staple and S.E. Butcher. Pseudoknots: Rna structures with diverse functions. *PLoS.Biol.*, 3(6):e213–, June 2005. 69

[SBH99] E. D. Scheeff, J. M. Briggs, and S. B. Howell. Molecular modeling of the intrastrand guanine-guanine dna adducts produced by cisplatin and oxaliplatin. *Mol Pharmacol*, 56(3):633–643, Sep 1999. 100, 111, 124

[SCB96] S.B. Smith, Y. Cui, and C. Bustamante. Overstretching b-dna: the elastic response of individual double-stranded and single-stranded dna molecules. *Science*, 271(5250):795–799, February 1996. 35

[SCD+05] Maria J Silva, Paula Costa, Anabela Dias, Marco Valente, Henriqueta Louro, and Maria G Boavida. Comparative analysis of the mutagenic activity of oxaliplatin and cisplatin in the hprt gene of cho cells. *Environ Mol Mutagen*, 46(2):104–115, Aug 2005. 95

[Sch03] H. Schiessel. The physics of chromatin. *Journal of Physics-Condensed Matter*, 15(19):R699–R774, May 2003. 9

[Sch06] H. Schiessel. The nucleosome: A transparent, slippery, sticky and yet stable dna-protein complex. *Eur.Phys.J.E.Soft.Matter*, pages –, February 2006. 9

[SDD07] S. Sharma, F. Ding, and N.V. Dokholyan. Multiscale modeling of nucleosome dynamics. *Biophys.J.*, 92(5):1457–1470, March 2007. 39, 40, 58

[SDN+06] S. Sharma, F. Ding, H.F. Nie, D. Watson, A. Unnithan, J. Lopp, D. Pozefsky, and N.V. Dokholyan. ifold: a platform for interactive folding simulations of proteins. *Bioinformatics*, 22(21):2693–2694, 2006. 15, 58, 78, 92

[SFB92] S.B. Smith, L. Finzi, and C. Bustamante. Direct mechanical measurements of the elasticity of single dna molecules by using magnetic beads. *Science*, 258(5085):1122–1126, November 1992. 35

[SFMC+06] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, pages –, July 2006. 27

[SLV88] M.J. Solomon, P.L. Larsen, and A. Varshavsky. Mapping protein-dna interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, June 1988. 34

[SNR+04] E.J. Sorin, B.J. Nakatani, Y.M. Rhee, G. Jayachandran, V. Vishal, and V.S. Pande. Does native state topology determine the rna folding mechanism? *J.Mol.Biol.*, 337(4):789–797, April 2004. 63, 73

[SO91] A.W. Shermoen and P.H. O'Farrell. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell*, 67(2):303–310, October 1991. 36

[SPW+97] M. R. Sawaya, R. Prasad, S. H. Wilson, J. Kraut, and H. Pelletier. Crystal structures of human dna polymerase beta complexed with gapped and nicked dna: evidence for an induced fit mechanism. *Biochemistry*, 36(37):11205–11215, Sep 1997. 97

[SPY+95] A. Sali, L. Potterton, F. Yuan, Vlijmen H. van, and M. Karplus. Evaluation of comparative protein modeling by modeller. *Proteins*, 23(3):318–326, November 1995. 47

[SSK94] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369(6477):248–251, May 1994. 65, 66

[SWL01] B. Spingler, D. A. Whittington, and S. J. Lippard. 2.4 a crystal structure of an oxaliplatin 1,2-d(gpg) intrastrand cross-link in a dna dodecamer duplex. *Inorg Chem*, 40(22):5596–5602, Oct 2001. 97, 106, 115, 116, 124, 126, 127, 132, 133

[SYKB07] B.A. Shapiro, Y.G. Yingling, W. Kasprzak, and E. Bindewald. Bridging the gap in rna structure prediction. *Curr.Opin.Struct.Biol.*, 17(2):157–165, April 2007. 56, 57, 91

[SZS05] J. Sun, Q. Zhang, and T. Schlick. Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *Proc.Natl.Acad.Sci.U.S.A*, 102(23):8180–8185, June 2005. 36

[SZW+99] S.K. Silverman, M. Zheng, M. Wu, Jr. Tinoco, I., and T.R. Cech. Quantifying the energetic interplay of rna tertiary and secondary structure interactions. *RNA.*, 5(12):1665–1674, December 1999. 76

[TB99] Jr. Tinoco, I. and C. Bustamante. How rna folds. *J.Mol.Biol.*, 293(2):271–281, October 1999. 63, 76

[TCS+97] L. B. Travis, R. E. Curtis, H. Storm, P. Hall, E. Holowaty, F. E. Van Leeuwen, B. A. Kohler, E. Pukkala, C. F. Lynch, M. Andersson, K. Bergfeldt, E. A. Clarke, T. Wiklund, G. Stoter, M. Gospodarowicz, J. Sturgeon, J. F. Fraumeni, and J. D. Boice. Risk of second malignant neoplasms among long-term survivors of testicular cancer. *J Natl Cancer Inst*, 89(19):1429–1439, Oct 1997. 95

[TFL96] Patricia M. Takahara, Christin A. Frederick, and Stephen J. Lippard. Crystal structure of the anticancer drug cisplatin bound to duplex dna. *Journal of the American Chemical Society*, 118(49):12309–12321, 1996. 97, 106, 115, 116, 124, 127, 132, 133

[TLW04] A. Thastrom, P.T. Lowary, and J. Widom. Measurement of histone-dna interaction free energy in nucleosomes. *Methods*, 33(1):33–44, May 2004. 52

[TZJE94] D. K. Treiber, X. Zhai, H. M. Jantzen, and J. M. Essigmann. Cisplatin-dna adducts are molecular decoys for the ribosomal rna transcription factor hubf (human upstream binding factor). *Proc Natl Acad Sci U S A*, 91(12):5672–5676, Jun 1994. 96

[UBGB94] S.I. Usachenko, S.G. Bavykin, I.M. Gavin, and E.M. Bradbury. Rearrangement of the histone h2a c-terminal domain in the nucleosome. *Proc.Natl.Acad.Sci.U.S.A*, 91(15):6845–6849, July 1994. 34

[UCD+04] B. Urbanc, L. Cruz, F. Ding, D. Sammond, S. Khare, S.V. Buldyrev, H.E. Stanley, and N.V. Dokholyan. Molecular dynamics simulation of amyloid beta dimer formation. *Biophys.J.*, 87(4):2310–2321, October 2004. 33

[VC00] A. Vaisman and S. G. Chaney. The efficiency and fidelity of translesion synthesis past cisplatin and oxaliplatin gpg adducts by human dna polymerase beta. *J Biol Chem*, 275(17):13017–13025, Apr 2000. 96

[VLP+99] A. Vaisman, S. E. Lim, S. M. Patrick, W. C. Copeland, D. C. Hinkle, J. J. Turchi, and S. G. Chaney. Effect of dna polymerases and high mobility group protein 1 on the carrier ligand specificity for translesion synthesis past platinum-dna adducts. *Biochemistry*, 38(34):11026–11039, Aug 1999. 96

[VMHC00] A. Vaisman, C. Masutani, F. Hanaoka, and S. G. Chaney. Efficient translesion replication past oxaliplatin and cisplatin gpg adducts by human dna polymerase eta. *Biochemistry*, 39(16):4575–4580, Apr 2000. 96

[VVU+98] A. Vaisman, M. Varchenko, A. Umar, T. A. Kunkel, J. I. Risinger, J. C. Barrett, T. C. Hamilton, and S. G. Chaney. The role of hmlh1, hmsh3, and hmsh6 defects in cisplatin and oxaliplatin resistance: correlation with replicative bypass of platinum-dna adducts. *Cancer Res*, 58(16):3579–3585, Aug 1998. 96

[vZ96] Holde K. van and J. Zlatanova. What determines the folding of the chromatin fiber? *Proc.Natl.Acad.Sci.U.S.A*, 93(20):10548–10555, October 1996. 8, 36

[WBK+07] Yibing Wu, Debadeep Bhattacharyya, Candice L King, Irene Baskerville-Abraham, Sung-Ho Huh, Gunnar Boysen, James A Swenberg, Brenda Temple, Sharon L Campbell, and Stephen G Chaney. Solution structures of a dna dodecamer duplex with and without a cisplatin 1,2-d(gg) intrastrand cross-link: comparison with the same dna duplex containing an oxaliplatin 1,2-d(gg) intrastrand cross-link. *Biochemistry*, 46(22):6477–6487, Jun 2007. 98, 99, 102, 106, 109, 111, 115, 124, 125, 126, 132, 133

[WCN+98] J. M. Woynarowski, W. G. Chapman, C. Napier, M. C. Herzig, and P. Juniewicz. Sequence- and region-specificity of oxaliplatin adducts in naked and cellular dna. *Mol Pharmacol*, 54(5):770–777, Nov 1998. 96

[WCSL01] M. Wei, S. M. Cohen, A. P. Silverman, and S. J. Lippard. Effects of spectator ligands on the specific recognition of intrastrand platinum-dna cross-links by high mobility group box and tata-binding proteins. *J Biol Chem*, 276(42):38774–38780, Oct 2001. 96, 129

[WHMA01] X. Wang, C. He, S.C. Moore, and J. Ausio. Effects of histone acetylation on the solubility and folding of the chromatin fiber. *J.Biol.Chem.*, 276(16):12764–12768, April 2001. 36

[Wid01] J. Widom. Role of dna sequence in nucleosome stability and dynamics. *Q.Rev.Biophys.*, 34(3):269–324, August 2001. 18

[WJPP01] M. T. Washington, R. E. Johnson, S. Prakash, and L. Prakash. Mismatch extension ability of yeast and human dna polymerase eta. *J Biol Chem*, 276(3):2263–2266, Jan 2001. 98, 125, 133

[WK98] J.L. Workman and R.E. Kingston. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu.Rev.Biochem.*, 67:545–579, 1998. 27

[WM00] L.D. Williams and III Maher, L.J. Electrostatic mechanisms of dna deformation. *Annu.Rev.Biophys.Biomol.Struct.*, 29:497–521, 2000. 35

[WMLA00] X. Wang, S.C. Moore, M. Laszckzak, and J. Ausio. Acetylation increases the alpha-helical content of the histone tails of the nucleosome. *J.Biol.Chem.*, 275(45):35013–35020, November 2000. 35

[WMW05a] K.A. Wilkinson, E.J. Merino, and K.M. Weeks. Rna shape chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in trna(asp) transcripts. *J.Am.Chem.Soc.*, 127(13):4659–4667, April 2005. 77

[WMW05b] K.A. Wilkinson, E.J. Merino, and K.M. Weeks. Rna shape chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in trna(asp) transcripts. *J Am.Chem Soc.*, 127(13):4659–4667, April 2005. 94

[WMW06] K.A. Wilkinson, E.J. Merino, and K.M. Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (shape): quantitative rna structure analysis at single nucleotide resolution. *Nat.Protoc.*, 1(3):1610–1616, 2006. 76

[WNL+05] C.M. Wood, J.M. Nicholson, S.J. Lambert, C.D. Chantalat, L.and Reynolds, and J.P. Baldwin. High-resolution structure of the native histone octamer. *Acta Crystallographica Section F*, 61:541–545, 2005. 34

[WPH+04] Yibing Wu, Padmanava Pradhan, Jody Havener, Gunnar Boysen, James A Swenberg, Sharon L Campbell, and Stephen G Chaney. Nmr solution structure of an oxaliplatin 1,2-d(gg) intrastrand cross-link in a dna dodecamer duplex. *J Mol Biol*, 341(5):1251–1269, Aug 2004. 97, 98, 102, 106, 109, 111, 115, 124, 125, 126, 132, 133

[WSd04] R.C. Wang, A. Smogorzewska, and Lange T. de. Homologous recombination generates t-loop-sized deletions at human telomeres. *Cell*, 119(3):355–368, October 2004. 54

[WSH05] R.R. Wei, P.K. Sorger, and S.C. Harrison. Molecular organization of the ndc80 complex, an essential kinetochore component. *Proc.Natl.Acad.Sci.U.S.A*, 102(15):5363–5367, April 2005. 50

[WSL01] C.L. White, R.K. Suto, and K. Luger. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *EMBO J.*, 20(18):5207–5218, September 2001. 6, 38, 47

[WvSW+05] S. Westermann, A. vila Sakar, H.W. Wang, H. Niederstrasser, J. Wong, D.G. Drubin, E. Nogales, and G. Barnes. Formation of a dynamic kinetochore-microtubule interface through assembly of the dam1 ring complex. *Mol.Cell*, 17(2):277–290, January 2005. 50

[YDD07] S. Yin, F. Ding, and N.V. Dokholyan. Eris: an automated estimator of protein stability. *Nat.Methods*, 4(6):466–467, June 2007. 39, 41, 43

[YMvH89] T.D. Yager, C.T. McMurray, and K.E. van Holde. Salt-induced release of dna from nucleosome core particles. *Biochemistry*, 28(5):2271–2281, March 1989. 27

[YPM94] S. Yao, J.P. Plastaras, and L.G. Marzilli. A molecular mechanics amber-type force field for modeling platinum complexes of guanine derivatives. *Inorganic Chemistry*, 33(26):6061–6077, 1994. 100, 111, 123, 124

[YvBR+95] D. Yang, S. S. van Boom, J. Reedijk, J. H. van Boom, and A. H. Wang. Structure and isomerization of an intrastrand cisplatin-cross-linked octamer dna duplex by nmr analysis. *Biochemistry*, 34(39):12912–12920, Oct 1995. 116

[ZBE+98] W. Zhang, J.R. Bone, D.G. Edmondson, B.M. Turner, and S.Y. Roth. Essential and redundant functions of histone acetylation revealed by mutation of target lysines and loss of the gcn5p acetyltransferase. *EMBO J*, 17(11):3155–3167, June 1998. 43

[ZBJE98] X. Zhai, H. Beckmann, H. M. Jantzen, and J. M. Essigmann. Cisplatin-dna adducts inhibit ribosomal rna synthesis by hijacking the transcription factor human upstream binding factor. *Biochemistry*, 37(46):16307–16315, Nov 1998. 96

[ZH03] C. Zheng and J.J. Hayes. Intra- and inter-nucleosomal protein-dna interactions of the core histone tail domains in a model system. *J.Biol.Chem.*, 278(26):24217–24224, June 2003. 32

[ZH04] C. Zheng and J.J. Hayes. Probing core histone tail-dna interactions in a model dinucleosome system. *Methods Enzymol.*, 375:179–193, 2004. 32

[Zhu05] X. Zhuang. Single-molecule rna science. *Annu.Rev.Biophys.Biomol.Struct.*, 34:399–414, 2005. 77

[ZK97] Y. Zhou and M. Karplus. Folding thermodynamics of a model three-helix-bundle protein. *Proc.Natl.Acad.Sci.U.S.A*, 94(26):14429–14432, December 1997. 15

[ZK99] Y. Zhou and M. Karplus. Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. *J.Mol.Biol.*, 293(4):917–951, November 1999. 19, 20

[ZLHH05] C. Zheng, X. Lu, J.C. Hansen, and J.J. Hayes. Salt-dependent intra- and internucleosomal interactions of the h3 tail domain in a model oligonucleosomal array. *J.Biol.Chem.*, 280(39):33552–33557, September 2005. 32

[ZLv98] J. Zlatanova, S.H. Leuba, and Holde K. van. Chromatin fiber structure: morphology, molecular determinants, structural transitions. *Biophys.J.*, 74(5):2554–2566, May 1998. 8

[ZMB91] R.P. Zinkowski, J. Meyne, and B.R. Brinkley. The centromere-kinetochore complex: a repeat subunit model. *J.Cell Biol.*, 113(5):1091–1110, June 1991. 54, 55

[ZMF+02] Zoran Z Zdraveski, Jill A Mello, Christine K Farinelli, John M Essigmann, and Martin G Marinus. Muts preferentially recognizes cisplatin- over oxaliplatin-modified dna. *J Biol Chem*, 277(2):1255–1260, Jan 2002. 96

[ZR01] Y. Zhang and D. Reinberg. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.*, 15(18):2343–2360, September 2001. 8, 31

[Zuk03] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, July 2003. 57