

# THE PACKAGING OF DNA IN CHROMATIN

BY GUOHUI ZHENG

A dissertation submitted to the  
Graduate School—New Brunswick  
Rutgers, The State University of New Jersey  
in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy  
Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Dr. Wilma K. Olson

and approved by

---

---

---

---

New Brunswick, New Jersey

January, 2010

## **ABSTRACT OF THE DISSERTATION**

### **The Packaging of DNA in Chromatin**

**by Guohui Zheng**

**Dissertation Director: Dr. Wilma K. Olson**

The mechanical properties of DNA play a key role in its biological processing, determining how the long, thin, double-helical molecule responds to the binding of proteins and functions in confined spaces within a cell. In eukaryotes, about 75 – 90% of genomic DNA exists in the form of nucleosomes, which are the fundamental units of DNA packaging in chromatin and the primary determinate of DNA accessibility. The structure of chromatin undergoes various changes that depend, at least in part, upon the requirements of gene expression and other functional environments. The dynamics of DNA packaging in chromatin is thus fundamental to numerous biological processes.

The flexibility of DNA is important in packaging DNA over lengths comparable to its persistence length during genetic processing and the sequence-dependent properties of DNA determine the positioning of nucleosomes in the genome and the sites of binding of enzymes and transcription factors. In addition, understanding the correlation between DNA flexibility and histone-DNA interactions inside the nucleosome is essential for unraveling currently unsolved mechanisms of gene regulation. Furthermore, although many experimental techniques have emerged to examine the overall structure of chromatin fibers, the internal arrangement of DNA and histones remains unclear. Thus an appropriate computational model able to incorporate experimental observations is key to interpretation of the folding and unfolding of chromatin.

The major goal of this thesis is to understand some of biophysical mechanisms involved in the packaging of DNA into chromatin using computational techniques at multi-scales: (i) to determine the sequence-dependent flexibility of DNA by developing DNA deformation analysis tools and databases; (ii) to design DNA spatial configurations using knowledge-based Monte-Carlo sampling; (iii) to map protein-DNA recognition inside nucleosomes in terms of realistic molecular treatments; and (iv) to interpret the internal structure of chromatin fibers and examine chromatin looping using novel modeling and simulation methods.

## Acknowledgements

I would like to express my deep gratitude to my adviser Prof. Wilma K. Olson for her passionate inspiration and insightful guidance. I want to thank her for offering me various opportunities to enhance my scientific skills and to explore great ideas in my Ph.D projects. I really appreciate her patience and enthusiasm in helping improve my English and writing skills.

I would like to also have my warm thanks to other members in Dr. Olson's group.



## Dedication

To my wife Ting He and my parents.

# Table of Contents

<b>Abstract</b> . . . . .	ii
<b>Acknowledgements</b> . . . . .	iv
<b>Dedication</b> . . . . .	v
<b>List of Tables</b> . . . . .	xi
<b>List of Figures</b> . . . . .	xii
<b>1. Introduction</b> . . . . .	1
1.1. DNA flexibility . . . . .	2
1.2. Nucleosome organization . . . . .	4
1.2.1. Histone fold . . . . .	5
1.2.2. Histone association . . . . .	5
1.2.3. The NCP . . . . .	6
1.2.4. Histone tails . . . . .	7
1.3. Chromatin folding . . . . .	8
1.3.1. Molecular models . . . . .	8
1.3.2. Computer simulation . . . . .	10
<b>References</b> . . . . .	18
<b>2. Web 3DNA — a webserver for the analysis, reconstruction, and visu- alization of three-dimensional nucleic-acid structures</b> . . . . .	23
2.1. Introduction . . . . .	23
2.2. Materials and methods . . . . .	25
2.2.1. Base coordinate frames . . . . .	25
2.2.2. Base-pair identification . . . . .	25

2.2.3.	Rigid-body parameters . . . . .	26
2.3.	Webserver . . . . .	27
2.3.1.	Analysis component . . . . .	27
2.3.2.	Reconstruction component . . . . .	29
2.3.3.	Visualization component . . . . .	31
2.3.4.	Tutorial . . . . .	32
2.4.	Technical Details . . . . .	33
2.4.1.	Structure Analysis . . . . .	33
2.4.2.	Model Reconstruction . . . . .	35
2.4.3.	Molecular Visualization . . . . .	37
2.5.	Concluding Remarks . . . . .	38
	<b>References . . . . .</b>	<b>47</b>
<b>3.</b>	<b>3DNALandscapes: a database for exploring the conformational features of DNA . . . . .</b>	<b>50</b>
3.1.	Introduction . . . . .	50
3.2.	Database Content . . . . .	52
3.2.1.	Structures . . . . .	52
3.2.2.	Backbones . . . . .	53
3.2.3.	Sugar-base side groups . . . . .	53
3.2.4.	Base pairs . . . . .	53
3.2.5.	Base-pair steps . . . . .	54
3.2.6.	Complementary-strand interactions . . . . .	55
3.3.	Web Interface . . . . .	56
3.3.1.	Structure filter . . . . .	56
3.3.2.	Parameter- and context-selection panels . . . . .	57
3.3.3.	Data report . . . . .	58
3.3.4.	Local summary and visualization . . . . .	59
3.4.	Concluding Remarks and Future Directions . . . . .	59

<b>References . . . . .</b>	<b>64</b>
 <b>4. Sequence-dependent flexibility of DNA . . . . .</b>	 <b>66</b>
4.1. Introduction . . . . .	66
4.2. Methods . . . . .	67
4.2.1. Non-redundant Structures . . . . .	67
4.2.2. Chain Model and Dimensions . . . . .	68
4.2.3. Deformation Energy . . . . .	69
4.2.4. Configurational Sampling . . . . .	71
4.2.5. Persistence Length . . . . .	71
4.2.6. J-Factor . . . . .	72
4.2.7. DNA Threading . . . . .	72
4.3. Results . . . . .	73
4.3.1. Knowledge-based potentials . . . . .	73
4.3.2. Intrinsic motions . . . . .	75
4.3.3. Persistence length . . . . .	75
4.3.4. Radial distribution . . . . .	77
4.3.5. Threading score . . . . .	78
4.4. Concluding Remarks . . . . .	79
 <b>References . . . . .</b>	 <b>88</b>
 <b>5. DNA simulation: How stiff is DNA? . . . . .</b>	 <b>91</b>
5.1. Introduction . . . . .	91
5.2. Methods . . . . .	92
5.2.1. DNA model . . . . .	92
5.2.2. Gaussian sampling . . . . .	93
5.2.3. DNA reconstruction . . . . .	94
5.2.4. DNA end-to-end distance and contour length . . . . .	95
5.2.5. Tether model . . . . .	96
5.2.6. DNA-tether interactions . . . . .	97

5.2.7. Monte-Carlo simulation . . . . .	98
5.3. Results and discussion . . . . .	99
5.3.1. Global fluctuations of DNA . . . . .	99
5.3.2. Effects of rigid tethers . . . . .	100
5.3.3. Simulated distributions of end-to-end distances . . . . .	102
5.4. Conclusions . . . . .	104
<b>References . . . . .</b>	<b>116</b>
 <b>6. Cylindrical view of the nucleosome core particle . . . . .</b>	 <b>119</b>
6.1. Introduction . . . . .	119
6.2. Methods . . . . .	120
6.2.1. NCP reference frame . . . . .	120
6.2.2. Coordinate systems . . . . .	121
6.3. Results . . . . .	123
6.3.1. Molecular organization . . . . .	123
6.3.2. Molecular contacts . . . . .	124
6.3.3. Distribution of charges . . . . .	125
6.3.4. Electrostatic potential . . . . .	127
6.4. Concluding remarks . . . . .	127
<b>References . . . . .</b>	<b>135</b>
 <b>7. Coarse-grained modeling of the chromatin . . . . .</b>	 <b>136</b>
7.1. Introduction . . . . .	136
7.2. Methods . . . . .	138
7.2.1. Nucleosome core-particle model . . . . .	138
7.2.2. Charge distribution model . . . . .	138
7.2.3. Linker DNA model . . . . .	139
7.2.4. Gaussian sampling . . . . .	140
7.2.5. Electrostatic interactions . . . . .	141
7.2.6. Rigid-body representations . . . . .	141

7.2.7. Excluded volume . . . . .	143
7.3. Model details . . . . .	143
7.3.1. Ion pairs . . . . .	143
7.3.2. Reduction of charges . . . . .	145
7.3.3. Clustering analysis . . . . .	145
7.3.4. Electrostatic potential . . . . .	147
<b>References . . . . .</b>	<b>159</b>
<b>8. Histone tails enhance distant communication in chromatin . . . . .</b>	<b>162</b>
8.1. Introduction . . . . .	162
8.2. Methods . . . . .	164
8.2.1. Long-range enhancer-promoter (E-P) interactions . . . . .	164
8.2.2. E-P distance . . . . .	164
8.2.3. Radius of gyration . . . . .	165
8.2.4. Neighbor density . . . . .	166
8.3. Results and discussion . . . . .	166
8.3.1. Distribution of E-P distances . . . . .	166
8.3.2. Simulated enhancement of E-P communication . . . . .	167
8.3.3. Distribution of the radius of gyration . . . . .	168
8.3.4. Nucleosome-nucleosome contacts . . . . .	169
8.3.5. Nucleosome-nucleosome flexibility . . . . .	170
8.4. Conclusion . . . . .	171
<b>References . . . . .</b>	<b>184</b>
<b>Appendix A. Determination of nucleosome step parameters . . . . .</b>	<b>186</b>
<b>References . . . . .</b>	<b>190</b>
<b>Vita . . . . .</b>	<b>191</b>

## List of Tables

2.1. Table of regular DNA and RNA helical models . . . . .	46
4.1. Persistence lengths of hypothetical, naturally straight DNA homopoly- mers with knowledge-based elastic properties for individual base-pair steps	86
4.2. Sequences and ring-closure properties of representative 94-bp DNA . . .	87
5.1. Observed vs. computed distances and fluctuations between gold nanocrys- tals attached to DNA chains . . . . .	115
7.1. Clusters of histone tail charges . . . . .	158
7.2. Clusters of histone core charges . . . . .	158
8.1. Size of Monte-Carlo simulations of chromatin-mediated systems . . . . .	182
8.2. The positions of the peaks in the distribution of E-P distances . . . . .	182
8.3. Comparison of the simulated enhancement of the E-P communication .	182
8.4. Peak positions and relative shifts in the distribution curves . . . . .	183
8.5. The partition of the distribution function of the radii of gyration . . . .	183

## List of Figures

1.1. Block diagram of local base-pair step parameters . . . . .	12
1.2. The two types of histone-dimer domains (H3-H4 and H2A-H2B) . . . .	13
1.3. Ribbon view of the core histones . . . . .	14
1.4. The (H3-H4) <sub>2</sub> tetramer . . . . .	15
1.5. The views down and along the side of the nucleosome core particle . . .	16
1.6. Models of the internal organization of the 30-nm chromatin fiber . . . .	17
2.1. The front page of the w3DNA analysis component . . . . .	39
2.2. Screenshots illustrating the information provided in the analysis of nucleic- acid-containing structures . . . . .	40
2.3. The front page of the w3DNA ‘Reconstruction’ component . . . . .	41
2.4. Representations of nucleic-acid-containing structures generated with the reconstruction component of the w3DNA server . . . . .	42
2.5. The front page of the w3DNA ‘Visualization’ component . . . . .	43
2.6. An intermediate page of the ligand-decorated DNA reconstruction sub- component . . . . .	44
2.7. Examples of the unique representations of nucleic-acid structures avail- able through the w3DNA server . . . . .	45
3.1. Screenshots illustrating some of the information about the DNA sugar- phosphate torsion angles . . . . .	61
3.2. Screenshots showing some of the information about the arrangements of DNA base-pair steps . . . . .	62
3.3. Screenshots showing some of the data provided in the local summary and visualization pages . . . . .	63



4.1. Collective scatter plots in the roll-twist $(\theta_2, \theta_3)$ plane of base-pair step parameters . . . . .	81
4.2. Sequence-dependent motions along the longest principal axes of the 10 unique DNA base-pair steps . . . . .	82
4.3. Contour surfaces in the roll-tilt $(\theta_2, \theta_1)$ plane . . . . .	83
4.4. Distributions of the end-to-end distances $r$ for a series of 94-bp DNA molecules . . . . .	84
4.5. Deformation profiles of representative DNA sequences ‘threaded’ on the central 60 base-pair steps . . . . .	85
5.1. (A) Atomic-level representations of a 10-bp DNA duplex with gold nanocrystals (large spheres) attached via short tethers to the 3’-ends of complementary strands . . . . .	107
5.2. Chain-length dependence of (A) the average end-to-end distances and contour lengths . . . . .	108
5.3. Effects of tethers on the end-to-end variance of a 15-bp ideal, inextensible DNA duplex . . . . .	109
5.4. Scatter plots of the covariance of DNA and nanocrystal end-to-end distances . . . . .	110
5.5. Chain-length dependence of (A) the variance of the DNA-DNA and Au-Au end-to-end distances . . . . .	111
5.6. Simulated probability density distributions of the distances $r_{Au}$ between gold nanocrystals attached via ‘stiff’, extended tethers to the ends of ideal, inextensible DNA duplexes . . . . .	112
5.7. Chain-length dependence of (A) the average and (b) the associated variance of the end-to-end distance of long, ideal, inextensible, twisted worm-like chains . . . . .	113
5.8. Simulated probability density distributions of the end-to-end distance $r_{Au}$ between gold nanocrystals attached via ‘stiff’, extended tethers . . .	113

5.9. Simulated probability density distributions of the end-to-end distance $r_{Au}$ between gold nanocrystals attached via ‘stiff’, extended tethers to the ends of sequence-dependent DNA duplexes . . . . .	114
5.10. Simulated probability density distributions of the end-to-end distance $r_{Au}$ between gold nanocrystals attached via ‘flexible’, extended tethers to the ends of ideal, inextensible DNA duplexes . . . . .	114
6.1. Scatter plot of the cylindrical radii of NCP atoms . . . . .	129
6.2. Scatter plot of the cylindrical phase angle $\theta$ of nucleosome atoms . . . . .	130
6.3. Scatter plot of the cylindrical height of nucleosome atoms . . . . .	130
6.4. Scatter plots of nucleosome atoms on the cylindrical $(R, \theta)$ plane . . . . .	131
6.5. Scatter plots of nucleosome atoms on the cylindrical $(Z, \theta)$ plane . . . . .	131
6.6. Atomic contacts between histone proteins and the nucleosomal DNA . . . . .	132
6.7. Atoms of the (H3-H4) tetramer and the H2A-H2B dimers in contact with other protein atoms in the NCP . . . . .	132
6.8. Sequences of core histones with tail regions underscored . . . . .	133
6.9. Cylindrical view of histone charges. DNA is represented at the all-atom level to provide a base line . . . . .	133
6.10. Number of charges within a given radial cutoff . . . . .	134
6.11. Distribution of charges on histones . . . . .	134
7.1. Two-dimensional map of the charge distribution on NCP-147 . . . . .	150
7.2. Long-distance electrostatic effect of an ion pair . . . . .	151
7.3. Ion pairs within a 7-Å cutoff in NCP-147 . . . . .	152
7.4. Two-dimensional cylindrical map of the charged atoms of NCP-147 . . . . .	153
7.5. Two-dimensional cylindrical map of charge clusters . . . . .	154
7.6. Centers of clustered charges within NCP-147 . . . . .	155
7.7. Potential of the NCP in the $(Z, \theta)$ plane . . . . .	156
7.8. Potential of the NCP in the $(r, \theta)$ plane . . . . .	157
8.1. The experimental systems used for the measurement of transcription-activation levels in both chromatin and naked DNA systems . . . . .	172

8.2. Schematic of the looping of DNA mediated by long-range interactions of NtrC and RNA polymerase . . . . .	172
8.3. Configurations of two 25-nucleosome chromatin fibers . . . . .	173
8.4. Distributions of E-P distances mediated by simulated chromatin fibers .	174
8.5. Distributions of E-P distances mediated by simulated protein-free DNA	175
8.6. Relative likelihood of looping probability for chromatin- vs. free-DNA chains . . . . .	176
8.7. Distributions of radius of gyration . . . . .	176
8.8. Contact number density histograms . . . . .	177
8.9. Break-down of the contact-number densities for each nucleosome in a selected ensemble of samples . . . . .	178
8.10. Frequency maps of sequential contacts between nucleosomes within simulated 50-nucleosome chromatin fibers . . . . .	179
8.11. Frequency of sequential contacts plotted against the sequential separation interval $m$ . . . . .	180
8.12. Distribution of the shearing between consecutive nucleosomes . . . . .	181
8.13. Distribution of the bending of consecutive nucleosomes . . . . .	181
A.1. Illustration of nucleosome step parameters . . . . .	189

# Chapter 1

## Introduction

DNA is one of the most fundamental biological materials; it stores genetic information and guides the biological machinery involved in the production of proteins and other biomolecular products [1]. In addition to the genetic message, DNA base sequence carries a multitude of structural and energetic signals related to its biological packaging and processing. In the cell, DNA does not exist as a linear double helix. Instead, it is tightly bent in the confines of the cell, and frequently exists in a circular form [2]. In viruses, DNA is highly packaged and constrained in a small capsid, under conditions of high internal pressure; in eukaryotic cells, about 75 – 90% of genomic DNA is packaged in nucleosomes [3]. In all cases, DNA exists along with many different proteins and other components, which in turn regulate genetic processes, such as DNA replication, transcription, and repair.

The nucleosome plays a key role in the regulation of gene transcription in eukaryotes. It is the basic repeating unit of DNA packaging in chromatin and the primary determinant of DNA accessibility in the cell [4]. In the nucleosome, DNA is tightly wrapped around an assembly of specific proteins called histones. The presence of nucleosomes poses barriers for transcription factors to access the DNA and therefore blocks transcription, replication, and other processes. On the other hand, the nucleosome also possesses dynamic properties that allow for the remodeling DNA-histones associations and temporarily release DNA to transcription factors and other genetic regulators. However, this requires the assistance of various protein complexes under specific conditions [5, 6, 7].

Chromatin is formed by the connection and association of repeating nucleosome units. It also contains various proteins that regulate the assembly of chromatin, the remodeling

of nucleosomes, and the transcription of DNA. The structure of chromatin undergoes various changes, when depend, at least in part, on the requirements of transcription. The dynamic folding and unfolding of chromatin, involving various functional environments, are fundamental to the regulation of gene expression [8, 9]. Understanding the mechanisms of chromatin remodeling along with DNA methylation has become a central mission of epigenetics, which refers to changes in gene expression without modifying the underlying DNA sequence. Epigenetics, related to many human disease and aging, is becoming the center of modern medical and biological research.

### 1.1 DNA flexibility

The flexibility of DNA is important in the packaging over lengths comparable to its persistence length during genetic processes such as recombination, transcription, etc. [10, 11, 12]. DNA exhibits flexibility in bending, twisting, stretching, and shearing. Among these, DNA bending is the most notable and commonly relevant to DNA-protein interactions, as many DNA-binding proteins bend DNA. Another important mechanical property of DNA is twisting, a rotational motion about the DNA axis, giving rise to variations in the numbering helical turns of DNA [13]. Bending and twisting together play a key role in DNA supercoiling and packing. DNA stretching and shearing, which are translational movements along and perpendicular to the helical axis, present only a small degree of flexibility in DNA deformation. However, examination of X-ray structures has shown that shearing coupled with bending is key to DNA folding on the nucleosome [14].

DNA flexibility is sequence-dependent [15, 16]. That is, distinct DNA sequences can lead to large difference in the relative ease of bending, twisting, and shearing. This property of double-helical DNA plays a crucial role in its recognition by proteins, which are able to act on specific sites in a given sequence by way of direct readout through base-amino acid contact and indirect readout through DNA deformation [17, 18, 19]. An important example of such sequence-dependent DNA-protein recognition is nucleosome positioning, which refers to the selection of specific sites on DNA by the histone proteins to form nucleosomes. A DNA sequence with periodically spaced ( $\sim 10$  bp) AA/TT/TA

base-pair steps and an intervening oscillating occurrence of GC-rich base-pair steps can significantly enhance the binding of histones [20]. This pattern and similar motifs have been observed in both natural and synthesized sequences [20, 21, 22].

The flexibility of DNA can be characterized in various ways. A knowledge-based approach of flexibility measurement was originally proposed by Olson et al. in 1998 [15]. This method examines the flexibility of DNA in terms of the six base-pair step parameters (tilt, roll, twist, shift, slide, rise) (Fig. 1.1) that characterize the rotational and translational arrangements of adjacent base pairs in a set of crystal structures. In this early work, however, only 93 crystal structures of protein-DNA complexes and 63 samples of unbound B-DNA molecules were available in the Protein Data Bank (PDB) [23] and Nucleic Acid Database (NDB) [24] for the analysis. Moreover, even among this small set of samples, there exists some redundant structures, which could bias the statistical inferences. After a decade, the number of high-resolution DNA-containing structures in the public databases, the PDB and NDB, has grown to a much larger number, providing a more reliable data source for us to re-examine the DNA flexibility in an unbiased way. Chapter 2-4 present related techniques, databases, and interfaces of advance to such an analysis. Here, we not only re-examine the elasticity of DNA but also offer our data and tools to the public through user-friendly web-based interfaces. The new web-based 3DNA interface (w3DNA), with which one is able not only to analyze the deformation of DNA but also to construct and visualize three-dimensional structures of nucleic acids is described in Chapter 2. The web-based 3DNALandscapes database — which presents not only the base-pair and base-pair-step parameters used for characterizing DNA flexibility but also other conventional structural information, such as hydrogen bonds, dihedral angles, groove widths, etc. — is presented in Chapter 3. In Chapter 4, we summarize some of the statistical results from the analysis of the rigid-body parameters in a non-redundant set of structures and computational applications based on this data.

The flexibility of DNA can be also estimated with newly emerging single-molecule techniques. Since the early 1990s, a number of force-experiment techniques have been designed to manipulate and measure single molecules of DNA, such as optical tweezers,

magnetic tweezers, bio-membrane force probes, glass micro-needles, and so on. The common feature of the above techniques is that, the two ends of a DNA molecule are respectively attached to a molecular surface and a force sensor. By measuring the displacement of the force sensor, under different manipulations of DNA, the mechanical properties of DNA can be estimated to some extent. A recent development in single-molecule techniques is a molecular ‘ruler’, which can measure the distance between two test points on a molecule in solution using small angle X-ray scattering [25]. This new method was recently applied to measure the end-to-end distance of short-length DNA labeled by soluble gold nanocrystal probes at the ends of the helix. The sample model used in the interpretation of this work suggested that the labeled DNA has a significantly larger stretching flexibility than that extracted from force-extension experiments and the analysis of high-resolution crystal structures. In Chapter 5, we use our knowledge-based Monte-Carlos simulation to carefully interpret the system and show that the classical elastic rod model can still account for the scattering measurements. That is, the DNA yields similar flexibility to that observed in X-ray structures as reported in Chapter 4.

## 1.2 Nucleosome organization

The nucleosome is formed as soon as a DNA is synthesized in the nucleus. The proteins involved in forming the nucleosome are called histones, and make up about half the mass of a eukaryotic chromosome [1, 26]. The main part of a nucleosome is called the nucleosome core particle (NCP), which contains four types of core histones: H3, H4, H2A, and H2B, each of which has two copies. The association of these eight core histones provides a disc-link-shaped ramp for  $\sim 147$  bp of DNA to wrap around it [9]. The linker DNA, a length of about 20 – 50 bp, connects successive NCPs. It is commonly associated with linker histone, H1 or its variant H5.

### 1.2.1 Histone fold

The core histones adopt a common folding pattern called the helix-loop-helix motif, that is, three  $\alpha$ -helices connected by two loops. This structural motif, called the histone fold, was first named by Moudrianakis et al. (1991) [27, 4, 28]. Fig. 1.3 illustrates the protein folding of the eight core histones in the highest resolution ( $1.9\text{\AA}$ ) nucleosome crystal structure (PDB ID: 1KX5) found to date. Each histone fold consists of a long central  $\alpha$ -helix associated with two short helices (C-terminal helix and N-terminal helix) at either end. These three elemental helices are connected by two short loops in such a way that the short terminal helices are folded back and rotated over the central helix [29]. Most histones have a long flexible tail, which can adopt a variety of conformations and can even fold in an  $\alpha$ -helix, such as in H3, or a  $\beta$ -strand, such as in H2A. Histone tails are thought to play essential roles in nucleosome assembly and the higher-order organization of nucleosomes.

### 1.2.2 Histone association

The first level of core histone association is two heterodimeric pairs: H3-H4 and H2A-H2B. They share a so-called “handshake” structural motif [27], that is, two histones associate in a head-to-tail manner (Fig. 1.2). In the dimer, N- and C-terminal helices from different histones are brought close together by a short two-stranded  $\beta$ -bridge. In addition, loops in the N- and C-terminal sides also interact between histones. The dimer is organized in a two-fold symmetry, with respect to a symmetric axis that is different from the NCP dyad axis.

The second level of histone associations is the  $(\text{H3-H4})_2$  tetramer organization of two copies of H3-H4 heterodimers (Fig. 1.4). The overall structure of the tetramer resembles a partially twisted open horseshoe, placing the two H3-H4 dimers in a two-fold symmetric arrangement about the NCP dyad axis [27, 4]. The tetramer is stabilized by a four-helix bundle made up of two C-terminal helices and two central helices from the H3 histones ( $\text{H3}\alpha$  and  $\text{H3}\beta$ ) in the two dimer halves. Fig. 1.4 shows a ribbon



diagram of the histone tetramer, along with the central 61 base-pair steps of the nucleosomal DNA, in order to present a general picture of the position of the tetramer. The  $(\text{H3-H4})_2$  tetramer is the central histone component of the NCP, and is thought to be the first unit to compact DNA into chromatin [30]. The histone tetramer can recognize specific positions of DNA and bind the double-helical molecule without association with other histones, in a process known as nucleosome positioning. Biochemical experiments also show that a single  $(\text{H3-H4})_2$  tetramer appears to wrap a DNA of around 146 bp with the same positioning as the entire histone octamer [31].

The third level of the association is the core histone octamer, where the  $(\text{H3-H4})_2$  tetramer associates cooperatively with two H2A-H2B dimers. The octamer resembles a left-handed disc-like-shaped ramp. DNA wraps onto the ramp and yields a left-handed superhelix with an axis perpendicular to the disc plane. In the octamer organization, one H2A-H2B dimer binds to one side of the  $(\text{H3-H4})_2$  tetramer, forming the top face of the disc plane, whereas the other H2A-H2B dimer binds to the other side of the tetramer on the bottom face. The two H2A-H2B dimers are related by a two-fold symmetry axis, with interactions between the C-terminal halves of H2B and H4, which form a four-helix bundle similar to the association between histones H3 $\alpha$ -H3 $\beta$  in the tetramer [32]. The interface of the H2B-H4 bundle, however, is less hydrophobic than that of the H3 $\alpha$ -H3 $\beta$  bundle. Thus the association of the H2A-H2B dimer with the  $(\text{H3-H4})_2$  tetramer is weaker than the association of the two halves of the tetramer. This difference between associated interfaces may be a reason for the instability of the histone octamer compared to the tetramer at low salt concentration [4]. Nonetheless, in terms of interacting areas, the interface between the H2A-H2B dimer and the tetramer is larger than that between the two halves of the tetramer, but it is more open and accessible to solvent [27, 4].

### 1.2.3 The NCP

The NCP is formed by wrapping a DNA fragment of around 145 – 147 bp onto the left-handed octamer ramp like thread around a spool [1, 33]. The overall shape of the core particle roughly resembles a short cylinder with a diameter of about 105Å and

height around  $65\text{\AA}$  (Fig. 1.5). The nucleosomal DNA is folded in about 1.65 turns along the superhelical pathway with the central base-pair positioned at the pseudo two-fold dyad of the core particle. The DNA sequence is divided nearly evenly into two parts by the dyad symmetry axis. Examination of available three-dimensional crystal structures of the NCP, shows that (i) the DNA bends such that the major groove faces inward towards the center of the NCP at the dyad position and (ii) starting at positions displaced by an interval of around 5 bp with respect to the dyad, the minor groove faces inward and major groove faces outward around the octamer with a periodicity of about 10 bp. When the minor groove faces inward, the DNA binds to histones via interactions with either the C-terminal helices or the helix-connecting loops of the histone folds. It is thought that, the tight bending of DNA around the octamer is mediated by insertions of arginine/lysine side chains into the DNA minor groove [4, 29, 34].

#### 1.2.4 Histone tails

There are histone regions that extend beyond the NCP disk. These regions, called histone tails, are located at the termini (N- and C- termini) of the histone chains. About one quarter of the total mass of the core histones is contributed by their tails [35], and about one third of the amino-acid residues in the N-terminal tails are either lysines or arginines, which are positively charged [36]. All core histones have an N-terminal tail, whereas H2A and H2B also have a C-terminal tail [4]. These core histone tails exit the NCP mainly through the minor grooves of the DNA. Once exiting the nucleosome disc, the tails are believed to behave as random-coil segments directed away from the core particle [37]. The histone tails play a crucial role in the chromatin system at many levels, including stabilization of the folding of oligonucleosome arrays into chromatin fibers, assistance in the fiber-fiber interactions involved in higher-order organization, and regulation of the accessibility of nucleosomal DNA to other DNA-binding proteins [35, 36, 38, 39, 40, 41, 42]. It has been shown that, by removing all core histone tails, nucleosome arrays cannot be organized into the 30-nm chromatin fiber, even in the presence of bound linker histones [40]. Furthermore, the H3/H4 tails play a major role in regulating DNA accessibility, which is confirmed by the fact that deletion of the

H3/H4 tails alone leads to increased binding of transcription factors [42]. A similar conclusion, obtained from equilibrium dynamic experiments, suggests that removal of the histone-tail domains would greatly raise transcriptional site exposure [41].

To carefully examine the histone and DNA organization in the NCP, we have developed a novel shape-based reference frame, also called a cylindrical reference frame, which can be easily used to describe atomic spatial locations and interactions in a quantitative way. Chapter 6 reports the definition of the cylindrical reference frame and its application in mapping NCP atoms, contacts, and charges.

## 1.3 Chromatin folding

### 1.3.1 Molecular models

Nucleosomes further associate into chromatin fibers. The hierarchy of higher-order nucleosome organization can be divided into three levels, in a manner that mimics the way one defines the hierarchy of protein structures. Nucleosome core particles, joined by DNA linkers, form a linear nucleosome array under low salt condition with a diameter of about 10 nm, which is also the diameter of the individual nucleosomes. In this level, the array is called the 10-nm chromatin fiber or beads-on-a-string configuration, which is analogous to primary structure. Under higher ionic strength, the 10-nm fiber can be folded into a much more highly compacted form with a diameter around 30nm, which is referred to as the 30-nm fiber, and is analogous to secondary structure. Beyond this point, the 30-nm fibers can associate into higher-level organization through direct or long-distance interactions. Such fiber-fiber associations are thought to be analogous to tertiary structure [43, 44, 45]. In vivo, most chromatin is maintained in the form of the 30-nm fiber [44]. When it is isolated from the nuclei without any treatment and examined by electron microscopy, chromatin appears to be a fiber of diameter around 30 nm. If chromatin is treated in such a manner that partially unfolds it, then the electron microscopic image shows a fiber of 10 nm in diameter, resembling a “beads on a string” configuration [1].

Although many imaging techniques have emerged to examine the overall structure of

chromatin fibers, the internal arrangement of DNA and histones is still unclear. In the past three decades, a number of folding models have been constructed to postulate the topology of internal organization, such as the solenoidal [46], helical ribbon [47], super bead [48, 49], twisted-ribbon [50], zig-zag [51, 52] and crossed-linker [53, 54] models. Among them, the two most widely cited models are the solenoidal and zig-zag models (Fig. 1.6). The solenoidal model is also called a one-start helix, in which a linear nucleosome array is coiled to form a helical structure with around six nucleosomes per turn; the pseudodyad axis of each nucleosome is almost perpendicular to the solenoid helix axis, and the entry-exit site of DNA in each nucleosome is also arranged to face inward toward the solenoid axis. The zig-zag model is normally referred to as a two-start helix. In the two-start helix model, each linker DNA connects two nucleosomes located on the opposite sides of the fiber with respect to the fiber axis. The plane of each nucleosome disk is nearly perpendicular to the fiber direction [44, 55]. The linker DNA in the one-start solenoidal helix has to be curved in order to connect neighboring nucleosomes, whereas the linker DNA in the two-start zig-zag helix is assumed to be fully straight. Although these ideal models can account for some experimental data, the detailed spatial arrangements of nucleosomes and linker DNA are still under debate, and there are many experimental phenomena that cannot be accounted for by these models [43, 56].

Recently, the X-ray structure of a tetranucleosome has been determined at a low (9Å) resolution [57]. This is the first crystal structure of an oligonucleosome, and has significant implications for understanding the organization of chromatin fibers. The tetranucleosome consists of two stacks of nucleosome core particles, and three linker DNAs connecting the nucleosomes in a zig-zag manner. The structure supports the concept of a two-start zig-zag model rather than a solenoidal model. However, as revealed in the three-dimensional structure, the linker DNA is not fully straight. One of the three linkers is straight and the other two are bent [57]. In fact, this is consistent with observations using electron microscopy (EM) [50] and optical laser tweezers [58], which also suggest that in vivo chromatin may favor irregular zig-zag models [43]. However, a recent study of Robinson and Rhodes [59] using the constraints obtained from an

EM analysis, suggests that in the presence of the linker histone, one per nucleosome, the chromatin fiber forms a left-handed one-start helix with 5.4 nucleosomes per helical turn. Variation of the DNA linkers changes the topology of the fiber [60]. Long DNA linkers allow the binding of linker histones and the formation of a highly regular and compact fiber with a diameter of 33–35 nm; nevertheless, short DNA linkers lead to highly organized nucleosome-nucleosome stacking within a small fiber diameter of about 21 nm and less binding of linker histones [60]. Thus, the chromatin fiber could accommodate various folding structures under different biophysical environments.

### 1.3.2 Computer simulation

A variety of computational models have been developed to investigate chromatin folding, with different levels of details. Woodcock et al. [52] take into account the variation of DNA linker lengths, but treat the DNA linkers as fixed and strictly straight. This model is based on a two-angle model, in which the linker entry-exit angle and the rotation angle between consecutive nucleosomes are allowed to change. Olson and coworkers [61] introduced flexibility in the linker DNA with the nucleosome constrained, and account respectively for nucleosome-nucleosome attractive and nucleosome-DNA repulsive interactions by a simple square-well potential and excluded volume effects. A worm-like-chain model is used to model the linker DNA and the nucleosomes are represented as isotropic spheres. Langowski et al. [62, 63, 64] and Schiessel et al. [65, 55] incorporate the two-angle model and flexibility of linker DNA into their computations, modeling the nucleosomes as ellipsoids that interact via a Gay-Berne potential [66]. Another more realistic approach invented by Schlick et al. [67] uses a so-called Discrete Surface Charge Optimization (DiSCO) representation of the nucleosome [68], where the nucleosome is represented by effective charges located on an irregular cylindrical surface and on the tails. All of these models treat the chromatin in a “static” state, normally seeking a low-energy folded configuration. However, the Brownian dynamics of chromatin fibers is also of interest in studying real biological systems. We present herewith a novel coarse-grained Monte-Carlo method to simulate the chromatin fiber at

the DNA base-pair level, detailed in Chapter 7. This method has been successfully applied to examine long-range interactions between gene promoters and enhancers, which is introduced in Chapter 8.

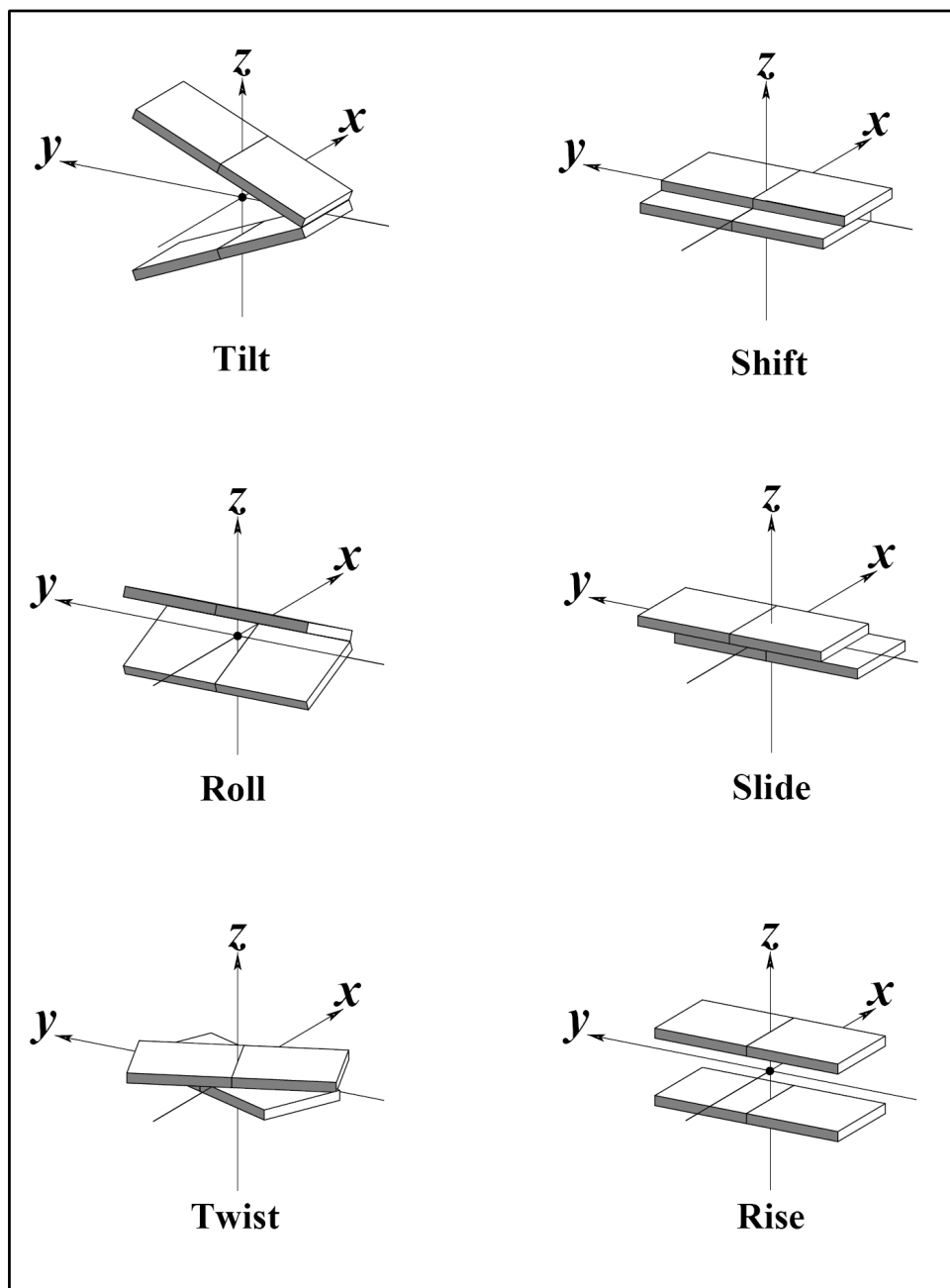


Figure 1.1: Block diagram of local base-pair step parameters. Watson-Crick base pairs are illustrated by blocks with the minor-groove edge color-coded in gray. The displayed reference frame is the standard mean base-pair plane [69]. The  $x$ -axis points toward the major groove and the  $y$ -axis toward the sequence strand.

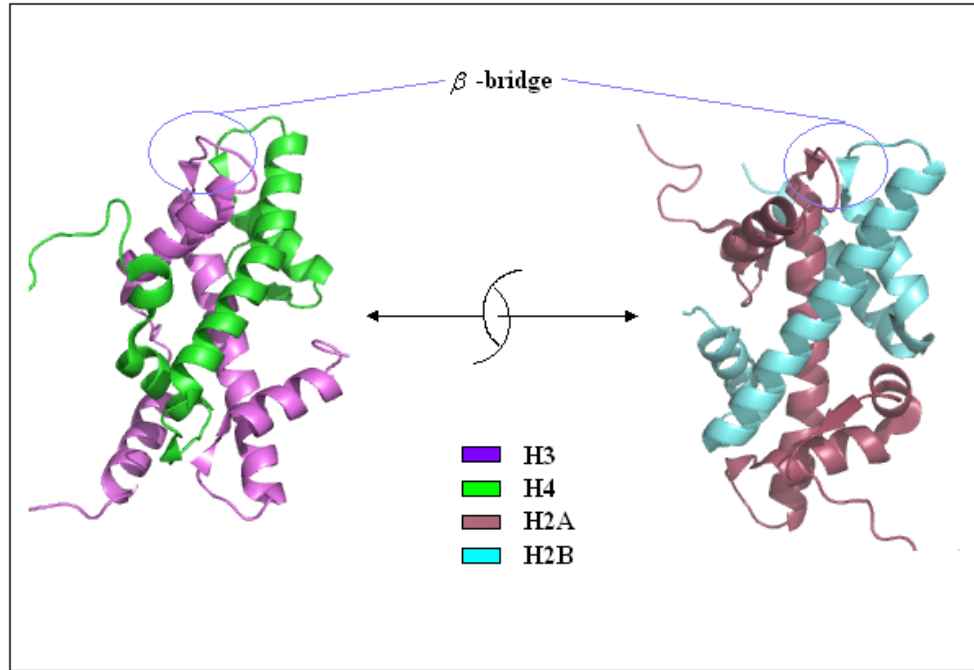


Figure 1.2: The two types of histone-dimer domains (H3-H4 and H2A-H2B). The N-terminal and C-terminal helices from the two paired histones are brought close together, by a short two-strand  $\beta$ -bridge. The two paired histones associate with a two-fold symmetry as well, although the axis of symmetry is not the dyad axis of the nucleosome core particle. This common folding of the two dimers is called a “handshake” motif.



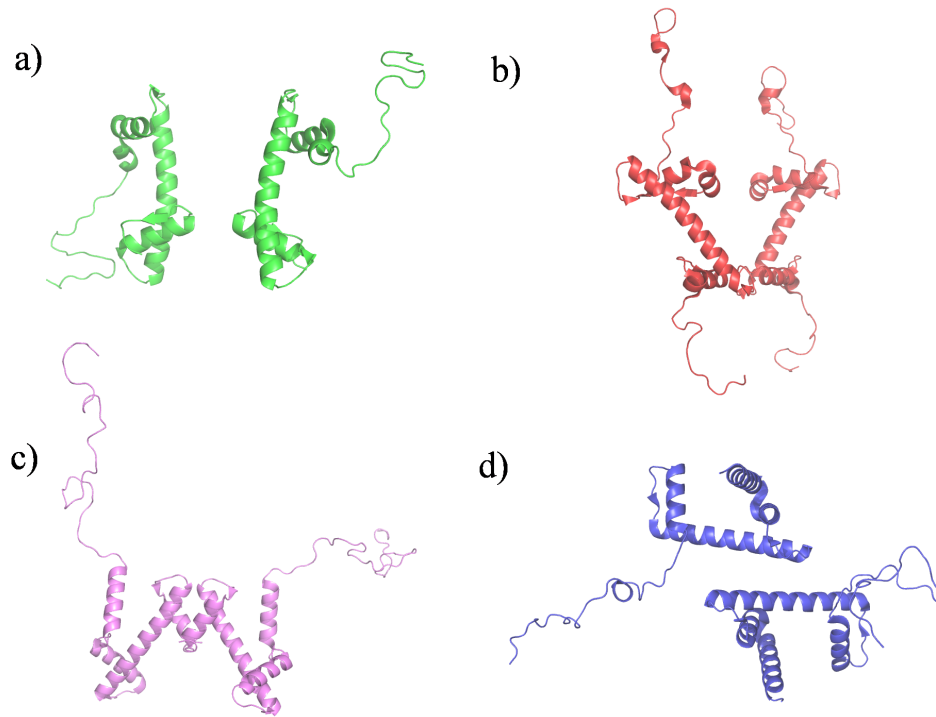


Figure 1.3: Ribbon view of the core histones. Each pair contains two copies of a core histone: a) H4, b) H2A, c) H3, and d) H2B. Each histone shares the same structural motif, called the histone fold. That is, three  $\alpha$ -helices are connected by two loops in a helix-loop-helix manner.

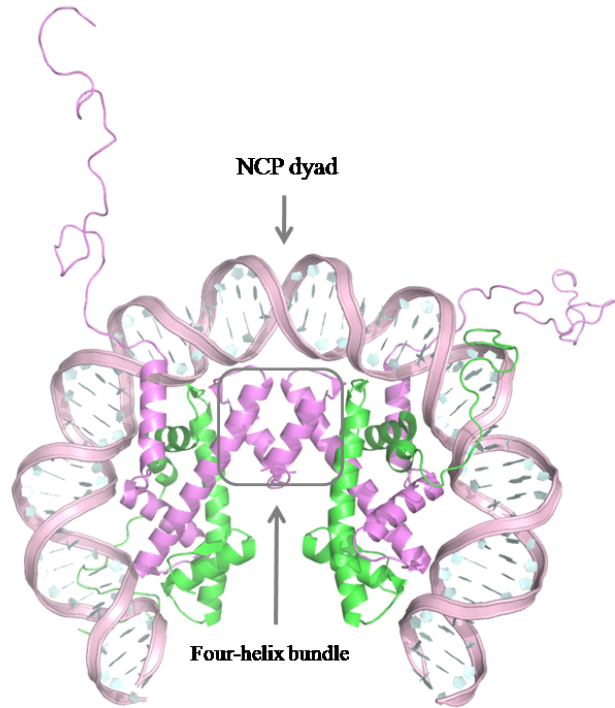


Figure 1.4: The  $(\text{H3-H4})_2$  tetramer. Two  $(\text{H3-H4})$  heterodimers associate via a four-helix bundle to form the tetramer. The dimers are organized in a two-fold symmetry about the NCP dyad axis. Color code: H3 in magenta, H4 in green, and DNA in pink.

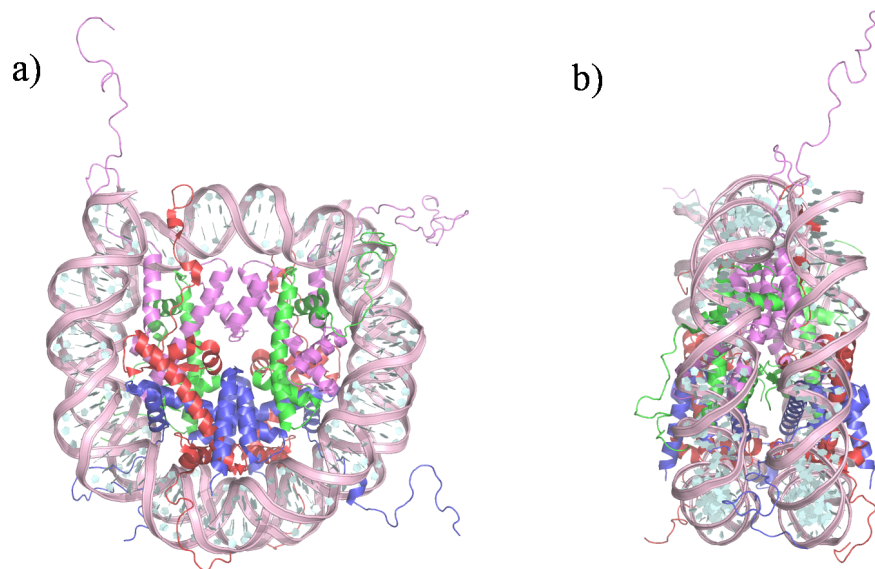


Figure 1.5: The views down and along the side of the nucleosome core particle. A DNA of 147 bp wraps onto the left-handed octamer ramp in about 1.65 turns like a thread around a spool, with the central base-pair positioned at the pseudo two-fold dyad of the core particle. Molecular chains are color-coded as follows: DNA in pink, H3 in magenta, H4 in green, H2A in red, and H2B in blue.

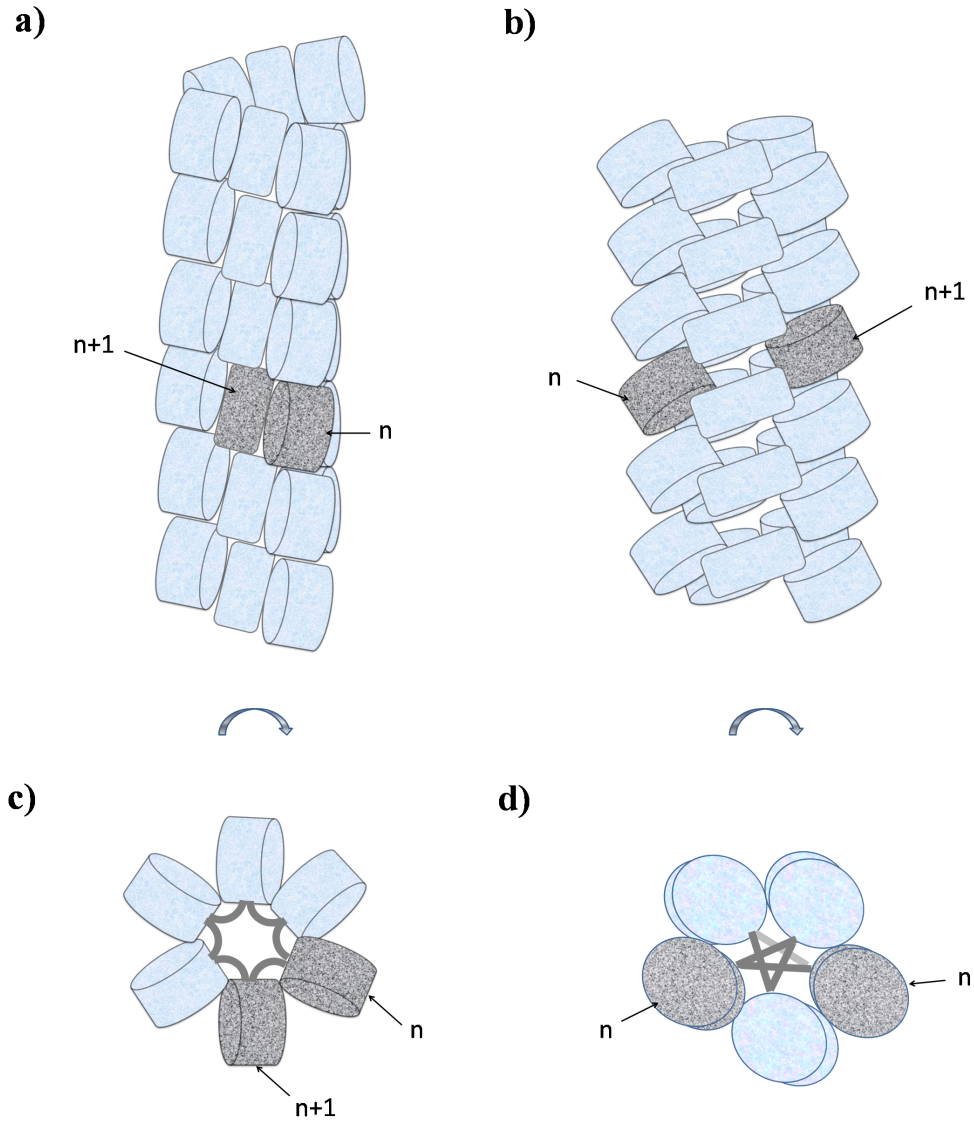


Figure 1.6: Models of the internal organization of the 30-nm chromatin fiber. (a) The solenoidal model. The linear nucleosome array is coiled to form a helical structure with around six nucleosomes per turn, and each linker DNA is curved. (b) The zig-zag model. Each linker DNA connects two nucleosomes located on opposite sides of the fiber by a straight path. Pictures are re-drawn based a modification from Ref. [55].

## References

- [1] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002) *Molecular Biology of the Cell*, 4th Edition, Garland Science, New York.
- [2] Travers, A. A. and Thompson, J. M. T. (2004) An introduction to the mechanics of DNA. *Phil. Trans. R. Soc. Lond. A*, **362**, 1265–1279.
- [3] Purohit, P. K., Kondev, J., and Phillips, R. (2003) Mechanics of DNA packaging in viruses. *Proc. Natl. Acad. Sci., U.S.A.*, **100**(6), 3173–3178.
- [4] Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
- [5] Annunziato, A. (2008) DNA packaging: Nucleosomes and chromatin. *Nature Education*, **1**, 1.
- [6] Li, B., Carey, M., and Workman, J. L. (2007) The role of chromatin during transcription. *Cell*, **128**(4), 707–719.
- [7] Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005) Rapid spontaneous accessibility of nucleosomal DNA. *Nature Structural and Molecular Biology*, **12**(1), 46–53.
- [8] Bradbury, E. M. and van Holde, K. E. (2004) Chromatin structure and dynamics: a historical perspective. In J. Zlatanova and S.H. Leuba, Editors, *Chromatin Structure and Dynamics: State-of-the-Art*, Page 1-11, Elsevier Science & Technology.
- [9] Wolffe, A. P. (1998) *Chromatin: Structure and Function*, Academic Press.
- [10] Crothers, D. M. and Steitz, T. A. (1992) Transcriptional activation by Escherichia coli CAP protein. In S. L. McKnight and K. R. Yamamoto, Editors, *Transcriptional Regulation*, Page 501-534, Cold Spring Harbor Laboratory Press.
- [11] Drew, H. R. and Travers, A. A. (1985) DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.*, **186**(4), 773–790.
- [12] Goodman, S. D. and Nash, H. A. (1989) Functional replacement of a protein-induced bend in a DNA recombination site. *Nature*, **341**(6239), 251–254.
- [13] Bates, A. D. and Maxwell, A. (2005) *DNA Topology*, 2nd Edition, Page 18-19, Oxford University Press, USA.
- [14] Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K., and Zhurkin, V. B. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**(3), 725–738.

- [15] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**(19), 11163–11168.
- [16] Drew, H. R. and Travers, A. A. (1985) Structural junctions in DNA: the influence of flanking sequence on nuclease digestion specificities. *Nucleic Acids Res.*, **13**(12), 4445–4467.
- [17] Arauzo-Bravo, M. J., Fujii, S., Kono, H., Ahmad, S., and Sarai, A. (2005) Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition. *J. Am. Chem. Soc.*, **127**(46), 16074–16089.
- [18] Packer, M. J., Dauncey, M. P., and Hunter, C. A. (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.*, **295**(1), 85–103.
- [19] Xiong, Y. and Sundaralingam, M. (2001) Protein-nucleic acid interaction: major groove recognition determinants. In *Encyclopedia of Life Science*, Pages 1-8, Macmillan Publishers Ltd, Nature Publishing Group.
- [20] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I., Wang, J., and Widom, J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
- [21] Widom, J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, **34**(3), 269–324.
- [22] Travers, A. A. and Drew, H. R. (1997) DNA recognition and nucleosome organization. *Biopolymers*, **44**(4), 423–433.
- [23] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- [24] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R., and Schneider, B. (1992) The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
- [25] Mathew-Fenn, R. S., Das, R., and Harbury, P. A. (2008) Remeasuring the double helix. *Science*, **32**(5900), 446–449.
- [26] Bergand, J. M., Tymoczko, J. L., and Stryer, L. (2002) *Biochemistry*, 5th Edition, Page 875, W.H. Freedman and Company New York.
- [27] Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E., and Moudrianakis, E. N. (1991) The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. Natl. Acad. Sci., U.S.A.*, **88**, 10148–10152.
- [28] Ramakrishnan, V. (1995) The histone fold: evolutionary questions. *Proc. Natl. Acad. Sci., U.S.A.*, **92**, 11328–11330.

- [29] Harp, J. M., Hanson, B. L., Timm, D. E., and Bunick, G. J. (2000) Asymmetries in the nucleosome core particle at 2.5 Å resolution. *Acta Cryst.*, **D56**, 1513–1534.
- [30] Banks, D. and Gloss, L. (2004) Folding mechanism of the (H3-H4)<sub>2</sub> histone tetramer of the core nucleosome. *Protein Sci.*, **13**, 1304–1316.
- [31] Dong, F. and van Holde, K. E. (1991) Nucleosome positioning is determined by the (H3-H4)<sub>2</sub> tetramer. *Proc. Natl. Acad. Sci., U.S.A.*, **88**, 10596–10600.
- [32] Ramakrishnan, V. (1997) Histone structure and the organization of the nucleosome. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 83–112.
- [33] Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, S. L., and Darnell, J. (2004) *Molecular Cell Biology*, 5th Edition, Page 425-426, W.H. Freedman and Company, New York.
- [34] Richmond, T. J. and Davey, C. A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
- [35] Zheng, C. and Hayes, J. J. (2003) Intra- and inter-nucleosomal protein-DNA interactions of the core histone tail domains in a model system. *J. Biol. Chem.*, **287**(26), 24217–24224.
- [36] Hansen, J. (2002) Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 361–392.
- [37] Luger, K. and Richmond, T. (1998) The histone tails of the nucleosome. *Current Opinion in Genetics & Development*, **8**, 140–146.
- [38] Angelov, D., Vitolo, J. M., Mutskov, V., Dimitrov, S., and Hayes, J. J. (2001) Preferential interaction of the core histone tail domains with linker DNA. *Proc. Natl. Acad. Sci., U.S.A.*, **98**(12), 6599–6604.
- [39] Fletcher, T. M. and Hansen, J. C. (1995) Core histone tail domains mediate oligonucleosome folding and nucleosomal DNA organization through distinct molecular mechanisms. *J. Biol. Chem.*, **270**(43), 25359–25362.
- [40] Gordon, F., Luger, K., and Hansen, J. C. (2005) The core histone N-terminal tail domains function independently and additively during salt-dependent oligomerization of nucleosomal arrays. *J. Biol. Chem.*, **280**(40), 33701–33706.
- [41] Polach, K. J., Lowary, P. T., and Widom, J. (2000) Effects of core histone tail domains on the equilibrium constants for dynamic DNA site accessibility in nucleosomes. *J. Mol. Biol.*, **298**, 211–223.
- [42] Zheng, C. and Hayes, J. J. (2003) Structures and interactions of the core histone tail domains. *Biopolymers*, **68**, 529–546.
- [43] Adkins, N. L., Watts, M., and Georgel, P. T. (2004) To the 30-nm chromatin fiber and beyond. *Biochimica et Biophysica Acta*, **1677**, 12–23.

- [44] Widom, J. (1998) Structure, dynamics, and function of chromatin in vitro. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 285–327.
- [45] Zlatanova, J. and Leuba, S. H. (2003) Chromatin fibers: one-at-a-time. *J. Mol. Biol.*, **331**, 1–19.
- [46] Finch, J. T. and Klug, A. (1976) Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 1897–1901.
- [47] Worcel, A., Strogatz, S., and Riley, D. (1981) Structure of chromatin and the linking number of DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 1461–1465.
- [48] Renz, M., Nehls, P., and Hozier, J. (1977) Involvement of histone H1 in the organization of the chromosome fiber. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 1879–1883.
- [49] Zentgraf, H. and Franke, W. W. (1984) Differences of supranucleosomal organization in different kinds of chromatin: cell type-specific globular subunits containing different numbers of nucleosomes. *J. Cell Biol.*, **99**, 272–286.
- [50] Woodcock, C. L., Frado, L. Y., and Rattner, J. B. (1984) The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J. Cell Biol.*, **99**, 42–52.
- [51] Subirana, J. A., Muoz-Guerra, S., Aymam, J., Radermacher, M., and Frank, J. (1985) The layered organization of nucleosomes in 30 nm chromatin fibers. *Chromosoma*, **91**, 377–390.
- [52] Woodcock, C. L., Grigoryev, S. A., Horowitz, R. A., and Whitaker, N. (1993) A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 9021–9025.
- [53] Widom, J. (1989) Toward a unified model of chromatin folding. *Annu. Rev. Biophys.*, **18**, 365–395.
- [54] Williams, S. P., Athey, B. D., Muglia, L. J., Schappe, R. S., Gough, A. H., and Langmore, J. P. (1986) Chromatin fibers are left-handed double helices with a diameter and mass per unit length that depend on linker length. *Biophys. J.*, **49**, 233–248.
- [55] Langowski, J. and Schiessel, H. (2004) Theory and computational modeling of the 30nm chromatin fiber, In J. Zlatanova and S.H. Leuba, Editors, *Chromatin Structure and Dynamics: State-of-the-Art*, Page 397-420, Elsevier Science & Technology.
- [56] Ausio, J. and Abbott, D. W. (2004) The role of histone variability in chromatin stability and folding, In J. Zlatanova and S.H. Leuba, Editors, *Chromatin Structure and Dynamics: State-of-the-Art*, Page 241-290, Elsevier Science & Technology.
- [57] Schalch, T., Duda, S., Sargent, D. F., and Richmond, T. J. (2005) X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, **436**, 138–141.
- [58] Cui, Y. and Bustamante, C. (2000) Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 127–132.



- [59] Robinson, P. J. and Rhodes, D. (2006) Structure of the ‘30 nm’ chromatin fibre: a key role for the linker histone. *Current Opinion in Structural Biology*, **16**, 336–343.
- [60] Routh, A., Sandian, S., and Rhodes, D. (2008) Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc. Natl. Acad. Sci., U.S.A.*, **105**(26), 8872–8877.
- [61] Katritch, V., Bustamante, C., and Olson, W. K. (2000) Pulling chromatin fibers: Computer simulations of direct physical micromanipulations. *J. Mol. Biol.*, **295**, 29–40.
- [62] Wedemann, G. and Langowski, J. (2002) Computer simulation of the 30-nanometer chromatin fiber. *Biophys. J.*, **82**, 2847–2859.
- [63] Langowski, J. (2006) Polymer chain models of dna and chromatin. *Eur. Phys. J. E*, **19**(3), 241–249.
- [64] Langowski, J. and Heermann, D. W. (2007) Computational modeling of the chromatin fiber. *Semin. Cell Dev. Biol.*, **18**(5), 659–667.
- [65] Mergell, B., Everaers, R., and Schiessel, H. (2004) Nucleosome interactions in chromatin: fiber stiffening and hairpin formation. *Phys. Rev. E*, **70**(1), 011915–011923.
- [66] Gay, J. G. and Berne, B. J. (1981) Modification of the overlap potential to mimic a linear site-site potential. *J. Chem. Phys.*, **76**(6), 3316–3319.
- [67] Sun, J., Zhang, Q., and Schlick, T. (2005) Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(23), 8180–8185.
- [68] Beard, D. A. and Schlick, T. (2001) Computational modeling predicts the structure and dynamics of chromatin fiber. *Structure*, **9**, 105–114.
- [69] Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, R. C., Heinemann, U., Lu, X. J., Neidle, S., and Shakked, Z., et al. (1998) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.

## Chapter 2

### **Web 3DNA — a webserver for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures**

The w3DNA (web 3DNA) server is a user-friendly web-based interface to the 3DNA suite of programs for the analysis, reconstruction, and visualization of three-dimensional (3D) nucleic-acid-containing structures, including their complexes with proteins and other ligands. The server allows the user to determine a wide variety of conformational parameters in a given structure — such as the identities and rigid-body parameters of interacting nucleic-acid bases and base-pair steps, the nucleotides comprising helical fragments, etc. It is also possible to build 3D models of arbitrary nucleotide sequences and helical types, customized single-stranded and double-helical structures with user-defined base-pair parameters and sequences, and models of DNA ‘decorated’ at user-defined sites with proteins and other molecules. The visualization component offers unique, publication-quality representations of nucleic-acid structures, such as ‘block’ images of bases and base pairs and stacking diagrams of interacting nucleotides. The w3DNA web server, located at <http://w3dna.rutgers.edu>, is free and open to all users with no login requirement. This chapter includes the introduction of the web server published in *Nucleic Acids Res.* [1] along with a brief technical description of the web server construction.

#### **2.1 Introduction**

DNA and RNA contain layers of biological information, interspersed between or superimposed on the text written in the three-letter codes that provide instructions for making proteins. For example, the structure of the constituent nucleotides governs

access to the sites on DNA and RNA targeted by enzymes and regulatory proteins. Understanding how the nucleic acids fold and how proteins and other ligands recognize and deform the 3D structure are important for comprehending the dynamics of the cell. Interest in understanding the relationship between the global folding of nucleic acids and the sequence-dependent arrangements of the constituent bases and base pairs has stimulated the development of new approaches to analyze and depict DNA and RNA structures. The characterization and visualization of such structures requires detailed knowledge of both the spatial disposition of the constituent bases and bases pairs and the conformation of the intervening sugar-phosphate backbone. Models that take advantage of this information are useful in the formulation of nucleic-acid binding ligands, the interpretation of various nucleic-acid configurational properties, etc.

The 3DNA suite of programs [2, 3] was designed for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid-containing structures, including their complexes with proteins and other ligands. At its core, the software uses a simple matrix-based scheme [4, 5, 6, 7, 8] to calculate the complete set of rigid-body parameters that characterize the orientation and displacement of the base pairs, base-pair steps, and single-stranded nucleotide steps that make up a DNA or RNA structure. The description of structure is geometrically straightforward and the computation of parameters is mathematically rigorous [2, 7, 8], allowing for the exact construction of molecular models based on the derived parameters. Although the software has gained wide use in the scientific community over the past decade, its command-line-driven style is not especially user-friendly, for either novices, i.e. non-Linux/Unix users, or educational purposes.

Here, we report a new, web-based interface that offers user-friendly access to some of the most popular features of the 3DNA package, including: (i) the conformational analysis of arbitrary nucleic-acid-containing structures; (ii) the construction of nucleic-acid models from derived conformational parameters and classic fiber-diffraction models; and (iii) the visualization of local and global nucleic-acid structure from novel and precisely controlled spatial perspectives. The server also contains a database of pre-analyzed nucleic-acid-containing structures stored in the Protein Data Bank (PDB) [9]

and Nucleic Acid Database (NDB) [10] to facilitate user access. The data include conformational information for the asymmetric and biological units of crystal structures and the complete sets of structures determined in NMR studies. Treatment of individual molecular models or ensembles of simulated structures is also possible. The server functions robustly and includes a well-documented tutorial of the program functionalities. To the best of our knowledge, there are no other web servers with the same integrated structural-analysis, modeling, and visualization capabilities.

## 2.2 Materials and methods

### 2.2.1 Base coordinate frames

The rigid-body parameters commonly used to characterize the 3D arrangements of the bases and base pairs in a nucleic-acid structure quantify the pairwise orientation and displacement of local, orthogonal reference frames embedded in the constituent nucleotides. The set of parameters and the coordinate frames used in the 3DNA software follow established, community-developed guidelines [11, 12]. The software performs a least-squares fitting of a standard planar base structure with an embedded coordinate frame on its experimental counterpart, following the approach of Babcock et al. [13, 14], to place the requisite reference frames on the bases in a structure.

### 2.2.2 Base-pair identification

The identification of interacting residues is based on the computed spatial disposition of the bases, in particular: (i) the distance  $d$  between the origins of the reference frames embedded in pairs of bases; (ii) the magnitude of the vertical offset of the base planes, the so-called Stagger (see text below); (iii) the angle  $\Lambda$  between the normals of the base planes; (iv) the distance  $d_{\text{N1-N9}}$  between the glycosidic base atoms, i.e. the purine N9 and pyrimidine N1 atoms linked to the sugar-phosphate backbone; and (v) the presence of one or more pairs of nitrogen/oxygen base atoms within a ‘hydrogen-bonding’ distance  $d_{\text{HB}}$ . The default values employed in the webserver calculations —  $d \leq 15\text{\AA}$ ; Stagger  $\leq 1.5\text{\AA}$ ;  $\Lambda \leq 30^\circ$  or  $\geq 150^\circ$ ;  $d_{\text{N1-N9}} \geq 4.5\text{\AA}$ ; and  $d_{\text{HB}} \geq 5.5\text{\AA}$

— identify both canonical and non-canonical nucleotides interactions [2, 3, 15], e.g. Watson-Crick [16], Hoogsteen [17], and other base pairs.

### 2.2.3 Rigid-body parameters

The w3DNA server reports three sets of rigid-body parameters: (i) the six base-pair parameters describing the spatial arrangements of associated bases — three angles called Buckle, Propeller, and Opening and three displacements called Shear, Stretch, and Stagger; (ii) the six base-pair-step parameters specifying the configurations of spatially adjacent base pairs — two bending angles called Tilt and Roll, the dimeric rotation angle Twist, two in-plane dislocations termed Shift and Slide, and the vertical displacement Rise; and (iii) the six parameters that relate the positions of successive base pairs relative to a local helical frame — the angles Inclination and Tip and the distances  $x$ -displacement and  $y$ -displacement describing the orientation and translation of the base planes with respect to the helical axis, and the rotation about and displacement along the helical axis, referred to as Helical Twist and Helical Rise [2, 11]. The numerical values describe the deviations of the base pairs in a given structure from the planar Watson-Crick base pairs in an ideal B-DNA helix, where the base-pair parameters, the dimeric bending components, and in-plane dislocations of adjacent base pairs are null [12]. A fourth set of rigid-body variables — the dinucleotide Tilt, Roll, Twist, Shift, Slide, and Rise — specifies the arrangements of adjacent bases along individual strands. The computations of rigid-body parameters use the mathematical definitions of El Hassan and Calladine [6]. The identification of the helical axis between adjacent base pairs follows the methodology introduced by Babcock et al. [14].

The reported output also includes the areas of overlap of adjacent bases and base pairs and the positioning of phosphorus atoms within each base-pair step. The former values quantify the stacking of neighboring base pairs, and the latter discriminate between A and B double-helical steps [18]. The base-pairing information is complemented by more conventional structural data, such as the identities and lengths of hydrogen bonds, the distances and angles between atoms in hydrogen-bonded and adjacent nucleotides, the torsion angles along the chain backbone, the amplitude and phase angle of sugar

pseudorotation (i.e. puckering geometry), the glycosyl torsions orienting the sugars and bases, and the widths of the major and minor grooves.

## 2.3 Webserver

### 2.3.1 Analysis component

The analysis component (Figure 2.1) of the w3DNA server determines the aforementioned conformational parameters for the paired bases, stacked base pairs, and sequential bases in a user-uploaded, PDB-formatted coordinate file, i.e. the standard listing of chemical information and atomic positions reported for the atoms in a structure (see the RCSB PDB website for a detailed description). The input of a PDB/NDB ID, i.e. the identifiers used respectively in the Protein Data Bank and the Nucleic Acid Database to denote individual structures, yields the same information. A simple keyword/author search and pop-up links to the PDB and NDB search engines and to the NDB Atlas facilitate the selection of archived structures.

**Output page.** The output page, illustrated in Figure 2.2 for the structure, deduced from multidimensional heteronuclear NMR spectroscopic studies, of a 13 base-pair DNA duplex bound to the human TTAGGG-repeat binding factor TRF1 (PDB ID: 1IV6) [19], contains four sections: (i) a brief summary of the structure; (ii) a schematic representation of the 3D fold; (iii) a link to the complete listing of 3DNA-derived parameters; and (iv) a set of interactive tables for selected parameters.

**Structural summary.** The structural summary includes the PDB ID and the NDB ID (if any), the methodology used to determine the structure, the resolution (if an X-ray structure), the deposition date (if the structural file is curated in the PDB or NDB), the author(s), the name of the compounds that make up the structure, and the links to several useful websites. If the input coordinate file contains more than one model, the summary also lists the number of models in the file, 20 in the case of the TRF1-DNA complex presented in Figure 2.2, and provides a link that gives the user the option to analyze multiple models.

**Structural representation(s).** The structural representation on the output page is

a composite image, with color-coded ‘blocks’ superimposed on the bases, an atomic depiction of backbone atoms, and color-coded tubes connecting the phosphorus atoms along individual strands. Proteins, if present, are represented by violet ribbons, and small molecules by ball-and-stick images. The same 3DNA-generated representations are found on the PDB and NDB websites. Each illustrated structure is automatically projected in the plane containing the two longest principal axes of the nucleic-acid fragment, but can be viewed from different viewpoints as described below. These and all other molecular images generated on the webserver can be saved by clicking on the appropriate download link.

Files, like 1IV6, with multiple models include a large gallery of small image icons depicting up to 50 structures in the file. Moving the mouse across different icons reveals the structural differences among the models. The location of the mouse determines the model that is enlarged on the output page. Clicking the icon generates the complete output for the selected structure. Icons of the same style allow the user to reorient and view the one model offered for most X-ray crystal structures in different principal-axes planes.

**Derived parameters and interactive tables.** The listing of derived parameters in the output file can be viewed on the web or downloaded. The 3DNA user’s manual, found at <http://3dna.rutgers.edu>, includes a brief description of each type of parameter. The parameter tables contain information about base sequence, interactions, and structure. Users can click each link to show/hide contents. The composition of base pairs and the rigid-body parameters relating sequential and paired bases and neighboring base pairs are presented in interactive, Grid-View tables, with angles expressed in degrees and distances in Ångstrom units. Data can be sorted by pressing an arrow at the top of each column. A simple quick search facilitates the examination of long nucleotide fragments. The example in Figure 2.2 shows the information included in the table of local base-pair step parameters — the numerical identities and chemical composition of the first 10 of the 12 base-pair steps TRF1-bound DNA, the rigid-body parameters describing each step, and the tetrameric sequence context in which the step occurs. The user can control the number of steps that are displayed, with up to 100

entries per page.

### 2.3.2 Reconstruction component

The reconstruction component (Figure 2.3) allows the user to build three-dimensional models of arbitrary sequence and helical type, including: (i) 55 different fiber-diffraction models of regular DNA, RNA, and hybrid DNA/RNA helices; (ii) customized single- and double-stranded structures with bases, base pairs, and base-pair steps arranged according to user-supplied rigid-body parameters; (iii) curved DNA structures constructed from fragments of canonical A-, B-, and C-type helices; and (iv) models of DNA ‘decorated’ at user-defined locations with proteins and other molecules in the arrangements found in known NMR and crystal complexes. The various structures provide useful starting points for atomic-level calculations.

**Fiber-diffraction models.** The 55 helical models include single-, double-, and multi-stranded structures based on the fiber-diffraction studies of Arnott and co-workers [20, 21] (43 models), Alexeev et al. [22] (two models), van Dam and Levitt [23] (two models), and Premilat and Albiser [24, 25, 26, 27, 28, 29] (eight models). Model choices are listed on a pull-down menu and described more fully in a table provided in the user tutorial. The models fall into two categories: generic helices that accommodate arbitrary base sequences of any length and non-generic helices that allow only the repetition of a pre-defined sequence. The example presented in 2.4(a) is a non-generic, triple-helical RNA complex made up of two 100-nt fragments of poly rU and a fragment of poly rA of the same length, held in place by Watson-Crick [16] and Hoogsteen [17] A·U pairing. The collection of fiber models includes 39 DNA double- or triple-helical structures, 12 RNA single-, double-, triple-, or quadruple-helical structures, and 4 DNA-RNA hybrid duplexes

**Customized models.** The input files of sequence information and rigid-body parameters needed to generate customized nucleic-acid models are of two types, depending on the nature of the desired structure. The construction of a folded, single-stranded structure, such as one adopted by an RNA molecule, requires the set of base step parameters describing the spatial disposition of successive nucleotides. Building a double-stranded



structure entails detailed specification of both the base-pair parameters between interacting nucleotides and the base-pair-step parameters between stacked pairs. Details of the necessary format are found on the tutorial page. The user selects the desired model type — either a full atomic model with an approximate, rigidly attached backbone or a model containing only base and P atoms, both in PDB format — from a pull-down menu.

The curved DNA pathways formed by the concatenation of regular A-, B-, and C-type models depend upon the chosen length, helical composition, and spacing of the structural components. For example, the slight zig-zag of the DNA duplex in Figure 2.4(b) reflects the opposing directions of dimeric bending and dislocation (Roll and Slide) in the A- and C-DNA fragments on either side of the central 35 base-pair stretch of B DNA. The all-atom backbones introduced in the model mirror the choice of helical types. The constructs accommodate any base sequence. The sequence within each conformational segment can be specified in two ways, as an arbitrary string of bases or as a string of repeated base-paired units.

**Ligand-decorated DNA.** The construction of protein- or ligand-decorated DNA models, such as the HU-bound DNA in Figure 2.4(c), entails specification of the DNA chain length and sequence, the number of bound species, the locations at which the molecules are bound, and the requisite protein- or ligand-bound DNA structural templates, such as the crystal complex of DNA with *Anabaena* HU [30] (PDB ID: 1P71) shown here. The bound fragments adopt the conformational parameters of the selected complexes, specified by a PDB or NDB ID or uploaded as a customized PDB-formatted structure. The unbound DNA, including rigidly attached backbone atoms, assumes the user-selected helical form (A, B, or C DNA). The binding positions correspond to the locations along the DNA of the central base pair or base-pair step of the chosen ligand-bound DNA structures. The location of the center point depends upon the length of the bound duplex, namely the middle base pair of a bound fragment with an odd number of bases pairs and the central base-pair step of a fragment with an even number of pairs. The software checks the user request for the potential overlap of proteins/ligands on the selected sequence and returns an error message if the proposed binding sites cover the

same base pair(s). Only double-stranded structures can be treated and only all-atom models are generated. The DNA is built in two stages, with bases first positioned in accordance with the rigid-body parameters of the protein-bound and free chain segments and the atoms of the sugar-phosphate backbone and associated ligands subsequently superimposed on the base framework. The resulting models reveal the interdependence of the bound species, chosen sites of binding, unbound DNA conformation, and overall macromolecular fold. For example, the undertwisted HU-bound DNA binding sites must be spaced at non-integral helical turns along B-form DNA to generate a planar, zig-zag pathway like the one shown in Figure 2.4(c).

**Output features.** All molecular constructs share the same three output features: (i) a composite representation of the overall structure in the above described principal-axis frame; (ii) a coordinate file in PDB format, which can be downloaded for further study; and (iii) a link to visualize the final structure via WebMol [31] (best done on a computer running a Java Runtime Environment) or Jmol (<http://www.jmol.org/>). In the interest of computational efficiency, models are limited in size to 1000 base pairs or 2000 nucleotides

### 2.3.3 Visualization component

The visualization component (Figure 2.5) creates vector-based drawings and scenes that can be rendered as raster-graphics images, allowing for easy generation of publication-quality figures. The server takes a user-uploaded PDB-formatted file or a PDB/NDB ID, and returns novel representations of the structure or parts of it. The images include: (i) composite block/tube/backbone representations of the type used to illustrate nucleosomal DNA [32] (PDB ID: 1KX5) in Figure 2.7(a); (ii) stacking diagrams of associated base pairs like that shown for neighboring C·G and A·U pairs in Figure 2.7(b); and (iii) composite block/ribbon/backbone representations of structural ensembles, such as the NMR-based models of the 5S RNA-TFIIA complex [33] (PDB ID: 2HGH) depicted in Figure 2.7(c).

**Composite images.** The composite images include informative color-coding of the nucleic acid. The user can choose the parts of the structure to be plotted, such as the

nucleic-acid atoms in the nucleosome complex in Figure 2.7(a) or the protein ribbons, and can rotate the structure as a whole by arbitrary amounts about one of the principal axes of the nucleic-acid structure. The axes — designated  $x$ ,  $y$ , and  $z$  — correspond respectively to the directions of the longest, intermediate, and shortest principal axes of the system.

**Stacking diagrams.** The stacking diagrams depict the hydrogen bonds between paired bases and reveal the overlap and relative disposition of stacked bases. The associated base pairs are automatically oriented in a top-down view such that the long axis of the step is horizontal and the leading strand lies on the left of the image, i.e. the average (middle) base-pair plane of the step coincides with the plane on which the structure is projected. The software identifies all stacked base pairs in the file and provides a list of the identified steps. The user specifies the step of interest and whether the bases should be labeled, as in Figure 2.7(b), or unlabeled in the diagram.

**Ensemble visualization.** The ensemble-visualization tool generates a composite block/ribbon/backbone image of a user-selected set of models in a file with multiple NMR-based or computer-generated structures, such as the 15 coordinate files of 5S RNA-TFIID structures depicted in Figure 2.7(c). The function returns error messages if applied to structures with a single model. The user selects the starting and ending models from a supplied list of model numbers.

### 2.3.4 Tutorial

The tutorial includes step-by-step instructions and worked-out examples to help the user take advantage of the available functions of the w3DNA server. The information pages address each of the functional categories i.e., analysis, reconstruction, and visualization. The server also provides links to the 3DNA Forum, a website where users pose and respond to assorted questions dealing with the use and application of the software, and to additional citations for users interested in learning more about (i) the content and capabilities of the software, (ii) the standard coordinate frame used in the determination of rigid-body parameters, (iii) the conformational parameter typical of nucleic-acid structures, and (iv) the differences among programs used in the analysis of nucleic-acid

structures.

## 2.4 Technical Details

### 2.4.1 Structure Analysis

The ‘Analysis’ component performs the analysis of nucleic-acid structures using three utilities of 3DNA: (i) ‘find\_pair’, (ii) ‘analyze’, and (iii) ‘blocview’. The function ‘find\_pair’ offers versatile options to locate bases, find possible base pairs, and identify reference frames and the helical regions of base pairs, given a PDB data file. The ‘analyze’ function calculates various nucleic-acid conformational parameters that are presented on the output page of the ‘Analysis’ component. The nucleic-acid structural parameter summary file on the web is a direct outcome of the 3DNA ‘analyze’ function. The visualization function ‘blocview’ generates a schematic image with a base block representation. This function is used at multiple points on the w3DNA webserver, as described below. More details about using these functions are found in the corresponding documentation of the 3DNA software.

**Structural Information Database.** To facilitate data processing of the w3DNA server, a background database was developed. The database contains a variety of tables which include: (1) PDB and NDB identifiers of all available nucleic-acid-containing structures deposited in the NDB and PDB uploaded to date; (2) basic descriptions of each structures, including authors, compounds, resolution, and deposit date; (3) nucleic-acid sequences of structures obtained directly from their pdb-formatted coordinate files; (4) base, base-pair, and base-pair step parameters of the nucleic-acid part of structures, which are pre-calculated with the 3DNA software. A set of programs has been written to automatically download structural coordinate files from the NDB and PDB, perform all calculations of conformational parameters, and load data into the corresponding database tables. The database, developed using the MySQL package, is free to academic users. Besides the database, a separated file folder, which contains various output files associated with the structure, such as block-representation images, conformational parameter files, etc., was created for each structure.

**Flexigrid Data Tables.** On the output page of the ‘Analysis’ component, nucleic-acid local structural parameters (base-step parameters, base-pair parameters, and base-pair-step parameters) are presented in tables using the framework of Flexigrid [34]. The Flexigrid is an open-source javascript-based template for querying and presenting data in a grid view with resizable columns and scrolling data to match the headers, plus the capability to connect to an xml-based data source. It contains features, including but not limited to, sortable column headers, paging, searches, and so on. Every time, when a structure identifier is requested, the server queries corresponding data from the aforementioned database and loads them to the Flexigrid template. With this style of grid view, one can easily sort, page, and search data to locate the values of interest.

**Biological Unit Versus Asymmetric Unit.** There are more than 800 nucleic-acid-containing structures in the PDB having a biological unit different from the asymmetric unit, as of April 2009. The biological unit, also called biological molecule, is the macromolecule that has been shown or is believed to be functional. In contrast, the asymmetric unit is the smallest portion of a crystal structure. By applying the crystallographic symmetry, an asymmetric unit can be rotated, translated, and twisted with displacement to make up the entire crystal. Depending on the space group of the crystal, the biological unit can be one of the three forms: (a) one copy of the asymmetric unit (all of which are equal); (b) multiple copies of the asymmetric unit; and (c) a portion of the asymmetric unit [35]. In the PDB, the first two cases occur most frequently. If the biological unit is different from the asymmetric unit, normally the asymmetric unit is a part of the biological unit. This can be found by comparing coordinate files of the biological and asymmetric units, both of which can be downloaded from the NDB. One may be interested in examining both units of a structure if they are different. The w3DNA web interface provides options for users to browse conformational analyses for both units.

**NMR Analysis.** An NMR coordinate file often contains multiple models of the molecular structure. About one quarter (976) of the total nucleic-acid-containing structures (4047) in the PDB/NDB are based on NMR data as of April 2009. The background server includes an analysis of each model of an NMR structure as thorough as that of

a normal crystal structure, with every output feature of the ‘Analysis’ component and with a specific identifier for each NMR model. On the w3DNA interface, the user can request to view the conformational parameters of a particular model of an NMR structure by clicking the main block representation of the model. The associated parameter data specified by the identifier are then queried and loaded to the output page. One can also view and download output files of all models of an NMR-based structure at the same page, given that the total model number does not exceed 50; otherwise, only the first 50 models will be analyzed and presented. This cutoff is designed to optimize the usage of space and memory of the server on which w3DNA is deployed.

### 2.4.2 Model Reconstruction

The ‘Reconstruction’ component of the webserver takes input in the form of sequence and geometric information, and generates atomic coordinates based on the user’s requirements. All three subunits of the ‘Reconstruction’ component require the pre-processing of the user’s requests, by which the server can eliminate input errors, interpret web forms into computer logic, and organize requests in the formatting of files needed by the 3DNA software. Two 3DNA functions are used: (1) ‘fiber’ and (2) ‘rebuild’. The tool ‘fiber’ generates DNA/RNA molecules by duplicating base pairs or DNA/RNA segments extracted from the 55 fiber models obtained by various researchers, including the canonical A-, B-, C- and Z-DNA double helices, triple-helical DNA and RNA, etc. (details below). The tool ‘rebuild’ takes the formatted input — which can be (1) the base-step parameters associated with a standard DNA base sequence, or (2) the base-pair and base-pair-step parameters describing the geometry of the leading sequence of a double-stranded molecule — and translates the input into rotational and translational matrices, with which nucleic-acid bases and base pairs are arranged in space and Cartesian coordinates are assigned to each atom based on standard models. More details about these two 3DNA functions are found in the associated parts of the 3DNA user manual. The ‘blocview’ function is also used here to present the reconstructed molecule.

**Fiber models.** Fiber-model reconstruction is based on the repetition of an experimentally determined helical repeating unit. The user can select from 55 fiber models, with the three most popular ones (A-, B-, and C-form DNA) placed at the top of the list of options under the ‘Fiber model’ subunit of the ‘Reconstruction’ component. Table 2.1 lists the 55 fiber models along with their conformational family, repeating sequence, strands, literature reference, and so on. Details about these 55 models can be found in the cited literatures.

**Protein Superposition.** The reconstruction of a protein-bound DNA model involves not only DNA generation but also superposition of protein atoms bound to the DNA. This cannot be done directly by any modules of the 3DNA software, which does not include protein superposition on the DNA. However, 3DNA offers a set of related tools that can significantly contribute to the development of the necessary modules for such reconstruction. Given a protein-DNA template structure, we first express the coordinates of the whole template onto the reference frame of the first base pair, which can be realized with the 3DNA programs ‘find\_pair’ and ‘frame\_mol’. Then, we extract the coordinates of the protein chains from the new coordinate file generated in the first step, by using the 3DNA program called ‘get\_part’ with the option ‘-p’, and save these data for later usage. With these preparatory steps, we then generate a whole-length DNA containing the user-provided sequence, with the input information such as the form of free DNA, the protein-DNA templates, and the binding sites. The 3DNA ‘rebuild’ function is used to obtain coordinates of this DNA. At this point, the generated DNA already contains the conformational distortion found at the protein-binding sites, because during the ‘rebuild’ process we have perturbed the bound DNA with the assumption that these parts of DNA would adopt the same base pair and base-pair-step parameters as in the protein-DNA templates upon the binding of proteins. For the completion of the protein-bound-DNA model, the protein coordinates have to be inserted into the same file as the DNA and placed appropriately at the binding sites. To this end, for every binding site, we do following work:

- (i) Locate the first base pair of the binding site. This base pair should have been

aligned with respect to the first base pair of the original corresponding protein-DNA template structure.

- (ii) Re-orientate the coordinates of the intermediate structure, which includes the DNA and inserted proteins (the latter occurs only when the current process is not for the first protein insertion).
- (iii) Insert the coordinates of the bound protein into the PDB-formatted coordinate file of the intermediate structure. The coordinates of the binding protein are the ones found and saved in the preparation phase. This insertion is rationalized by the fact that all coordinates have been expressed in the reference frame of the ‘first base pair’ identified in the first step.
- (iv) Iterate the above steps to have a final coordinate file that includes DNA and all proteins.

### 2.4.3 Molecular Visualization

The ‘Visualization’ component of the webserver utilizes three view tools from 3DNA: (i) ‘blocview’; (ii) ‘stack2img’; and (iii) ‘nmr\_ensemble’. The function ‘blocview’ has been described previously and is used here to generate a block representation of a part or the entire nucleic-acid-containing structure. In contrast to the block representation in the ‘Analysis’ component, the user is allowed to choose different options of the ‘blocview’ function here, such as the view angles. The function ‘stack2img’ generates a stacking-diagram image of a base pair with hydrogen bonds, filled base rings, and labels. Associated programs are written here to extract all base-pair steps of a structure and allow the user to choose a particular step for drawing a stacking diagram. The function ‘nmr\_ensemble’ generates a schematic image of an ensemble of structural models, each of which is displayed in block representation. Although it is primarily designed for an NMR ensemble, the ‘nmr\_ensemble’ feature also works for any ensemble of structures, such as a small molecular trajectory generated by molecular-dynamics simulations. Every inputted PDB file containing an ensemble of models is first analyzed by the server, which identifies the number of models in the ensemble and allows the user the



select the range of models for visualization. More details about these 3DNA functions can be found in the 3DNA publications [2, 3].

## 2.5 Concluding Remarks


The w3DNA server provides straightforward access to some of the most popular features of the 3DNA suite of programs. The server integrates various 3DNA utilities to carry out the pre- and post-processing of data necessary for the analysis and presentation of nucleic-acid structural information.

Other new subroutines working in the background allow the user to search for and manipulate input files, analyze structural data, generate the coordinates of molecular models, display assorted images, and manipulate tables on the fly. The various components make direct use of commands within w3DNA through graphic input options. The model reconstruction tools include new software for structure superposition and interactive visualization from multiple perspectives.

The server is intended for a broad range of users and educational purposes. Advanced users are encouraged to download the software package from the 3DNA web site and explore more of its functions.

Web 3DNA for analysis, reconstruction and visualization of nucleic-acid structures

[Analysis](#) [Reconstruction](#) [Visualization](#) Monday, May 18, 2009 04:08



Enter a PDB/NDB ID | [Upload a PDB file](#) | [Search PDB/NDB](#)

[\[Help\]](#)

[ex1: 3EXJ]  
[ex2: 1HRZ]

---

[About](#) [Tutorial](#) [Download](#) [Forum](#) [Contact](#) [Citation](#)

---

2008-2009 ©Dr. Wilma K. Olson Group, Rutgers, the State University of New Jersey  
New Brunswick, New Jersey, USA

Figure 2.1: The front page of the w3DNA analysis component. The user has the options to search for a structure in the PDB or NDB, analyze a DNA/RNA structure using its PDB or NDB identifier, or upload a customized pdb file to find the conformational parameters.

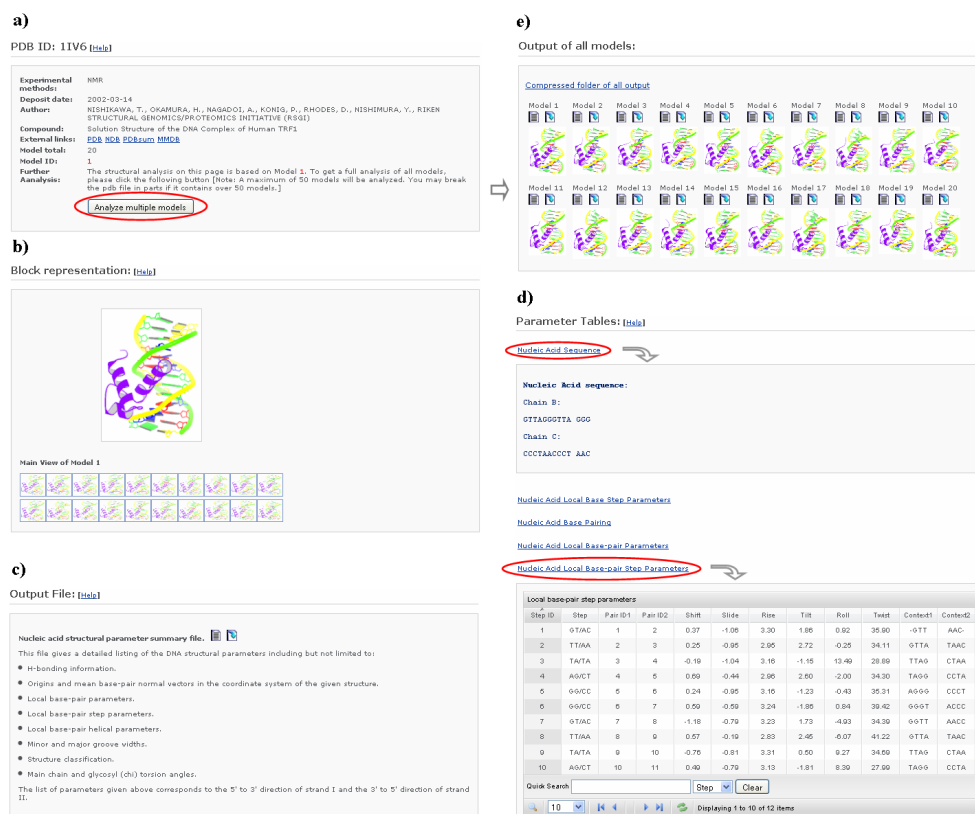


Figure 2.2: Screenshots illustrating the information provided in the analysis of nucleic-acid-containing structures with valid Protein Data Bank identifiers (PDBID), here the ensemble of 20 structures of the complex of DNA with the human TTAGGG-repeat binding factor TRF1 determined by multidimensional heteronuclear NMR spectroscopy [19] (PDB ID: 1IV6). The output comprises, but is not limited to: (a) a brief description of the structural file, including the author(s), compound(s), number of models, external links, etc., (b) a gallery of block representations of each model in the file, which by moving the mouse over the icons, reveals the fluctuations in the structural ensemble and by clicking a specific icon, points to the set of parameters describing the chosen model, (c) a summary file with a comprehensive list of structural parameters, which can be viewed or downloaded by clicking the appropriate icons, (d) tables of selected parameters, which can be displayed by clicking on one of the links and sorted by clicking on the headers of the columns in the selected table, (e) the page, redirected from (a) via the Analyze multiple models button, with links to the summary files for all of the models of the protein-DNA complex.

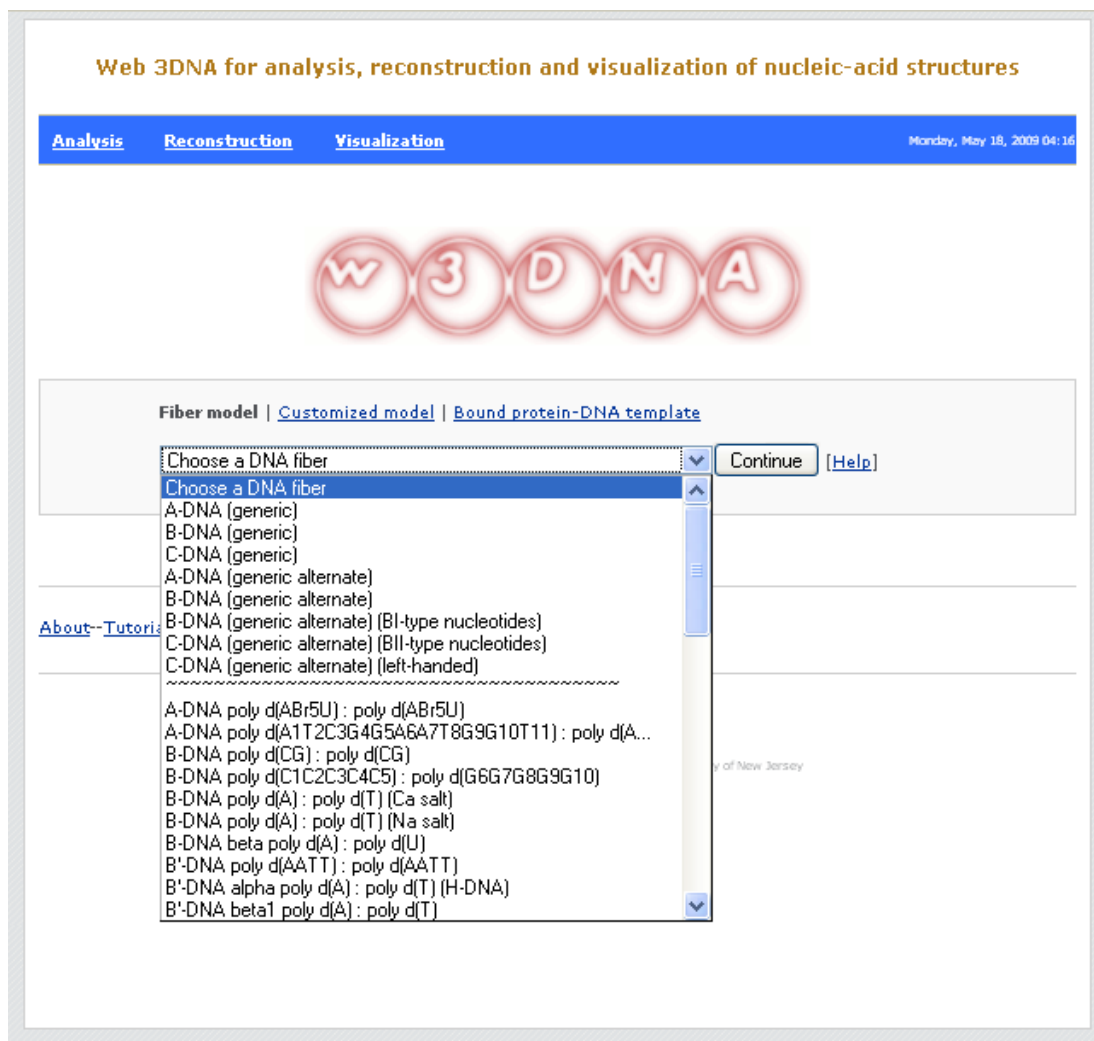


Figure 2.3: The front page of the w3DNA 'Reconstruction' component contains three subunits for different types of model building: (i) fiber model; (2) customized model; and (3) ligand-decorated DNA model. The drop-down options, displayed here, show the 55 different fiber-diffraction models that can be generated.

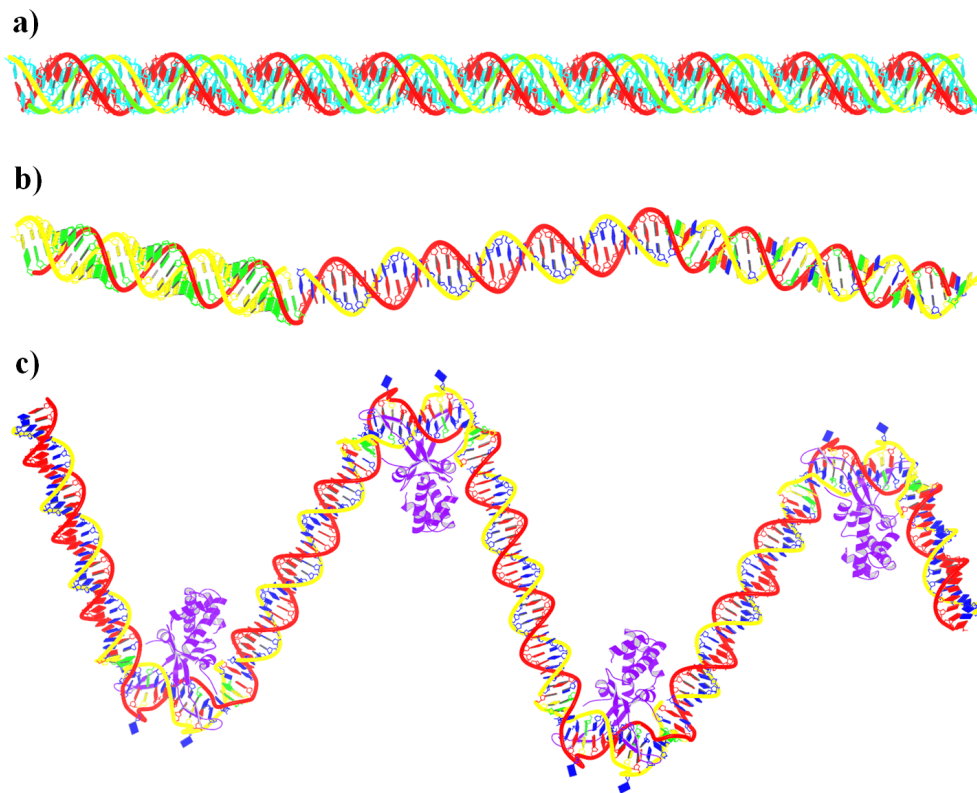


Figure 2.4: Representations of nucleic-acid-containing structures generated with the reconstruction component of the w3DNA server. (a) Two 100 nucleotide strands of poly(rU) complexed with the same length of poly(rA) in the classic 11-fold RNA poly(rU)·poly(rA)·poly(rU) triple helical structure [20, 21] (3DNA fiber model 32). (b) A 100 base-pair curved DNA block copolymer made up respectively of A-, B-, and C-form double-helical fragments (3DNA fiber models 1, 4, and 7) of  $G_{35} \cdot C_{35}$ ,  $A_{35} \cdot T_{35}$ , and  $(GA)_{15} \cdot (TC)_{15}$ . (c) A 210 base-pair B-form DNA ‘decorated’ with 4 HU proteins. The protein-bound steps — centered at base pairs 41, 89, 137, and 185 — are assigned the sequence and rigid-body parameters of the central 17 base pairs in the 1.90-Å crystal complex with *Anabaena* HU [30] (PDB ID: 1P71). The protein-free DNA steps are fixed in the canonical B form and assigned a homopolymeric repeating sequence  $(A_n \cdot T_n)$ , where  $n$  is respectively 31, 28, 28, 28, and 16 base pairs. Color-coded tubes on RNA and DNA trace the progression of the backbone defined by the phosphorus atoms: strand I (red); strand II (yellow); strand III (green). Color coding of nucleotide sequence conforms to the Nucleic Acid Database standard [10]: A (red); C (yellow); G (green); T (blue). Violet ribbons connecting protein  $C^\alpha$  carbons generated with MolScript (<http://www.avatar.se/molscript/>).

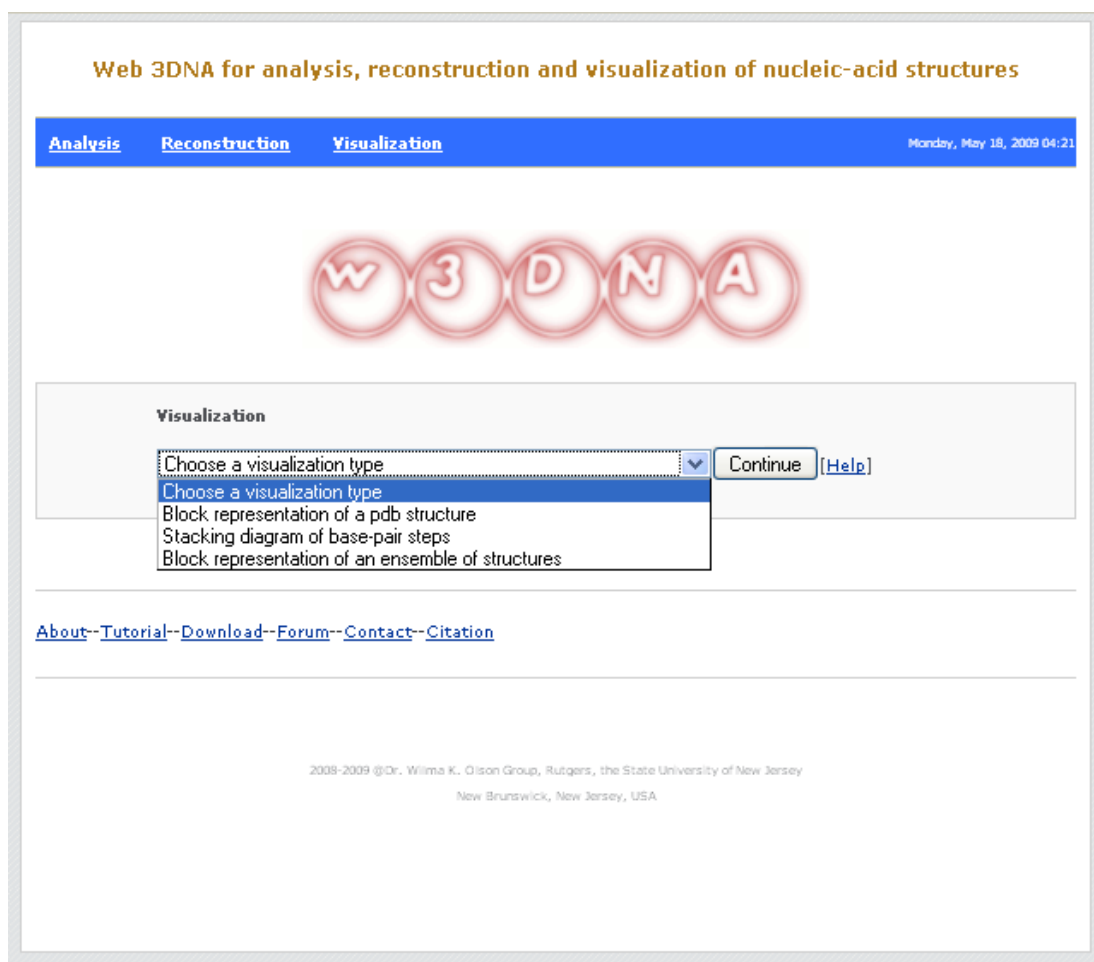



Figure 2.5: The front page of the w3DNA ‘Visualization’ component. The user has three options to use this component: (i) block representation of a nucleic-acid-containing structure; (ii) stacking diagram of base-pair steps; and (iii) block representation of an ensemble of nucleic-acid-containing structures.

**Web 3DNA for analysis, reconstruction and visualization of nucleic-acid structures**

---

[Analysis](#)   [Reconstruction](#)   [Visualization](#) Monday, May 18, 2009 04:24



**Current choice: 3 binding sites**

**Other choices:**

Choose a binding site number ▼ Continue

**Type in or paste a sequence:** [\[Help\]](#)

**Choose a form of DNA (unbounded)**

B-form ▼

[\[Help\]](#)

Binding position:	Template PDB ID:	--OR-- upload the template pdb file:
		Browse...
		Browse...
		Browse...

[PDB Search](#) (only simple double-stranded DNA containing structures can be used here)

☐ Preview with block representation (This option may slow down the reconstruction process.) [\[Help\]](#)

Continue

[About](#)--[Tutorial](#)--[Download](#)--[Forum](#)--[Contact](#)--[Citation](#)

---

2008-2009 © Dr. Wilma K. Olson Group, Rutgers, the State University of New Jersey  
New Brunswick, New Jersey, USA

Figure 2.6: An intermediate page of the ligand-decorated DNA reconstruction sub-component. The user is required to provide a DNA sequence, the form of the free DNA, the ligand binding positions, and the ligand-DNA structural templates.

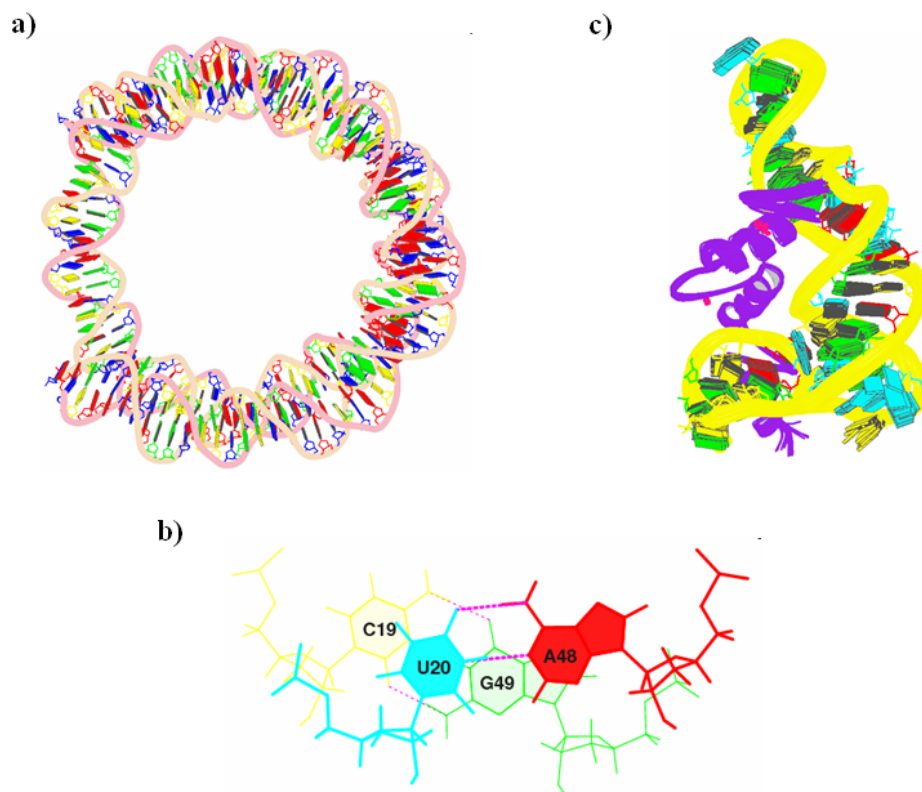


Figure 2.7: Examples of the unique representations of nucleic-acid structures available through the w3DNA server. (a) Color-coded composite block/backbone/tube representation of the DNA in the currently best-resolved nucleosome core-particle structure [32] (PDB ID: 1KX5). (b) Stacking diagram illustrating the overlap and hydrogen-bonding patterns of adjacent base pairs (C<sub>19</sub>·G<sub>49</sub> and U<sub>20</sub>·A<sub>48</sub>) in model 1 from the ensemble of NMR structures of the complex of a 55 nucleotide fragment of *X.laevis* 5S rRNA with three zinc fingers of transcription factor TFIIEA [33] (PDBID: 2HGH). (c) Schematic 'NMR-ensemble' image of the bases, RNA backbones, and protein ribbons in 15 of the 20 models of the aforementioned RNA-protein complex. Color coding identical to that in Figure 2.4, save for the depiction of U in aqua.



Table 2.1: Table of regular DNA and RNA helical models

Family	Molecule	Repeating sequence	Notes	Base/turn	Strands	Reference
A	DNA	generic		2	11	[20, 21]
B	DNA	generic		2	10	[20, 21]
C	DNA	generic		2	9.3	[20, 21]
A	DNA	generic alternate		2	11	[24]
B	DNA	generic alternate		2	10	[24]
B	DNA	generic alternate	BI nucleotides	2	10	[23]
C	DNA	generic alternate	BII nucleotides	2	9	[23]
C	DNA	generic alternate	left-handed	2	9.3	[25]
A	DNA	ABr5UABr5U		2	11	[20, 21]
A	DNA	ATCGGAATGGTTAGCCTTACCA		2	11	[20, 21]
B	DNA	CGCG		2	10	[20, 21]
B	DNA	CCCCCGGGGG		2	10	[20, 21]
B	DNA	AT	Ca salt	2	10	[22]
B	DNA	AT	Na salt	2	10	[22]
B	DNA	AU	b	2	10	[20, 21]
B'	DNA	AATTAATT		2	10	[20, 21]
B'	DNA	AT	alpha H DNA	2	10	[20, 21]
B'	DNA	AT	beta1	2	10	[20, 21]
B'	DNA	AICT	beta1	2	10	[20, 21]
B'	DNA	AT	beta2 H DNA beta	2	10	[20, 21]
B'	DNA	AU	beta2	2	10	[20, 21]
B'	DNA	AICT	beta2	2	10	[20, 21]
B*	DNA	AT	high temperature	2	11.4	[27]
C	DNA	GGTACC		2	9	[20, 21]
C	DNA	GGTACC		2	9	[20, 21]
C	DNA	AGCT		2	9	[20, 21]
C	DNA	AGCT		2	9	[20, 21]
D	DNA	AATATT		2	8	[20, 21]
D	DNA	CICI		2	8	[20, 21]
D	DNA	ATATATATATAT		2	8	[20, 21]
D.A	DNA	ATAT		2	8.2	[29]
D.B	DNA	ATAT		2	8	[29]
L	DNA	GCGC		2		[20, 21]
S	DNA	GCGC	CBGA, right-handed	2	12	[28]
S	DNA	GCGC	CAGB, right-handed	2	12	[28]
Z	DNA	GCGC		2	12	[20, 21]
Z	DNA	As4TAs4T		2	14	[20, 21]
	DNA	CIC		3	11	[20, 21]
	DNA	TAT		3	12	[20, 21]
	DNARNA	AdT	hybrid	2	11	[20, 21]
	DNARNA	dGC	hybrid	2	11.25	[20, 21]
	DNARNA	dIC	hybrid	2	10	[20, 21]
	DNARNA	dAU	hybrid	2	11	[20, 21]
A	RNA	AU		2	11	[20, 21]
A	RNA	XX		2	11	[20, 21]
A	RNA	s2Us2U	symmetric	2	11	[20, 21]
A	RNA	s2Us2U	asymmetric	2	11	[20, 21]
A	RNA	IC		2	12	[20, 21]
	RNA	XX		2	10	[20, 21]
	RNA	UAU		3	11	[20, 21]
	RNA	UAU		3	11	[20, 21]
	RNA	UAU		3	12	[20, 21]
	RNA	III		4	11.5	[20, 21]
	RNA	eC (O2 ethyl)		1	6	[20, 21]

## References

- [1] Zheng, G., Lu, X. J., and Olson, W. K. (2009) Web 3DNAa web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.*, **37**(Web Server issue), W240–W246.
- [2] Lu, X. J. and Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
- [3] Lu, X. J. and Olson, W. K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protoc.*, **3**, 1213–1227.
- [4] Zhurkin, V. B., Lysov, Y. P., and Ivanov, V. I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, **6**, 1081–1096.
- [5] Bolshoy, A., McNamara, P., Harrington, R. E., and Trifonov, E. N. (1991) Curved DNA without AA: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci., U.S.A.*, **88**, 2312–2316.
- [6] Hassan, M. A. E. and Calladine, C. R. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA: a new local calculation scheme. *J. Mol. Biol.*, **251**, 648–664.
- [7] Lu, X. J., Hassan, M. A. E., and Hunter, C. A. (1997) Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *J. Mol. Biol.*, **273**, 668–680.
- [8] Lu, X. J., Hassan, M. A. E., and Hunter, C. A. (1997) Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP). *J. Mol. Biol.*, **273**, 681–691.
- [9] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- [10] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R., and Schneider, B. (1992) The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
- [11] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. H., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C., and Olson, W. K., et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 781–791.

- [12] Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, R. C., Heinemann, U., Lu, X. J., Neidle, S., and Shakked, Z., et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
- [13] Babcock, M. S. and Olson, W. K. (1994) The effect of mathematics and coordinate system on comparability and “dependencies” of nucleic acid structure parameters. *J. Mol. Biol.*, **237**, 98–124.
- [14] Babcock, M. S., Pednault, E. P., and Olson, W. K. (1994) Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J. Mol. Biol.*, **237**, 125–156.
- [15] Xin, Y. and Olson, W. K. (2009) BPS: a database of RNA base-pair structures. *Nucleic Acids Res.*, **37**, D83–D88.
- [16] Watson, J. D. and Crick, F. H. (1953) A structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
- [17] Hoogsteen, K. (1963) The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine.. *Acta Crystallogr.*, **16**, 907–916.
- [18] Lu, X. J., Shakked, Z., and Olson, W. K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
- [19] Nishikawa, T., Okamura, H., Nagadoi, A., Konig, P., Rhodes, D., and Nishimura, Y. (2001) Solution structure of a telomeric DNA complex of human TRF1. *Structure*, **9**, 1237–1251.
- [20] Chandrasekaran, R. and Arnott, S. (1989) The structures of DNA and RNA helices in oriented fibers. In W. Saenger, Editor, *Landolt-Bornstein Numerical Data and Functional Relationships in Science and Technology*, Group VII/1b, Nucleic Acids, Page 31-170, Springer-Verlag, Berlin.
- [21] Arnott, S. (1999) Polynucleotide secondary structures: an historical perspective. In S. Neidle, Editor, *Oxford Handbook of Nucleic Acid Structure*, Page 1-38, Oxford University Press, Oxford, UK.
- [22] Alexeev, D. G., Lipanov, A. A., and Skuratovskii, I. Y. (1987) The structure of poly(dA)·poly(dT) as revealed by an X-ray fibre diffraction. *J. Biomol. Struct. Dynam.*, **4**, 989–1011.
- [23] van Dam, L. and Levitt, M. H. (2000) BII nucleotides in the B and C forms of natural-sequence polymeric DNA: a new model for the C form of DNA with 40° helical twist. *J. Mol. Biol.*, **304**, 541–561.
- [24] Premilat, S. and Albiser, G. (1983) Conformations of A-DNA and B-DNA in agreement with fiber X-ray and infrared dichroism. *Nucleic Acids Res.*, **11**, 1897–1908.
- [25] Premilat, S. and Albiser, G. (1984) Conformations of C-DNA in agreement with fiber X-ray and infrared dichroism. *J. Biomol. Struct. Dynam.*, **2**, 607–613.

- [26] Premilat, S. and Albiser, G. (1986) DNA models for A, B, C and D conformations related to fiber X-ray, infrared and NMR measurements. *J. Biomol. Struct. Dynam.*, **3**, 1033–1043.
- [27] Premilat, S. and Albiser, G. (1997) X-ray fibre diffraction study of an elevated temperature structure of poly(dA)poly(dT). *J. Mol. Biol.*, **274**, 64–71.
- [28] Premilat, S. and Albiser, G. (1999) Helix-helix transitions in DNA: fibre X-ray study of the particular cases poly(dG-dC) and poly(dA) 2poly(dT). *Eur. Biophys. J.*, **28**, 574–582.
- [29] Premilat, S. and Albiser, G. (2001) A new D-DNA form of poly(dA-dT)-poly(dA-dT): an A-DNA type structure with reversed Hoogsteen pairing. *Eur. Biophys. J.*, **30**, 404–410.
- [30] Swinger, K. K., Lemberg, K. M., Zhang, Y., and Rice, P. A. (2003) Flexible DNA bending in HU-DNA cocrystal structures. *EMBO. J.*, **22**, 3749–3760.
- [31] Walther, D. (1997) WebMola Java-based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
- [32] Davey, C. A., Sargent, D. F., Luger, K., Mader, A. W., and Richmond, T. J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, **319**, 1097–1113.
- [33] Lee, B. M., Xu, J., Clarkson, B. K., Martinez-Yamout, M. A., Dyson, H. J., and Case, D. A. (2006) Induced fit and "lock and key" recognition of 5 S RNA by zinc fingers of transcription factor IIIA. *J. Mol. Biol.*, **357**, 275–291.
- [34] Marinas, P. P. (2009) Flexigrid for jQuery, <http://www.flexigrid.info>.
- [35] Introduction to biological units and the PDB archive, <http://www.pdb.org>.

## Chapter 3

### **3DNA Landscapes: a database for exploring the conformational features of DNA**

3DNA Landscapes, located at: <http://3DNAscapes.rutgers.edu>, is a new database for exploring the conformational features of DNA. In contrast to most structural databases, which archive the Cartesian coordinates and/or derived parameters and images for individual structures, 3DNA Landscapes enables searches of conformational information across multiple structures. The database contains a wide variety of structural parameters and molecular images, computed with the 3DNA software package and known to be useful for characterizing and understanding the sequence-dependent spatial arrangements of the DNA sugar-phosphate backbone, sugar-base side groups, base pairs, base-pair steps, groove structure, etc. The data comprise all DNA-containing structures — both free and bound to proteins, drugs and other ligands — currently available in the Protein Data Bank. The web interface allows the user to link, report, plot and analyze this information from numerous perspectives and thereby gain insight into DNA conformation, deformability and interactions in different sequence and structural contexts. The data accumulated from known, well-resolved DNA structures can serve as useful benchmarks for the analysis and simulation of new structures. The collective data can also help to understand how DNA deforms in response to proteins and other molecules and undergoes conformational rearrangements.

#### **3.1 Introduction**

In addition to the genetic message, DNA base sequence carries a multitude of structural and energetic signals related to its biological packaging and processing. These codes govern how the double-helical molecule deforms in response to proteins and other

ligands and when and where the genetic information is expressed. DNA is not just a passive substrate of cellular proteins but an active player with physical properties capable of influencing the three-dimensional organization of genetic sequences and the activity of regulatory proteins and processing enzymes. Understanding the pathways and capabilities of DNA deformation is thus crucial for deciphering the codes behind the regulation, organization and dynamics of various genomes. Acquiring this knowledge requires a systematic view of the structural landscapes accessible to DNA as it deforms in solution and adjusts to interactions with other molecules. This information, in turn, offers reliable benchmarks for predictions of nucleic-acid interactions and structures.

3DNALandscapes is a new database for exploring the conformational features of DNA. The database has been designed to study DNA backbone, side-group, base-pair, base-pair-step and complementary-strand geometry statistically, using information derived from multiple structures with the 3DNA software package [1, 2, 3] in combination with other currently available data resources, such as structural classifications and descriptions found in the Protein Data Bank (PDB) [4] and Nucleic Acid Database (NDB) [5]. We have also constructed a web interface to link, report, plot and analyze the structural parameters in the database. The main component of the web interface is a search function that enables the user to collect structural data and generate statistical reports on the fly.

The PDB and NDB contain a number of derived nucleic-acid conformational parameters, including the base-pair and base-pair-step parameters obtained with 3DNA. Although these databases include some of the information stored in 3DNALandscapes, not all of the information is contained in either of them. In addition, the PDB and NDB are designed to be structure-centric, meaning that data from a single structure are easy to obtain. Gathering data for a specific parameter or parameter set across multiple nucleic-acid structures is difficult or impossible with these interfaces. The collective information in 3DNALandscapes provides insights into the intrinsic sequence-dependent structure and deformability of DNA [6, 7] as well as useful benchmarks for the analysis and simulation of other DNA structures [8, 9, 10].

## 3.2 Database Content

The database is managed by a MySQL platform [11]. Data are stored in a rational schema that organizes tables of information in a hierarchical fashion. The highest level of the schema contains basic structural information, such as molecular classifications, sequences and resolution. The next level divides the data into five categories: backbone, sugar-base side-group, base-pairing, base-pair-step and complementary-strand information. The lowest level of the schema contains the derived parameters associated with the backbones, side groups, base pairs, base-pair steps and complementary-strand interactions.

### 3.2.1 Structures

The first release of the database contains derived information for all DNA-containing structures — both free and bound to proteins, drugs and other ligands — deposited in the Protein Data Bank as of October 2009. The composite data come from 6615 structural models, taken from the complete sets of atomic coordinates reported in 2084 X-ray crystallographic and 586 nuclear magnetic resonance (NMR) investigations. Among those structures, 1429 occur in complexes with proteins, 973 associate with drugs and other small molecules and 2004 contain only bound water or metal ions. The X-ray-based entries reflect the coordinates of the biological units rather than the asymmetric structural units. Individual models within the ensembles of NMR-derived structures contain unique internal identifiers assigned as the database is loaded.

The structures are classified in terms of the DNA conformational assignments made by the 3DNA software, e.g. fraction or number of base-pair steps in A and B double-helical forms. Individual entries also include the resolution (in the case of X-ray models), literature citations and other features stored in the original structural files. The DNA sequences and associated chain names and residue numbers are extracted in the 3DNA analysis for subsequent use in locating specific nucleotides, base pairs and base-pair steps in a given model. The data-collection procedure records the chemical composition and nucleotide surroundings of the base pairs and base-pair steps so that effects

of base modification and sequence context can be studied. That is, the base-pair and dimeric entries contain the identities of the base pairs that precede and follow the designated unit, thereby marking the relevant set of conformational data in the context of the trimer that contains the base pair and the tetramer than contains the dimer step. The annotation thus takes account of the base pairs and base-pair steps at the ends of helices.

### 3.2.2 Backbones

Features of the DNA chemical framework stored in the database include the standard set of internal torsional parameters associated with the nucleotide units along individual strands [12] and related intrastrand distances. These quantities include the five acyclic torsion angles —  $\alpha$  (O3'-P-O5'-C5'),  $\beta$  (P-O5'-C5'-C4'),  $\gamma$  (O5'-C5'-C4'-C3'),  $\delta$  (C5'-C4'-C3'-O3'),  $\epsilon$  (C4'-C3'-O3'-P),  $\zeta$  (C3'-O3'-P-O5') — along the sugar-phosphate backbone and the distances  $d_{P-P}$  between phosphorus atoms on successive nucleotides. The distances are expressed in Ångstrom units and the angles are assigned values over the range  $(-180^\circ, +180^\circ)$ .

### 3.2.3 Sugar-base side groups

Description of the spatial arrangements of the sugar and base units follows conventional guidelines (12). The stored conformational data include: (i) the glycosyl torsion angle  $\chi$  (O4'-C1'-N9-C4) or  $\chi$  (O4'-C1'-N1-C2), respectively, describing the orientation of a purine (R) or pyrimidine (Y) with respect to the sugar ring; (ii) the five internal sugar-ring torsion angles —  $\nu_0$  (C4'-O4'-C1'-C2'),  $\nu_1$  (O4'-C1'-C2'-C3'),  $\nu_2$  (C1'-C2'-C3'-C4'),  $\nu_3$  (C2'-C3'-C4'-O4') and  $\nu_4$  (C3'-C4'-O4'-C1'); and (iii) the phase angle  $P$  and amplitude  $\tau_{max}$  of sugar pseudorotation derived from the latter quantities [13].

### 3.2.4 Base pairs

The 3DNA analysis identifies 91280 hydrogen-bonded base pairs — 70120 canonical (Watson-Crick) pairs and 21160 noncanonical pairs — in the above set of structures. The Watson-Crick pairs include all A·T and G·C associations with the requisite



hydrogen-bond (H-bond) patterns. All other base pairs, including partially distorted Watson-Crick pairs with missing H bonds, are classified as noncanonical. Structures with three or more strands include the close base-base associations of all interacting strands. The accepted base pairs meet simple geometric criteria [14] and contain two or more H bonds, at least one of which involves a proton donor-acceptor interaction between nitrogens or oxygens on the two bases.

The spatial disposition of the bases in each pair is described by three types of data: (i) the identities and lengths of the H bonds; (ii) the six rigid-body parameters that relate local coordinate frames embedded on the interacting bases; and (iii) the virtual distances and angles between selected atoms on the bases and attached sugars. The set of H bonds includes the interactions between the flagged bases as well as those with the sugar-phosphate backbone and the bifurcated (three-center) H-bonds between contacted residues. The base-pair parameters — three angles called Buckle, Propeller and Opening and three distances called Shear, Stretch and Stagger [15] — follow the matrix-based definitions originated by Zhurkin et al. [16] and described in detail by El Hassan and Calladine [17]. The virtual parameters include the distances  $d_{C1' \cdots C1'}$  between the C1' atoms attached to paired bases and the angles  $\lambda_R$  and  $\lambda_Y$  formed by the C1'  $\cdots$  C1' line with the R(C1'-N9) and Y(C1'-N1) glycosidic bonds, respectively.

### 3.2.5 Base-pair steps

Structural characterization of the 66549 base-pair steps formed by sequential base pairs includes: (i) the six rigid-body parameters specifying the orientation and displacement of the constituent base pairs; (ii) the six local helical parameters relating the positions of the base pairs; (iii) the area of overlap of the stacked base pairs; (iv) the displacement of the phosphorus atoms on interacting strands along the local dimeric and helical coordinate frames; (v) the distances between the C1' atoms in the dimeric unit; and (vi) the conformational families to which the steps belongs. The coordinate frames on the bases, base pairs and base-pair steps follow established conventions [18]. The six base-pair-step parameters — three rotations (Tilt, Roll, Twist) and three translations (Shift, Slide, Rise) [15] — are analogs of the six base-pair parameters [16, 17]. The six local

helical parameters — Inclination, Tip, Helical Twist,  $x$ -displacement,  $y$ -displacement and Helical Rise [18] — are defined, following Babcock et al. [19], in terms of the single rotational operation that brings the coordinate frames on the base pairs into alignment. The base-pair overlap is the area shared by the four polygons formed by projecting the ring atoms of the bases on the mean base-pair plane [1]. The stored data include the contributions to the overlap from the bases on the same and opposing strands and the corresponding values obtained for larger polygons constructed from the ring and exocyclic base atoms. The projections of the P atoms ( $x_P$ ,  $y_P$ ,  $z_P$ ) along the coordinate axes of the dimeric step distinguish A- from B-type DNA [7] as well as potential intermediate AB steps along the A→B conformational pathway [10]. The corresponding projections along the axes of the local helical frame [ $x_P(h)$ ,  $y_P(h)$ ,  $z_P(h)$ ] distinguish the TA-like steps [1], i.e. the conformational form of DNA [20] found in complexes with the TATA-box protein and other proteins. The  $z_P$  and  $z_P(h)$  values are used to determine the conformational family of the dimer steps. The intrastrand  $C1' \cdots C1'$  distances also distinguish different conformational types.

### 3.2.6 Complementary-strand interactions

Finally, the conformational data include the widths of the major and minor grooves, i.e. the long-range distances between phosphorus atoms on interacting strands that expose the respective non-H-bonded edges of Watson-Crick base pairs. The recorded values are based on the direct and refined formulations of El Hassan and Calladine [21] and are assigned to the relevant base-pair step. The direct values correspond to the distances between  $P_i$ , the phosphorus atom on the leading strand of base-pair step  $i$ , and specific phosphorus atoms on the other strand,  $P_{i-3}$  across the minor groove and  $P_{i+4}$  across the major groove, typically the shortest cross-strand  $P \cdots P$  distances in B-DNA helices. The refined values allow for the variation in helical structure that alters the identities of the atoms in closest cross-strand contact. Thus, the two measures of groove width may differ markedly if the helix undergoes large distortions.

### 3.3 Web Interface

The web interface, located at: <http://3DNAsclapes.rutgers.edu> and constructed in the CodeIgniter PHP web application framework [22], parallels the organization of the database. The software contains three major components: a structure filter; a series of parameter- and context-selection panels; and a data report. The tabulated data also include links to a local summary and visualization page for each of the structural fragments from which the listed quantities are extracted. The user must first specify a set of structures in the structure filter, then select the type of structural information to be considered and finally view the summaries of the analysis in the statistical reporter.

#### 3.3.1 Structure filter

The structure-filter page offers two options for the user to define a set of structures. First, one can make selections based on a combination of the following features: the experimental method used to determine the structure; the molecular contents; the resolution cutoff; and the conformational characteristics of the constituent base-pair steps. By specifying the experimental method, the user can examine structures obtained by X-ray, NMR or both approaches. The choice of molecular contents refers to the other molecules present in the experimental structure: proteins; drugs or other small molecules; bound water; metal ions. The conformational option allows the user to select structures with given fractions or numbers of base-pair steps that have local conformational features characteristic of A-, B-, AB-, TA- or Z-type helices. That is, the structure-filtering algorithm uses the values of various parameters, determined with the 3DNA software, to characterize individual base-pair steps in a given structure rather than group the structure as a whole in terms of its global appearance. Thus, A-type base-pair steps might occur in what appears at the global level to be a B-DNA duplex and vice versa. This information is useful in understanding how ligands induce local conformational changes in DNA or how large-scale reorganization of structure preserves fundamental local structural propensities. The resolution cutoff affects only the collection of X-ray structures. The NMR structures in the database have an arbitrarily

assigned resolution of zero, which will lie always within the cutoff limit.

The second option lets the user enter a list of PDB or NDB structural identifiers (IDs), which the server checks for accuracy. This option allows the user to perform searches elsewhere, such as the integrated search at the NDB or the advanced search at the PDB, and then import the findings into the 3DNA Landscapes interface for conformational analysis. After clicking ‘next’, a list of structures with brief descriptions is displayed in a table with sorting and paging capabilities. The user can edit the structures generated in the automated search by denoting the PDB identifiers of the files to be removed or added.

Finally, the user has the option in either selection process of choosing a representative structure (the first structure) or the complete ensemble of structures associated with the NMR-based files. It worth noting that the selection of ensembles can lead to time delays in the analysis and visualization of large quantities of data and also may bias the statistical results. The choice can be useful, however, if the user is interested in the conformational trends associated with the DNA included in a single NMR structure file.

### **3.3.2 Parameter- and context-selection panels**

The parameter- and context-selection panels allow the user to choose the conformational parameters of interest and the nucleotide units that meet certain conditions within the set of selected structures. The parameter list includes the aforementioned quantities associated with the DNA backbones, sugar-base side groups, base pairs, base-pair steps and complementary strands. The set of conditions includes the chemical context, sequence context and conformational category.

Thus, the user can specify whether or not to include nucleotides containing modified bases or those found in non-canonical base pairs. One can also select the identities of the bases that flank particular chemical moieties, such as the base pairs that precede and follow a base pair or base-pair step. Only parameters associated with the specified chemical unit in the given sequential context are retained. This option allows the user to study the effects of neighboring base pairs on the local conformation of DNA. The

user can also narrow the search by specifying the conformational character of base-pair steps. This action restricts the selection of parameters to the backbones and base pairs that constitute the base-pair steps of a particular conformational type, e.g. only A-DNA steps (as opposed to the structures with a given proportion or number of A-like steps, which can be chosen with the Structure Filter).

### 3.3.3 Data report

The data report contains a table of the selected conformational data, a gallery of plotted images and a brief statistical report. The grid-view table at the top of the report lists all entries for the chosen parameters and contains hyperlinks, which direct the user to the local summary and visualization pages described below. The information in the table can be sorted by column entries and exported as a data file. The graphical gallery includes histograms and, in some cases, scatter plots of the distribution of the collected data. The histograms (Figure 3.1) illustrate the information included in individual columns, while the scatter plots (Figure 3.2) reveal the pairwise correlations of selected parameters, such as the coupling of bending and twisting in DNA base-pair steps (via Roll and Twist) [9]. The scatter plots also include ellipses, derived from the covariance, that encircle most of the plotted data [6]. Related parameters are plotted on a common scale for ease of comparison, and all images can be downloaded. The statistical report includes the number of examples, average values, minima and maxima for the data associated with the chosen sequences, such as the rigid-body parameters of specific base pairs or base-pair steps (Figure 3.1). The report also includes the option to determine the statistics for the chosen parameters in different trimeric or tetrameric sequence contexts. The analyses of rigid-body parameters of both base pairs and base-pair steps include the covariance matrices and derived sequence-dependent elastic constants. These knowledge-based parameters can be used to study many DNA bending and packaging problems, such as DNA cyclization [8] and nucleosome-positioning [9, 10] propensities.

### 3.3.4 Local summary and visualization

The local summary and visualization pages (Figure 3.3) give a detailed listing of the sequence context, H-bonding interactions, conformational parameters and atomic-level representations of each of the base pairs or base-pair steps incorporated in the data report. Each page contains three sections. The first section gives the complete sequence, the location(s) of the selected base pair(s), the number of H bonds between paired bases and the base-pair types (Watson-Crick or noncanonical) in the structural example. The second section lists the values of all conformational parameters associated with the given base pair or base-pair step, including the torsional angles about the glycosidic linkage and the attached sugar-phosphate backbones. The last section contains a two-dimensional stacking diagram of the base pair or base-pair step generated with 3DNA and a link to three-dimensional visualization and manipulation of the same unit with the JAVA-based Jmol software [23].

## 3.4 Concluding Remarks and Future Directions

3DNALandscapes allows a user to gather information and gain insight about DNA sequence-dependent conformation and deformability from known-high-resolution structures. In contrast to other structural databases [4, 5], which archive the Cartesian coordinates and/or derived parameters for individual structures, 3DNALandscapes enables searches and summarizes conformational data from multiple structures that meet selected criteria. To the best of our knowledge, there are no other databases with these unique capabilities.

The information collected in 3DNALandscapes also provides useful benchmarks for the analysis and simulation of other DNA structures. The data that characterize existing structures can be compared with new experimentally derived or computer-simulated DNA structures. The database can be used in combination with the 3DNA software tools [1, 2] or the w3DNA web interface for such analyses [3]. The knowledge-based potentials provided through 3DNALandscapes can be used in various computer applications, such as the simulation of fluctuating DNA polymers [8, 24] or the analysis

of nucleosome positioning on DNA [9, 10]. The access to large volumes of derived conformational information may stimulate new types of analyses and lead to new understanding of DNA structure and deformability.

We plan to connect 3DNALandscapes to the w3DNA server. We are currently investigating ways to identify nonredundant DNA-containing structures automatically and will include this information in future releases of 3DNALandscapes. We will update the database at regular intervals as new structures are added to the Protein Data Bank and Nucleic Acid Database.

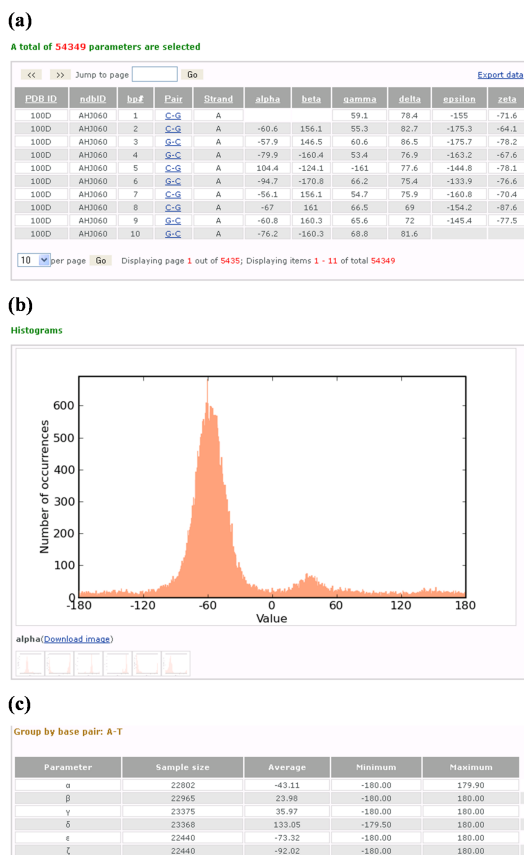


Figure 3.1: Screenshots illustrating some of the information about the DNA sugar-phosphate torsion angles collected from a search of the paired nucleotides in all DNA-containing crystal structures of 3Å or better resolution. (a) A table of 54349 entries from 1895 different structures arranged in seven columns that respectively list the Protein Data Bank and Nucleic Acid Database identifiers of the structures (PDB ID, NDB ID), the residue number (bp#), the chemical identities of the paired bases (Pair), the strand identity of the first of the two listed bases (Strand) and the values of the six torsion angles ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$  and  $\zeta$ ) in the specified nucleotides. The data can be sorted by clicking on the headers of the columns and also exported into a tab-delimited file. (b) A close-up of one of the downloadable histograms — here the distribution of the torsion angle  $\alpha$  about the O3'–P–O5'–C5' chemical bond sequence — automatically generated for each of the angles in the above data set. Moving the mouse across the different icons reveals the corresponding distributions for the other angles. (c) Summary of statistical information, including the number of examples, average values, minima and maxima for the torsion angles in the data set. The report is divided into groups based on the type and composition of base pairs, here the canonical A·T Watson-Crick pair.



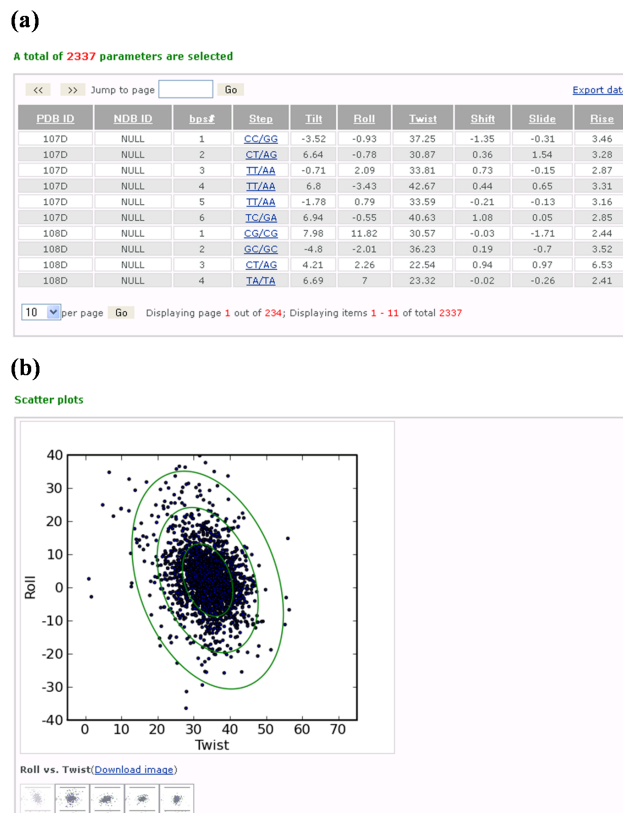


Figure 3.2: Screenshots showing some of the information about the arrangements of DNA base-pair steps extracted from a search of all DNA-containing structures derived by solution NMR spectroscopic measurements. (a) A table of 2337 sets of base-pair-step parameters relating unmodified base pairs in 586 representative models from 586 PDB files arranged in columns that respectively list the PDB identifiers (PDB ID), the base-pair-step numbers (bps#), the chemical identities (Step) and the (Tilt, Roll, Twist, Shift, Slide and Rise) values describing the steps. Data can be manipulated and downloaded as described in Figure 3.1. (b) One of the downloadable two-dimensional scatter plots, here Roll versus Twist, automatically generated for the above steps in the data set. Ellipses are projections of six-dimensional ‘equipotential’ surfaces derived, following [6], from the plotted data. Contours correspond to ‘energies’ where parameters deviate from mean values by no more than  $n$  times the root-mean-square deviation, where  $n = 1 - 3$ . Moving the mouse across the different icons reveals four other such plots: Roll versus Slide; Twist versus Slide; Tilt versus Shift; Twist versus Rise.

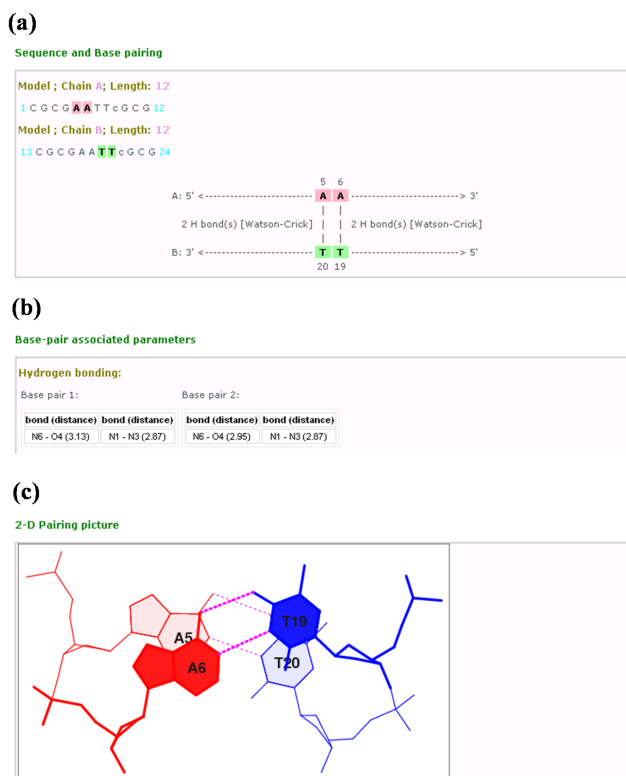


Figure 3.3: Screenshots showing some of the data provided in the local summary and visualization pages of each of the base-pair steps, here the  $A_5A_6 \cdot T_{19}T_{20}$  dimer from the 2.25-Å crystal structure of the Dickerson-Drew dodecamer complexed to netropsin (PDB ID: 101D) [25]. (a) A map of the interactions and locations of the paired bases in the selected base-pair step, including information on the number of hydrogen bonds and type of base pair. (b) A summary of some of conformational parameters associated with the base-pair step, here the identities and lengths of the hydrogen bonds in each base pair. (c) A downloadable stacking diagram that illustrates the overlap and hydrogen-bonding patterns of the base pairs ( $A_5 \cdot T_{20}$  and  $A_6 \cdot T_{19}$ ) in the selected step.

## References

- [1] Lu, X.-J. and Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
- [2] Lu, X.-J. and Olson, W. K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protoc.*, **3**, 1213–1227.
- [3] Zheng, G., Lu, X.-J., and Olson, W. K. (2009) Web 3DNA — a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.*, **37**, w240–w246.
- [4] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- [5] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R., and Schneider, B. (1992) The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
- [6] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**(19), 11163–11168.
- [7] Lu, X.-J., Shakked, Z., and Olson, W. K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
- [8] Czapla, L., Swigon, D., and Olson, W. K. (2006) Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theory Comput.*, **2**, 685–695.
- [9] Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K., and Zhurkin, V. B. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**(3), 725–738.
- [10] Balasubramanian, S., Xu, F., and Olson, W. K. (2009) DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. *J. Mol. Biol.*, **96**(6), 2245–2260.
- [11] MySQL <http://www.mysql.com>.
- [12] IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) (1983) Abbreviations and symbols for the description of conformations of polynucleotide chains. *Eur. J. Biochem.*, **131**, 9–15.

- [13] Altona, C. and Sundaralingam, M. (1972) Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Am. Chem. Soc.*, **94**, 8205–8212.
- [14] Xin, Y. and Olson, W. K. (2009) BPS: a database of RNA base-pair structures. *Nucleic Acids Res.*, **37**(Database issue).
- [15] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. H., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C., and Olson, W. K., et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 781–791.
- [16] Zhurkin, V. B., Lysov, Y. P., and Ivanov, V. I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, **6**, 1081–1096.
- [17] El Hassan, M. A. and Calladine, C. R. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA: a new local calculation scheme. *J. Mol. Biol.*, **251**, 648–664.
- [18] Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, R. C., Heinemann, U., Lu, X.-J., Neidle, S., and Shakked, Z., et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
- [19] Babcock, M. S. and Olson, W. K. (1994) The effect of mathematics and coordinate system on comparability and “dependencies” of nucleic acid structure parameters. *J. Mol. Biol.*, **237**, 98–124.
- [20] Guzikevich-Guerstein, G. and Shakked, Z. (1995) A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nat. Struct. Biol.*, **3**, 32–37.
- [21] El Hassan, M. A. and Calladine, C. R. (1998) Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.*, **282**, 331–343.
- [22] CodeIgniter <http://codeigniter.com>.
- [23] Jmol <http://www.jmol.org>.
- [24] Olson, W. K., Colasanti, A. V., Czapla, L., and Zheng, G. (2008) Insights into the sequence-dependent macromolecular properties of DNA from base-pair level modeling. In G. A. Voth, Editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Page 205–223, Taylor and Francis Group, Boca Raton, FL.
- [25] Goodsell, D. S., Kopka, M. L., and Dickerson, R. E. (1995) Refinement of netropsin bound to DNA: bias and feedback in electron density map interpretation. *Biochemistry*, **34**, 4983–4993.

## Chapter 4

### Sequence-dependent flexibility of DNA

The sequence-dependent structural properties of DNA play important roles at the mesoscopic level. In this chapter, we report statistically derived sequence-dependent features of DNA deformability, based on a non-redundant data set of 239 crystal structures of DNA in complexes with proteins. Computer calculation and simulation are performed to examine the effects of sequence on DNA flexibility in terms of the local molecular motions, persistence length, radial distribution, and nucleosome positioning.

#### 4.1 Introduction

The micromechanical behavior of DNA is sequence-dependent. That is, the sequence context of DNA can affect the extent to which the long, threadlike molecular fluctuates. For instance, a specific sequence with periodically repeating chemical features can bind much more tightly than a random DNA sequence of the same length to the histone octamer [1]. Also, placement of selected sequence motifs, such as TA base-pair steps, can considerably enhance the cyclization of short DNA [2]. Examination of the sequence-dependent properties of DNA is essential for understanding mechanisms involved in genetic processes and packaging.

The sequence-dependent features of DNA can be evaluated at multiple scales. At the local dinucleotide level, density distributions of base-pair step parameters, which describe the dimeric deformations of DNA, directly measure the local flexibility of the double helix in terms of the degree to which successive base pairs undergo translational and rotational motions. Obtaining such distributions with respect to different sequence contexts provides information about the sequence-dependent deformability of DNA. At the global level, the DNA persistence length and end-to-end distance distribution can

be used to assess DNA curvature and flexibility. The persistence length is a quantitative measurement of the extension of a polymer chain, and is commonly used to characterize basic mechanical properties of the polymer. Based on this concept, the DNA persistence length is the length over which the direction of DNA is maintained [3]. The DNA end-to-end distance measures the magnitude of displacement between the first and last base pairs of the double helix and implies the bending curvature of the DNA as a whole when compared to its contour length. Given this, distributions of DNA end-to-end distances statistically represent the ease of DNA bending and fluctuation. The sequence-dependent mechanics of DNA can be also assessed by hypothetically packaging the double helix into a nucleosome. A threading method, recently developed [4] to investigate the ease of DNA packaging onto the nucleosome, depends upon how the underlying sequence dictates its conformational features.

Obtaining the distributions of base-pair step parameters requires a reliable and comprehensive data resource. We have developed an informative database containing a variety of DNA structural parameters, including base-pair step parameters, extracted from all currently available DNA structures in the absence or presence of proteins (Chapter 3). This database serves as a groundwork for the knowledge-based analysis of DNA sequence-dependent flexibility, reported in this chapter and used in calculations of DNA persistence length, end-to-end distances, and threading scores. All of these calculations are based on an elastic DNA model which allows the specification of DNA base composition.

## 4.2 Methods

### 4.2.1 Non-redundant Structures

A non-redundant pool of protein-DNA crystal complexes of 2.5 Å or better resolution extracted from the Nucleic Acid Database [5], was identified by Y. Li [6] for the purpose of reducing sample bias in statistical inferences of DNA properties. The selection and classification of these complexes were based on an integration of information from sequence alignment, structural alignment, and the SCOP (Structural Classification

of Proteins) protein-folding-domain classification database [7]. Over-represented complexes were then filtered out from each classified group, in order to obtain a balanced sample of spatial and functional forms. The resulting dataset includes 101 structures of double-helical DNA bound to enzymes, 121 duplexes in the presence of regulatory proteins, 16 complexes with structural proteins, and one DNA associated with a multifunctional protein [8].

#### 4.2.2 Chain Model and Dimensions

Base-pair level models of DNA are constructed from the serial products of generator matrices  $\mathbf{A}_n$  that incorporate the displacement vectors  $\mathbf{r}_n$  and the rotation matrices  $\mathbf{T}_n$ , which relate coordinate frames on successive base pairs:  $\mathbf{A}_{1:N} = \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_{N-1} \mathbf{A}_N$  [9], where

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{T}_n & \mathbf{r}_n \\ \mathbf{0} & 1 \end{bmatrix}, \mathbf{A}_{1:N} = \begin{bmatrix} \mathbf{T}_{1:N} & \mathbf{r}_{1:N} \\ \mathbf{0} & 1 \end{bmatrix}. \quad (4.1)$$

Values used to evaluate chain configuration — (i) the end-to-end vector  $\mathbf{r}_{1:N+1}$ , (ii) the cosine of the angle  $\gamma$  between the normals of terminal base pairs, and (iii) the twisting  $\tau$  of terminal base pairs — are embedded in  $\mathbf{A}_{1:N+1}$  [2]:

$$\mathbf{r}_{1:N} = \begin{bmatrix} \mathbf{I}_3 & 0 \end{bmatrix} \mathbf{A}_{1:N+1} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}, \quad (4.2)$$

$$\cos\gamma = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{A}_{1:N+1} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad (4.3)$$

$$\text{Tr}(\mathbf{T}_{1:N+1}) = \cos\tau(1 + \cos\gamma) + \cos\gamma. \quad (4.4)$$

Here  $\mathbf{I}_3$  is the identity matrix of order three and the  $\mathbf{0}$ s are null matrices of orders necessary to fill the  $3 \times 4$  premultiplication and  $4 \times 1$  postmultiplication vectors. A joining step  $N + 1$ , which is included in these expressions to test for terminal base-pair

overlap, is subsequently removed and circles are closed by a step  $c$  that connects the  $N$ th to the first base pair, i.e.,  $\mathbf{A}_{1:N}\mathbf{A}_c = \mathbf{I}_4$ , where  $\mathbf{I}_4$  is the  $4 \times 4$  identity matrix.

### 4.2.3 Deformation Energy

The deformational energy  $U$  of a configuration of DNA is the sum, over  $n$ , of the energy of interaction  $\Psi_n$  of the  $n$ th and  $(n + 1)$ th base pairs,  $U = \sum_{n=1}^N \Psi_n$ . Here  $\Psi_n$  is a function of the relative orientation, the displacement, and the chemical composition of base pairs  $n$  and  $n + 1$ , and  $N$  is the number of base-pair steps that make up the DNA. The known complementarity of Watson-Crick base pairs, i.e., the specific association of adenine with thymine (A·T) and guanine with cytosine (G·C), and the antiparallel directions of the sugar-phosphate chains place restrictions on the  $\Psi_n$ . That is, step parameters are defined such that tilt and shift  $(\theta_1, \theta_4)$  change signs in complementary strands [10], and the potential  $\Psi_n(\text{XZ})$  of dimer step XZ determines that of its complement  $\text{X}'\text{Z}'$  [11, 12].

The deformability of DNA is based on the range of configurational states found in a non-redundant set of 239 protein-DNA crystal complexes of 2.5 Å or better resolution, taken from the Nucleic Acid Database [5]. The dataset includes 101 structures of double-helical DNA bound to enzymes, 121 duplexes in the presence of regulatory proteins, 16 complexes with structural proteins, and one DNA associated with a multifunctional protein [8]. The structures have been filtered to exclude over-represented complexes in order to obtain a balanced sample of spatial and functional forms. The dinucleotide samples exclude chemically modified bases, terminal and penultimate base pairs, and side groups attached to nicked backbones. The working dataset also omits base pairs on nucleotides that are attached to modified or mispaired residues. The preferred arrangements and likely fluctuations of base-pair steps are derived from the average properties of the dimeric units in these structures [12].

The cost of deformation  $\Psi_n(\text{XZ})$  of a given base-pair step is expressed by a double summation of elastic terms over the six base-pair step parameters:

$$\Psi_n(\text{XZ}) = \frac{1}{2} \sum_{i=1}^6 \sum_{j=1}^6 f_{ij}(\text{XZ}) \Delta\theta_i^n \Delta\theta_j^n. \quad (4.5)$$



Here  $\Delta\theta_i^n = \theta_i^n - \theta_i^0(\text{XZ})$  is the imposed deviation of the  $i$ th step parameter  $\theta_i^n$  at the  $n$ th dinucleotide step from the equilibrium rest-state value  $\theta_i^0(\text{XZ})$  of the XZ dimer step, and the  $f_{ij}(\text{XZ})$  are stiffness constants determined by the XZ sequence. The rest-state values of the dinucleotide steps are equated to the average step parameters of the XZ dimers in the protein-DNA sample, i.e.,  $\theta_i^0(\text{XZ}) = \langle \theta_i(\text{XZ}) \rangle$  ( $i = 1 - 6$ ), and the stiffness constants are extracted from the pairwise covariance of these variables; that is, the covariance matrix with elements given by the differences between the mean squares and the squares of the means of all pairs of step parameters,  $\langle \theta_i(\text{XZ})\theta_j(\text{XZ}) \rangle - \langle \theta_i(\text{XZ}) \rangle \langle \theta_j(\text{XZ}) \rangle$ , and equal to the inverse of the  $6 \times 6$  force-constant matrix  $\mathbf{F}(\text{XZ})$  that contains the  $f_{ij}(\text{XZ})$  [12]. Such an approach accounts for both the sequence-dependent structure of DNA and the correlations of dinucleotide step parameters, which are especially important for “realistic” models of DNA. The model, however, omits consideration of (i) the sequence context of the given dimer, i.e., the spatial configuration of a given dimer is assumed to be independent of that of adjacent base-pair steps, (ii) the precise arrangement of complementary purine and pyrimidine bases, such as the propeller and buckle angles that effect base-pair non-planarity, (iii) the detailed arrangement of the sugar-phosphate backbone, and (iv) the “structure” of the surrounding chemical environment. Backbone and solvent atoms are implicitly treated in the energy terms so that their omission introduces no serious error when duplex deformations are limited to energies of the order of  $k_B T$ , where  $k_B$  is the Boltzmann constant and  $T$  the temperature, and the DNA remains within the limits of the B-DNA family. If distortions are large, such as in a “melted” helix, these atoms should be incorporated, along with complementary base-pair parameters, i.e., the rigid-body parameters that describe the orientation and displacement of paired purine and pyrimidine bases, in the DNA model. The energies derived from the protein-DNA crystal set must also be scaled to account for known configuration-dependent properties of DNA in solution (see below).

#### 4.2.4 Configurational Sampling

By denoting the configuration of base-pair step  $n$  by the vector  $\Theta_n$ , with components  $\theta_i^n$  ( $i = 1 - 6$ ) corresponding to the instantaneous values of the angular and translational parameters at the given step and defining  $\Theta^0(\text{XZ})$  as the vector that contains the intrinsic step parameters of dimer XZ, the potential of the step can be expressed in matrix form as  $\Psi_n(\text{XZ}) = (1/2)\Delta\Theta^T \mathbf{F}(\text{XZ})\Delta\Theta$ , where  $\Delta\Theta = \Theta_n - \Theta^0(\text{XZ})$ . To facilitate the sampling of representative chain configurations, each dimeric energy contribution  $\Psi_n(\text{XZ})$  is reexpressed in terms of a diagonal matrix  $\mathbf{D} = \mathbf{Q}\mathbf{F}(\text{XZ})\mathbf{Q}^T$  and a basis variable set  $\Omega_n = \mathbf{Q}\Theta_n$ , with elements  $\omega_i^n$  ( $i = 1 - 6$ ) given by linear combinations of the base-pair step parameters [2]. Here  $\mathbf{Q}$  is the eigenvector matrix specifying the directions of the principal axes of deformation, and the superscript  $T$  is used to denote the transpose. Elimination of the cross terms in the energy expression makes it possible to write the probability density function for a single base-pair step, including normalization, as a product of Gaussians. This function can be sampled with a standard Gaussian random-number generator [13] and a Boltzmann distribution of states can be collected without the necessity of using the Metropolis method [14]. Such an approach is superior to the Metropolis method in that it is computationally more efficient and does not suffer from correlations between sample points or incomplete coverage of phase space. Gaussian sampling cannot be used, however, if the potential function includes long-range electrostatic terms.

#### 4.2.5 Persistence Length

The persistence length  $a$  is computed from the projection of the mean end-to-end vector  $\langle \mathbf{r} \rangle$ , the so-called persistence vector [15], at infinite chain length along the initial direction of the chain; that is,  $a = \langle \mathbf{r}_\infty \rangle \cdot \mathbf{r}_1 / |\mathbf{r}_1|$ . If the dimeric chain units are independent,  $\langle \mathbf{r} \rangle$  can be determined from the product  $\mathbf{P}_N = \langle \mathbf{A}_1 \rangle \langle \mathbf{A}_2 \rangle \cdots \langle \mathbf{A}_{N-1} \rangle \langle \mathbf{A}_N \rangle$  of average generator matrices  $\langle \mathbf{A}_n \rangle$  [9]. The components of  $\langle \mathbf{r} \rangle$ , which accumulate in the far right column of  $\mathbf{P}_N$ , approach limiting values with increasing  $N$ , owing to the non-orthogonality of each  $\langle \mathbf{T}_n \rangle$  matrix of the flexible duplex [16]. Thus the persistence

length of DNA can be obtained by calculating the limiting value of the [3, 4] matrix element of  $\mathbf{P}_N$ :

$$a = \lim_{N \rightarrow \infty} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{P}_N \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad (4.6)$$

#### 4.2.6 J-Factor

The  $J$ -factor depends on the fraction  $M_c/M$  of configurations that meet the criteria of chain closure, namely that (i) the end-to-end vector  $\mathbf{r}$  is null,  $W(\mathbf{r} = \mathbf{0})$ , (ii) the terminal normals are aligned; that is, the cosine of the angle between the normals of the first and last base pairs is unity, given that the vector  $\mathbf{r}$  is null,  $\Gamma_r$  ( $\cos \gamma = 1$ ), and (iii) the end-to-end twist is zero, given that the normals are aligned and the vector  $\mathbf{r}$  is null,  $\Phi_{r, \cos \gamma}(\tau = 0)$  [17]. The product of these probability densities is approximated by choosing three corresponding bounds: (i) the magnitude of  $\mathbf{r}$  being less than  $r_0$ ; (ii) the cosine of the angle  $\gamma$  between the normals of terminal base pairs being greater than  $1 - \Gamma_0$ ; and (iii) the magnitude of the end-to-end twist being less than  $\tau_0$ . Thus, the  $J$  factor is given by

$$J = \frac{4\pi}{N_A} W(|\mathbf{r}| \leq r_0) \Gamma_r(\cos \gamma \geq 1 - \Gamma_0) \Phi_{r, \cos \gamma}(\tau \leq \tau_0) = \frac{1}{K} \frac{M_c}{M}, \quad (4.7)$$

where  $K = 4\pi N_A r_0^3 \Gamma_0 \tau_0 / 3$ ,  $N_A$  is Avogadro's number,  $M_c$  is the number of configurations that satisfy the three closure constraints, and  $M$  is the total sample size. The bounds used here —  $r_0 = 10 \text{ \AA}$ ,  $\Gamma_0 = 0.02$ ,  $\tau_0 = 11.5^\circ$  ( $\cos \tau_0 = 1 - 0.02 = 0.98$ ) — are very restrictive, constraining the trace of  $\mathbf{A}_{1:N}$  to values very close to 3 and the radial bound to distances no more than 5% of the contour length of the sampled DNA chains. Previous work [2] has shown that such bounds yield the most accurate results for  $M_c \geq 1000$ .

#### 4.2.7 DNA Threading

The nucleosome-binding affinity of a given DNA sequence is estimated by “threading” the constituent base pairs on the three-dimensional pathway found in the currently

best-resolved nucleosome core-particle structure [18] and calculating a knowledge-based deformation score in terms of the deviations of the base-pair step parameters that make up the structure from their preferred equilibrium values. The total “energy”  $U$  of the threaded sequence is expressed as a sum of quadratic terms  $U = \sum_{n=1}^N \Psi_n$ , where  $\Psi_n$  is given by Eqn. 4.5 and  $N$  is the number of base-pair steps that comprise the nucleosome template. Here  $\theta_i^n$  is the value imposed on the  $i$ th step parameter at the  $n$ th dinucleotide step of the assumed structure. This approach assumes that the core of histone proteins imposes exactly the same configurational constraints on DNA regardless of base-pair sequence and ignores the occurrence of gaps, for example, small “bubbles” of unbound duplex that may loop away from the surface of the nucleosome [19].

### 4.3 Results

#### 4.3.1 Knowledge-based potentials

The equi-potential surfaces in Figure 4.1 illustrate the sequence-dependent deformability and structural interdependence of DNA dimer steps in the non-redundant set of protein-DNA structures. The contour plots reveal the distinctive equilibrium (average) rest states of the 10 unique dimers [12, 20] and the strong coupling of rigid-body parameters found in most base-pair steps. By contrast, there is no sequence dependence, bending and twisting are uncoupled, and there are no translational deformations in the classical representation of DNA as an inextensible elastic rod. The ellipses in the figure are projections of the multi-dimensional potential surface of each dimer on the roll-twist plane obtained from the covariance of two different sets of  $(\theta_2, \theta_3)$  values: (i) a ‘refined’ set of dimer steps (dots), found by iteratively removing outlying states (open circles) of extreme bending, twisting, and stretching, i.e., states with one or more step parameters that deviate from their respective mean values by more than three times their root-mean-square deviations before culling [12]; and (ii) a ‘complete’ set of structural examples (dots and open circles). The contours correspond to deviations of parameters equal to two times the combined root-mean-square deviations of the  $\theta_i$  in the selected sets of data and thus encompass  $\sim 95\%$  of the reference points. The mean

values of roll and twist in the respective datasets are highlighted by thin (solid and dashed) lines, and the contour surfaces by curves of the same style.

As is clear from Figure 4.1, the coupling of roll and twist depends upon sequence and dataset. The positive values of  $f_{23}$  associated with the dimer steps produce energy pathways that involve a decrease in one angle and an increase in the other, mimicking the observed variation of step parameters. The extent and direction of parametric coupling reflect the sequence and choice of reference points. For example, whereas all other ‘complete’ dimers tend to deform more easily via roll than twist, the GC·GC steps in the dataset twist slightly more easily than roll. The extreme distortions of DNA found in the crystal complexes similarly reflect sequence: for example, the CA·TG steps show a propensity to take up the large negative roll values associated with the kinking of DNA into the minor groove, while the AA·TT, TA·TA, and CG·CG dimers tend to kink more easily via large positive roll into the major groove. The coupling of TA·TA parameters changes direction when outlying states of extreme roll and twist are included. Consideration of the outlying states also softens the knowledge-based potentials, with some of the most pronounced changes in deformability occurring at AA·TT, AT·AT, and TA·TA steps (note the larger areas spanned by the dashed contours of the ‘complete’ potentials compared to the solid contours of the ‘refined’ potentials for these steps). The relative deformability of CG·CG dimers compared to other base-pair steps also changes substantially if outlying states are considered.

In addition to the roll-twist correlations noted above, roll and twist are frequently coupled to slide, the local displacement of neighboring base pairs along their long axes. Roll-slide coupling is very sensitive to sequence: whereas the roll and slide of CA·TG, TA·TA, and GC·GC dimers show negative correlations in both the ‘refined’ and the ‘complete’ datasets, the parameters are positively correlated at most other base-pair steps [12]. By contrast, the  $f_{35}$  twist-slide constants are predominantly negative and the correlations of twist and slide are positive (data not shown).

### 4.3.2 Intrinsic motions

The molecular images in Figure 4.2 illustrate the pathways of preferred DNA deformation deduced from the known structures. The sets of low-energy librations, which lie along the longest principal axes of the knowledge-based potentials, i.e., in the direction of most probable configurational change, are reminiscent of the normal modes of vibration of small molecules. The illustrated motions involve combinations of roll and twist plus varying degrees of translation, dictated by the set of ‘complete’ potentials. As is clear from the images, correlations between roll and twist dominate the preferred movements of the 10 unique base-pair steps. The CA·TG step, however, incorporates significant translational changes along this lowest energy pathway, whereas the GA·TC and AT·AT steps involve essentially no base-pair displacement. The GG·CC step also includes variation in tilt, although the observed changes are substantially lower than those of roll. The illustrated moves correspond to one of the six directions of configurational sampling, i.e., linear combinations of base-pair step parameters, used in the Monte-Carlo simulation of polymeric structures.

### 4.3.3 Persistence length

The values of the persistence length in Table 4.1 show how the dimeric deformability of a given base-pair step influences the global properties of DNA. Each numerical value in the table gives the computed mean extension along the initial direction of a hypothetical, naturally straight homopolymer with a helical repeat of 10.5 bp per turn and local elastic properties corresponding to those deduced for the designated step in the specified structural sample. The force constants are scaled by a factor  $\xi$  so that the persistence length of a mixed-sequence DNA homopolymer is  $\sim 500\text{\AA}$ , or  $\sim 150$  bp, i.e.,  $f_{ij}^\dagger(XX) = \sum f_{ij}^\dagger(XZ)$ , where  $f_{ij}^\dagger(XX)$  is the force constant of the mixed-sequence repeating unit,  $f_{ij}^\dagger(XZ) = \xi f_{ij}(XZ)$  is the scaled force constant of the XZ base-pair step, and the summation is carried out over all 16 possible steps. The values of  $f_{ij}^\dagger(XZ)$  determine the range of step parameters sampled for the specified dimer and thus the average components of the generator matrices  $\langle \mathbf{A}_n \rangle$  used in Eqn. ?? to determine the

limiting values of  $a$ .

The values of  $\xi$  reveal the extent to which the sampled points mimic the average properties of DNA in solution. Interestingly, mixed-sequence homopolymers guided by the potentials of the ‘complete’ dataset have chain extension properties more closely resembling those known to characterize polymeric DNA than chains that are subject to the deformations associated with the ‘refined’ more B-like dataset. That is,  $\xi$  is closer to unity for the ‘complete’ homopolymer than the ‘reduced’ homopolymer, with the range of accessible configuration space increased by a factor of  $1.18 = 0.85^{-1}$  in the former case and  $2.0 = 0.5^{-1}$  in the latter case to yield a persistence length of  $\sim 500\text{\AA}$ . The persistence lengths of mixed-sequence homopolymers that conform to the unscaled potentials are greater than  $500\text{\AA}$ , i.e.,  $592\text{\AA}$  for the ‘complete’ potential and  $995\text{\AA}$  for the ‘refined’ potential. Thus, the occasional adoption of extreme configurational states like those induced by the binding of proteins appear, from this perspective, to be necessary to account for the observed persistence length of mixed-sequence DNA in solution.

The data in Table 4.1 further show that AC·GT and GC·GC steps have more pronounced stiffening effects at the polymeric level than other base-pair steps, with longer computed persistence lengths. As is clear from the contours of the scaled potentials in the roll-tilt( $\theta_1, \theta_2$ ) plane (Figure 4.2), these steps bend to a much lesser extent than the other dimers. Furthermore, the AC·GT step is even stiffer than the mixed-sequence homopolymeric repeating unit that yields a persistence length of  $\sim 500\text{\AA}$ . Similarly, TA·TA steps stand out as being highly bendable at both the global and local levels, although the local bending deformability, as measured by the area within the corresponding energy contours, is somewhat greater for CG·CG compared to TA·TA steps that obey the ‘complete’ potential. The degree of dimeric bending and the values of  $a$  based on the ‘complete’ potentials are much more sensitive to sequence than the corresponding values associated with the ‘refined’ functions. Figure 4.3 also includes the contour surface of a dimer subject to the classic elastic-rod model of DNA. Notably, none of the ‘real’ dimers exhibits the bending isotropy assumed in the ideal model. The well-known anisotropy of DNA bending, i.e., the preferential bending of base-pair steps via roll rather than tilt [21], is clear from the elliptical (as opposed to circular) shapes

of the derived contour surfaces.

#### 4.3.4 Radial distribution

The radial density functions in Figure 4.4 show how the placement of individual dimers affects the range of accessible configurations of a series of 94-bp DNA molecules compared to that of a mixed-sequence DNA of the same chain length. The molecules, which are detailed in Table 4.2, include two fragments, TA-94 and S5-94, taken from well characterized nucleosome-positioning sequences [22, 23] and found to form small minicircles [24, 25], and four sequences — E6-94, E8-94, E13-94, CA-94 — used as experimental controls in the determination of the  $J$  factor (19, 20). All of the sequences shift the distribution of the end-to-end distance  $r$  toward smaller values than those determined for the mixed-sequence chain. Moreover, the ends of the nucleosome-positioning sequences are more likely to come into close contact than the ends of the control sequences. That is, the tails of the distributions formed from the most compact arrangements of the positioning sequences lie closer to zero than the tails of the control sequences. The boundary delimiting the 10% shortest configurations,  $r_{0,1}$ , is smaller for the positioning sequences than the control sequences, and both limits are substantially smaller than the  $r_{0,1}$  boundaries for mixed-sequence DNA and an ideal 94-bp DNA model (Table 4.2). Furthermore, the values of  $r_{0,1}$  computed with the “complete” potentials are roughly proportional to the negative logarithm of the reported  $J$  factors. The likelihood of ring closure is lower for chains that obey the “refined” potentials, i.e., the value of  $r_{0,1}$  associated with a given sequence is larger compared to that obtained with the “complete” sequence. The values of  $r_{0,1}$ , however, do not take account of the orientational constraints (see Eqn. 4.7) that must be met for successful ring closure and considered in the calculations reported below.

The TA-94 and 5S-94 fragments stand out from the other sequences in Table 4.2 in containing flexible dimers with strong bending propensities and appreciable coupling of roll and twist, such as the TA and CG steps of the “complete” potential (highlighted in boldface), that recur approximately in phase with the  $\sim 10.5$ -bp double-helical repeat. The fragments with larger values of  $r_{0,1}$  and smaller  $J$  factors; that is, greater values



of  $-\log J$ , contain few such steps.

### 4.3.5 Threading score

The ‘cost’ of threading the same three sequences on the central 60 base-pair steps of the best-resolved nucleosome core-particle structure [18] is reported in Figure 4.5. The imposed distortions of DNA reflect the close contact with the (H3-H4)<sub>2</sub> tetramer that is believed to be critical to nucleosome positioning [26, 27]. The 34 settings of each sequence on the crystalline template are described in terms of the displacement, with respect to the central base-pair step, of the nucleotide that is placed on the twofold structural dyad. Here, since the sequences contain an even number of base pairs, the settings are numbered from -17 to +17, without a zero entry. In order to extract the contribution of dimeric deformability to positioning, the sequences are assigned an unsheared, naturally straight, B-like rest state with 10.5 bp per helical turn.

Although the crystallographic template accommodates the regularly repeating TA-94 sequence in several relatively low-cost settings (denoted by triangles in Figure 4.5), none of these corresponds to the +1 setting that aligns most closely with the observed positioning of the 601 sequence, from which TA-94 is derived. That is, none of the local TA-94 minima in the computed scoring profile is in register with the observed setting of nucleosomes on 601, regardless of the choice of scoring function.

The TA-94 fragment shares 83% sequence identity with base pairs 88 to 181 of the 232-bp 601 sequence, falling in the middle of the stretch found to position nucleosomes. That is, base pair 47 of TA-94 coincides in this alignment with the observed location of the dyad on 601 at base pair 134 (J. Widom, personal communication). By contrast, the predicted sites of nucleosome binding on TA-94 recur at 10-11-bp increments in settings where the naturally flexible TA·TA steps of the sequence easily take up the ‘kink-and-slide’ states of nucleosomal DNA [28], in which roll is negative ( $\theta_2 < -10^\circ$ ) and slide is highly positive ( $\theta_5 > 1.5\text{\AA}$ ). The 4-bp discrepancy in the predicted positioning of TA-94 vs. the observed positioning of 601 may reflect limitations of the model in dealing with the sequence-dependent features of the 601 sequence [28] and/or subtle differences in the sequence of TA-94, including the replacement of two GG·CC steps in 601 by

phased TA·TA steps in TA-94, that bias the positioning. Indeed, the same predicted nucleosome positions occur with the ‘complete’ and ‘refined’ potentials.

By contrast, there are no deep minima in the scoring profiles of the E6-94 control sequence, although the overall cost of nucleosomal deformation is lower for E6-94 than TA-94 (note the relative displacement of the ‘energy’ profiles with respect to the fixed cost of deforming a mixed-sequence homopolymer (dashed line) on the nucleosome). Thus, there are no intrinsic features in the E6-94 sequence that accommodate the known distortions of nucleosomal DNA in a particular setting.

The cost of deforming the CA-94 sequence on the nucleosome is much lower than that for the other sequences. The CA·TG steps, which repeat at 10-11 bp along CA-94, in phase with the double-helical repeat, accommodate the positive slide found at distorted nucleosomal steps much more easily than any other dimer. The lower cost of sliding contributes, in turn, to the low positioning scores despite the higher cost of bending CA·TG compared to TA·TA steps. In fact, deformations in slide make a contribution to the total positioning score that is comparable to, if not greater than, that from roll [28]. The slight displacement of the CA·TG steps on CA-94 relative to the positions of the TA·TA steps on TA-94 accounts for the 1-2-bp shift in the predicted settings of nucleosomes on CA-94 compared to TA-94.

#### 4.4 Concluding Remarks

The mathematics used to relate local base-pair structure to global chain configuration underlies successful “realistic” treatment of polymeric DNA. The distribution of accessible chain configurations governs the overall behavior of long DNA fragments, including the likelihood of loop formation and the ease of wrapping on the surface of the nucleosome. The naturally discrete representation of DNA described herein [11] is general in the sense that any functional description of DNA dimeric geometry can be employed; that is, not just the harmonic form of the knowledge-based potentials that have been extracted from the three-dimensional arrangements of DNA base-pair steps in high-resolution crystal structures [12]. The latter functions incorporate the intrinsic structure, the sequence-dependent fluctuations, the anisotropy of DNA deformations,

and the known correlations of base-pair step parameters.

These local dimeric features translate into measurable effects at the macromolecular level, giving useful new insights into the contribution of base sequence to the mesoscopic properties of DNA. The examples presented here show how judicious placement of flexible base-pair steps enhances the likelihood of ring closure and lowers the cost of deforming a DNA sequence on the surface of a nucleosome. Thus, the regular repetition of TA·TA steps in phase with the helical repeat of the TA-94 sequence promotes spontaneous ring closure and preferential positioning of nucleosomes on DNA. The bending flexibility of the TA·TA steps gives rise to a relatively high proportion of compact polymer configurations with chain ends close enough to effect cyclization. The relative ease of TA·TA bending, in combination with its coupled propensity to slide, lowers the cost of deforming particular settings of the sequence on the nucleosome [28]. By contrast, the regularly repeated CA·TG steps in the CA-94 sequence inhibit chain cyclization but enhance the wrapping of the sequence on the surface of the nucleosome. These steps, although not as easily deformed via roll as TA·TA dimers, readily take up the costly sliding deformations found in nucleosomal DNA. Other dimers, such as CG·CG steps, which easily bend but resist sliding in the positive sense observed on the nucleosome, could be used in the design of DNA molecules that would preferentially loop rather than form nucleosomes.

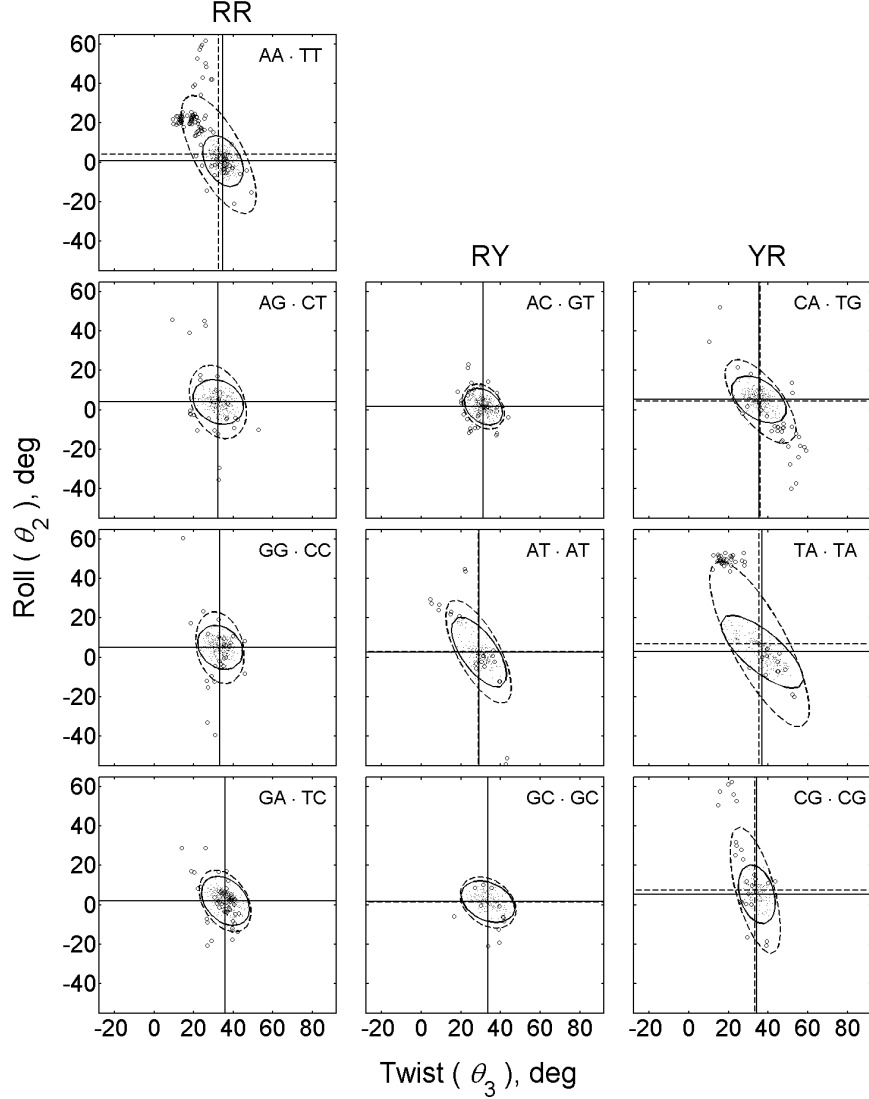


Figure 4.1: Collective scatter plots in the roll-twist ( $\theta_2, \theta_3$ ) plane of base-pair step parameters found in high-resolution protein-DNA crystal complexes and derived sequence-dependent potentials of the ten unique dimer steps. Dots correspond to the points used to derive the ‘refined’ potentials (solid contours) and open circles to the states of extreme bending, twisting, and stretching that are included with the preceding points in the ‘complete’ functions (dashed contours). Ellipses are projections of the multi-dimensional potentials on the  $\theta_2, \theta_3$  plane obtained from the  $2 \times 2$  covariance matrix of observed roll-twist values. Contours correspond to deviations of parameters equal to two times the combined root-mean-square deviations of  $\theta_2$  and  $\theta_3$ . Average values of roll and tilt are highlighted by thin (solid and dashed) lines. The three columns show the respective deformational patterns of individual purine-purine (RR), purine-pyrimidine (RY), and pyrimidine-purine (YR) steps.

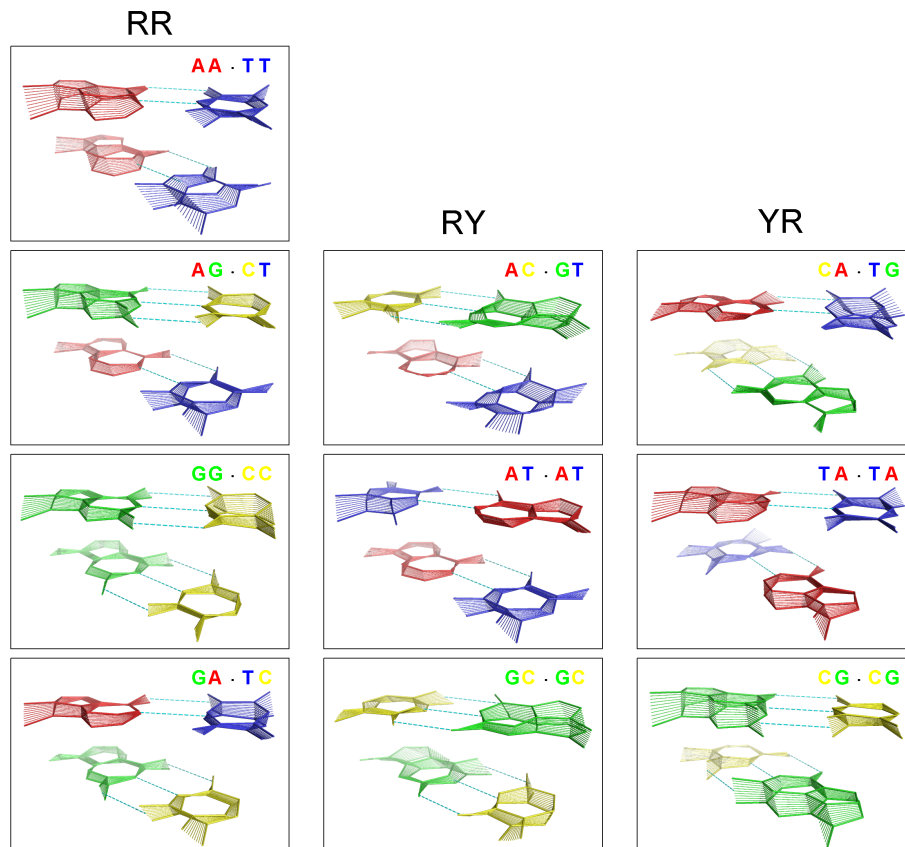


Figure 4.2: Sequence-dependent motions along the longest principal axes of the 10 unique DNA base-pair steps. Non-equilibrium forms are superimposed on the intrinsic (average) dimer structures. Perturbed states correspond to deformations, at increments of  $3\langle\lambda_1^2\rangle^{1/2}$ , along the longest principal axes of the “complete” knowledge-based potentials, where  $\lambda_1$  is the largest eigenvalue of the covariance matrix, and ‘energies’ range from zero to  $4.5m^2k_BT$  for displacements of  $\pm 3m\langle\lambda_1^2\rangle^{1/2}$ . Here  $m$  is set to 5 to enhance visualization of structural deformations. Base pairs are represented as ideal Watson-Crick pairs, with the hydrogen bonds of rest structures denoted by dashed lines. Bases are color-coded according to chemical identity: adenine (red); thymine (blue); guanine (green); cytosine (yellow). Motions are illustrated with respect to the ‘middle’ frame of each step and viewed into the minor groove of the upper 3′-base pair of each miniduplex. Note the correspondence of observed structural variability with the corresponding contour surfaces in Figure 4.1

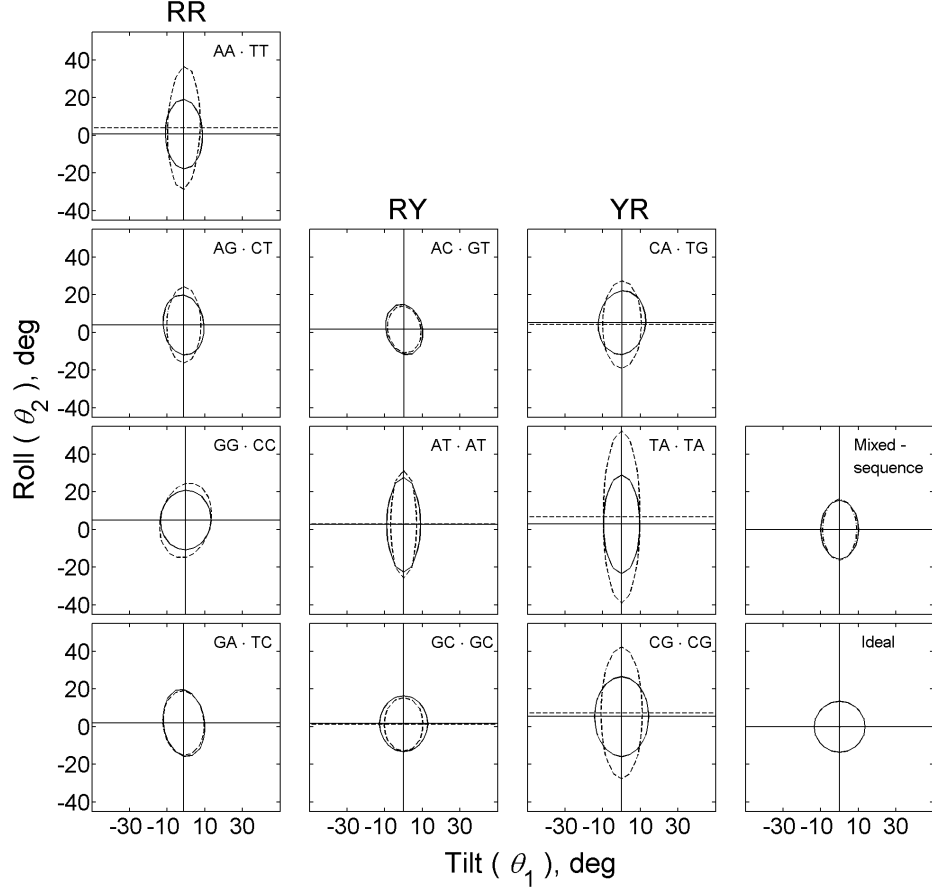


Figure 4.3: Contour surfaces in the roll-tilt  $(\theta_2, \theta_1)$  plane of scaled, knowledge-based potentials of the 10 unique base-pair steps (columns 1-3), the dimeric repeating unit of a naturally straight, mixed-sequence DNA homopolymer with force constants averaged over all 16 dimeric potentials and weighted to yield a persistence length  $a$  of  $\sim 500\text{\AA}$  (column 4, top), and the dimeric repeat of an ideal DNA elastic rod with the same value of  $a$  (column 4, bottom). See text and legend to Figure 4.1.

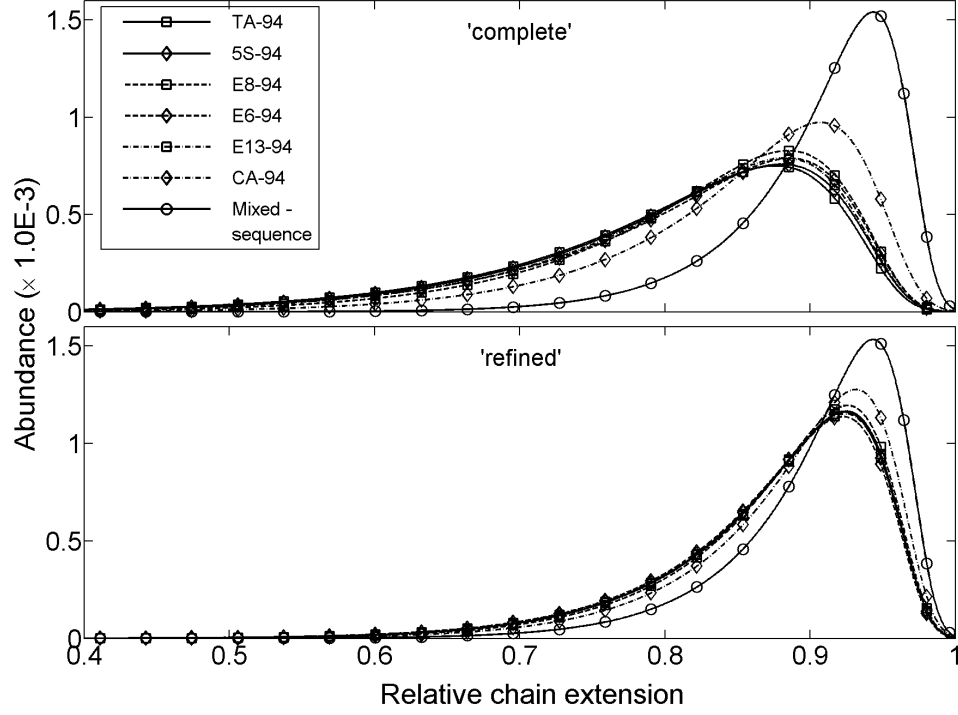


Figure 4.4: Distributions of the end-to-end distances  $r$  for a series of 94-bp DNA molecules compared to that of a mixed-sequence DNA of the same chain length. The double helix is assumed to be naturally straight in its equilibrium rest state with a 10.5 bp double-helical repeat. Fluctuations of local structure in polymeric sequences are based on Monte-Carlo sampling of the scaled, knowledge-based potentials of the 10 unique dimers and the mixed-sequence potential that yields a persistence length of  $\sim 500\text{\AA}$ . Radial distributions of  $2.5 \times 10^8$  sampled configurations are expressed in terms of relative chain extension,  $r/L_0$ , where  $L_0$  is the contour length of the fully extended polymer ( $93 \text{ bp steps} \times 3.4\text{\AA}/\text{step} = 316.2\text{\AA}$ ). Chain sequences are listed in Table 4.2.

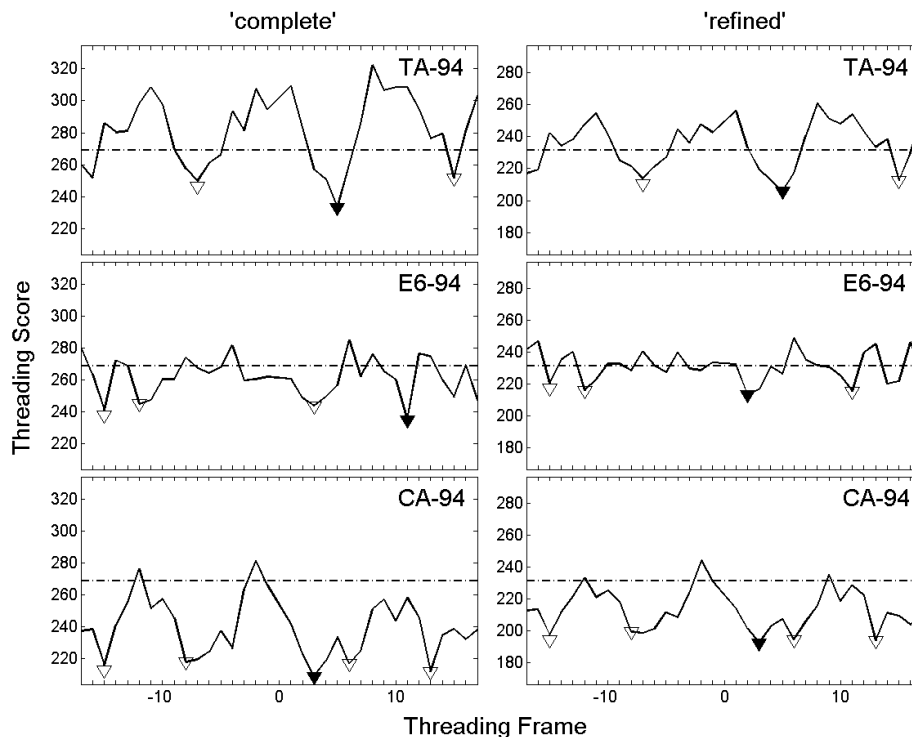


Figure 4.5: Deformation profiles of representative DNA sequences ‘threaded’ on the central 60 base-pair steps of the currently best-resolved nucleosome core-particle structure [18]: the TA-94 sequence derived from the 232-bp synthetic high-affinity ‘601’ sequence (top); the ‘random’ E6-94 sequence used as a control in ring-closure experiments (middle); and the CA-94 sequence with CA·TG dimer steps repeated at 10-11-bp intervals (bottom). The major minima in the ‘energy’ profiles, which are denoted by triangles (the filled triangle corresponding to the deepest minimum), are taken as ‘predicted’ nucleosomal dyad positions. The settings are numbered with respect to the center of each sequence; note that there is no zero position. The threading scores of the sequences (black lines) are compared at each test position with the score of a mixed-sequence homopolymer (dashed line). Data are reported for chains subject to both ‘complete’ and ‘refined’ dimeric potentials.



Table 4.1: Persistence lengths of hypothetical, naturally straight DNA homopolymers with knowledge-based elastic properties for individual base-pair steps. *a.* Persistence lengths obtained using Eqn. 4.6 with average generator matrices  $\langle \mathbf{A}_n \rangle$  based on Monte-Carlo samples of  $10^6$  states of the designated dimeric repeating unit subject to the specified knowledge-based potential. All steps assigned an unsheared, naturally straight, B-like rest state with 10.5 bp per helical turn, i.e.,  $\theta_1^0(\text{XZ}) = \theta_2^0(\text{XZ}) = 0^\circ$ ,  $\theta_3^0 = 34.3^\circ$ ,  $\theta_4^0(\text{XZ}) = \theta_5^0(\text{XZ}) = 0^\circ$ ,  $\theta_6^0(\text{XZ}) = 3.4^\circ$ . *b.* Factor used to scale the force constants of each set of knowledge-based potentials so that the persistence length of a mixed-sequence homopolymer is  $\sim 500 \text{ \AA}$  (see text).

Dimeric repeat	$a^a$ (“complete”, $\text{\AA}$ )	$a^a$ (“refined”, $\text{\AA}$ )
$\xi^b$	0.85	0.50
AA·TT	150	395
AG·CT	359	461
GG·CC	298	405
GA·TC	423	395
AC·GT	735	625
AT·AT	193	245
GC·GC	562	454
CA·TG	276	391
TA·TA	87	217
CG·CG	140	269
Mixed sequence	500.5	500.3
Ideal DNA	500.2	

Table 4.2: Sequences and ring-closure properties of representative 94-bp DNA molecules. *c*: “complete” samples; *r*: “refined” samples. Sequences and measured *J* factors taken from Refs. [24, 25]; predicted *J* factor of ideal DNA taken from Ref. [2].

DNA <sup>a</sup>	Base-pair sequence	$r_{0,1}^c$	$r_{0,1}^r$	$\log J$
TA-94	ggccgggtcgTAgcaagctcTAgcaccgct TAAacgcacgTAcgcgctgt cTAccgcgtt tTAaccgcca aTAggatTActTAcTAgtctcTAc	206	245	-9.0
5S-94	ggccgacatccctgaccctt TAAaTAgtT Aactttcatcaagcaagagc cTAcgaccaT Accatgctga aTATAccggt tctcgtccgatcac	208	245	-9.3
E6-94	ggccgtgcgcacgaaatgcTAtgccgaaga ttggatggacatgctTATAa aaggaatccc cagaggTAatccttgatctgatgatgcc gcc	211	243	-10.1
E8-94	ggccgtgcgTAgaaTActt tTAttTAtcg cctccacggtgctgatcccc tgtgctgtg gccgtgtTAtctcgagtTAgTAcgacgtcc gcc	218	247	-10.3
E13-94	ggccgtgctg tcggTAaggtgcgatggcct catcaaggcgccaTATAaga tcactcgTAg tgaaaaccTAcccttcattT Aatgttgatc gcc	211	244	-10.2
CA-94	ggccgtcccagcaagctccaggtgcgcca aacggctgcagacgcctgc acggcagccc aagcgcaccc agagccccctctccgaattcacc	231	251	-10.2
Mixed-sequence	xxxxxxxxxxxxxxxxxxxxxxxxxxxx	262	262	
Ideal-94	—————	259		-11.6

## References

- [1] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. Z., and Widom, J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
- [2] Czapla, L., Swigon, D., and Olson, W. K. (2006) Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theor. Comp.*, **2**, 685–695.
- [3] Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Interscience Publishers, New York.
- [4] Balasubramanian, S., Xu, F., and Olson, W. K. (2009) DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. *J. Mol. Biol.*, **96**(6), 2245–2260.
- [5] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R., and Schneider, B. (1992) The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
- [6] Li, Y. (2006) *Protein DNA Interaction from the Nucleic Acid Perspective*, Ph.D thesis, Rutgers University.
- [7] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- [8] Olson, W. K., Colasanti, A. V., Li, Y., Ge, W., Zheng, G., and Zhurkin, V. B. (2006) DNA simulation benchmarks as revealed by X-ray structures, In J. Sponer and F. Lankas, Editors, *Computational Studies of RNA and DNA*, Page 235-237, Springer, Dordrecht, The Netherlands.
- [9] Marky, N. L. and Olson, W. K. (1994) Configurational statistics of the DNA duplex: Extended generator matrices to treat the rotations and translations of adjacent residues. *Biopolymers*, **34**, 109–120.
- [10] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. H., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., and Olson, W. K., et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 781–791.
- [11] Coleman, B. D., Olson, W. K., and Swigon, D. (2003) Theory of sequence-dependent dna elasticity. *J. Chem. Phys.*, **118**, 7127–7140.

- [12] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**(19), 11163–11168.
- [13] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986) *Numerical Recipes in C*, New York, Cambridge University Press.
- [14] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
- [15] Flory, P. J. (1973) Moments of the end-to-end vector of a chain molecule, its persistence and distribution. *Proc. Natl. Acad. Sci., U.S.A.*, **70**, 1819–1823.
- [16] Olson, W. K., Marky, N. L., Jernigan, R. L., and Zhurkin, V. B. (1993) Influence of fluctuations on DNA curvature. A comparison of flexible and static wedge models of intrinsically bent dna. *J. Mol. Biol.*, **232**, 530–554.
- [17] Flory, P. J., Suter, U. W., and Mutter, M. (1976) Macrocyclization equilibria. I. Theory. *J. Am. Chem. Soc.*, **98**, 5733–5739.
- [18] Richmond, T. J. and Davey, C. A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
- [19] Kulic, I. M. and Schiessel, H. (2003) Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure*, **14**, 1527–1534.
- [20] Gorin, A. A., Zhurkin, V. B., and Olson, W. K. (1995) B-DNA twisting correlates with base pair morphology. *J. Mol. Biol.*, **247**, 34–48.
- [21] Zhurkin, V. B., Lysov, Y. P., and Ivanov, V. I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, **6**, 1081–1096.
- [22] Simpson, R. T. and Stafford, D. W. (1983) Structural features of a phased nucleosome core particle. *Proc. Natl. Acad. Sci., U.S.A.*, **80**, 51–55.
- [23] Lowary, P. T. and Widom, J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.
- [24] Cloutier, T. E. and Widom, J. (2004) Spontaneous sharp bending of double-stranded DNA. *Mol. Cell.*, **14**, 355–362.
- [25] Cloutier, T. E. and Widom, J. (2005) DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc. Natl. Acad. Sci., U.S.A.*, **102**, 3645–3650.
- [26] Hayes, J. J., Tullius, T. D., and Wolffe, A. P. (1990) The structure of DNA in a nucleosome. *Proc. Natl. Acad. Sci., U.S.A.*, **87**, 7405–7409.
- [27] Thastrom, A., Bingham, L. M., and Widom, J. (2004) Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.*, **338**, 695–708.

- [28] Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K., and Zhurkin, V. B. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**(3), 725–738.

## Chapter 5

### DNA simulation: How stiff is DNA?

#### 5.1 Introduction

The mechanical properties of DNA play a key role in its biological processing, determining how the long, thin, double-helical molecule responds to the binding of proteins and functions in confined spaces within a cell. Spectroscopic tools developed over the years to measure the distances between small, covalently linked chemical labels — e.g., fluorescent dyes [1, 2, 3] and nitroxide spin labels [4, 5, 6] — provide some of the best available estimates of the natural structure and deformability of DNA in solution. The observed intramolecular distances between these probes mirror the known helical pathway of DNA, but the fluctuations in distances detected in duplexes of a few helical turns substantially exceed those expected from the classic helical wormlike chain model [7] used to characterize the polymeric properties of DNA.

Recent studies of the small-angle X-ray scattering between gold nanocrystals attached to opposing ends of short DNA duplexes (Fig. 5.1A) reveal smaller variations in the distances between the tethered probes [8, 9]. The variation in distance with chain length, however, is greater than the uptake in end-to-end fluctuations expected from the simple geometric model used to interpret the data. The apparent discrepancy — attributed to intrinsic stretching fluctuations in DNA appreciably larger than those deduced from either single-molecule force-extension measurements [10, 11, 12] or analyses of high-resolution structures [13, 14] — has stimulated our interest in this system. Oversimplified models of polymeric behavior can sometimes be misleading [15]. Interpretation of the observed properties of even a short, chemically labeled duplex requires a model that conforms closely and in an identifiable manner with the structural and deformational characteristics of both the DNA and the tethered probes. A simple model

with direct control of the structural components can offer useful insights into the molecular system.

Here we investigate the subtle relationship between the local elastic properties of DNA, the fluctuations of tethered gold nanocrystals, and the overall configurational properties of short DNA duplexes of the type recently characterized by small-angle X-ray scattering. We explore the system directly by combining Gaussian sampling [16] of the likely spatial arrangements of the DNA base-pair steps with Metropolis-Monte-Carlo simulations [17] of movements in the tether. We take advantage of the multiple ‘time-step’ Monte-Carlo approach pioneered by Berne and co-workers [18] to treat these two very different types of molecular movement. We examine the chain-length-dependent fluctuations in end-to-end extension associated with the twisted wormlike chain behavior of double-helical DNA. We also consider the contributions of rigid and flexible tethers to the distances between gold nanocrystals on the ends of short, fluctuating duplexes. Finally, we compare the predicted spread of distances of different DNA-tether models with the observed fluctuations and present our findings in the context of the experimental data.

## 5.2 Methods

### 5.2.1 DNA model

DNA is modeled at the level of base-pair steps in terms of six rigid-body parameters: three angular variables termed tilt, roll, and twist and three variables called shift, slide, and rise with dimensions of distance [19]. A configuration of DNA is defined by the set of parameters at each base-pair step and is said to be relaxed when all parameters adopt their preferred equilibrium values.

The potential governing the fluctuations in base-pair steps is assumed to follow a quadratic expression of the form:

$$\Psi = \frac{1}{2} \sum_{i=1}^6 \sum_{j=1}^6 f_{ij} \Delta\theta_i \Delta\theta_j, \quad (5.1)$$

where the  $\Delta\theta_i$  are deviations of the base-pair-step parameters  $\theta_i$  from their intrinsic value  $\theta_i^0$ , and the  $f_{ij}$  are ‘stiffness’ constants. Local sequence-dependent structure and

deformability in DNA can be incorporated in the  $\theta_i^0$  and  $f_{ij}$  [13].

If the  $\Delta\theta_i$  at base-pair step  $n$  are collected in the  $6 \times 1$  vector  $\Delta\Theta_n$  and the  $f_{ij}$  in the  $6 \times 6$  force-constant matrix  $\mathbf{F}_n$ , eqn (7.1) takes the form:

$$\Psi_n = \frac{1}{2} \Delta\Theta_n^T \mathbf{F}_n \Delta\Theta_n, \quad (5.2)$$

with the total deformation energy  $U$  of DNA equal to the sum of the  $\Psi_n$  over all  $N$  base-pair steps:

$$U = \sum_{n=1}^N \Psi_n. \quad (5.3)$$

### 5.2.2 Gaussian sampling

We take advantage of the quadratic form of the energy in eqn (7.1) and the assumption that the base-pair steps fluctuate independently of one another to collect a Boltzmann distribution of dimeric states. We achieve this, as described elsewhere [16], by diagonalizing  $\mathbf{F}$  and sampling linear combinations of base-pair-step parameters along the principal axes of dimeric deformation.

We consider several simple models of DNA. We first treat the double helix as an ideal, inextensible, naturally straight molecule with an intrinsic helical repeat of 10.5 bp/turn. The tilt and roll angles are accordingly null and the twist is  $\sim 34.3^\circ$  in the rest state ( $\theta_1^0 = \theta_2^0 = 0$ ;  $\theta_3^0 = 34.3^\circ$ ). The translational parameters are ‘fixed’ at their intrinsic values ( $\theta_4^0 = \theta_5^0 = 0$ ;  $\theta_6^0 = 3.4\text{\AA}$ ) by the assignment of large force constants. The root-mean-square fluctuations in tilt are equated to those in roll, i.e.,  $\langle \Delta\theta_1^2 \rangle^{1/2} = \langle \Delta\theta_2^2 \rangle^{1/2}$ , so that bending is isotropic, and assigned values of  $4.84^\circ$  corresponding to a persistence length  $a = 2\Delta s / (\langle \Delta\theta_1^2 \rangle + \langle \Delta\theta_2^2 \rangle)$  of nearly  $500\text{\AA}$  (if  $\Delta s$ , the per residue base-pair displacement, is taken as  $3.4\text{\AA}$ ). The fluctuations in twist are assumed to be independent of the bending deformations so that the model corresponds to the classic twisted wormlike chain representation of DNA [7]. The assumed fluctuations in twist  $\langle \Delta\theta_3^2 \rangle^{1/2} = 4.09^\circ$  correspond to a global twisting constant  $C = k_B T / \langle \Delta\theta_3^2 \rangle$  somewhat larger in magnitude than the global bending constant  $A$ , i.e.,  $C/A = 1.4$ , where  $A = ak_B T$ . This choice of  $C$  is compatible with measurements of the equilibrium topoisomer distributions of DNA minicircles and the fluorescence depolarization anisotropy of ethidium bromide



molecules intercalated in DNA minicircles [20, 21].

We also consider more realistic representation that incorporate the known conformational properties of the DNA base-pair steps in knowledge-based elastic expressions of form of eqn. (7.1). The latter models allow for well-known features of DNA deformability such as anisotropic bending [22], the coupling of bending and shearing deformations [23], chain extensibility [24], etc., as well as the subtle differences in deformability among different base-pair steps. Thus, local chain units can stretch as well as bend and twist. The force constants, which are derived from the covariance of step parameters in high-resolution structures [13], are scaled such that a mixed-sequence chain, with all 16 base-pair steps equally weighted [14], has the same persistence length as the ideal DNA model. Simulated sequences with a high proportion of pyrimidine-purine steps are more deformable and those with a high proportion of purine-pyrimidine or purine-purine steps are stiffer than the ideal and mixed-sequence chains. For simplicity, we ignore the small, sequence-dependence differences in intrinsic step parameters and the effects of adjacent nucleotides, which have almost no effect on the extension of short DNA chains.

### 5.2.3 DNA reconstruction

Recovery of atomic information from the DNA base-pair-step parameters is essential for understanding and visualizing the modeled fluctuations in double-helical structure. Moreover, the computational treatment of tethered gold nanocrystals requires knowledge of the coordinates of the points to which the labels are attached. We thus make use of the *rebuild* algorithm from the 3DNA software package [25, 26] to construct atomic models of DNA from the rigid-body parameters. We ignore potential fluctuations in base-pair geometry, assuming that the four base pairs — A·T, T·A, G·C, C·G — adopt standard Watson-Crick arrangements [27].

Generation of an atomic-level model necessitates the transformation of the coordinate frame on each base pair (Fig. 5.1C) into the global DNA reference frame. This is achieved using a serial product of matrices  $\mathbf{A}_n$  that incorporate the  $3 \times 1$  displacement vector  $\mathbf{r}_n$  and the  $3 \times 3$  rotation matrix  $\mathbf{T}_{n,n+1}$ , which relate coordinate frames on

successive base pairs  $(n, n + 1)$ :

$$\mathbf{A}_{1:N} = \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_{N-1} \mathbf{A}_N, \quad (5.4)$$

where

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{T}_{n,n+1} & \mathbf{r}_n \\ \mathbf{0} & 1 \end{bmatrix}. \quad (5.5)$$

The dependence of  $\mathbf{T}_{n,n+1}$  and  $\mathbf{r}_n$  on the base-pair-step parameters  $\Theta_n$  follows the formulation introduced by Zhurkin *et al.* [22] and further developed by El Hassan and Calladine [28].

Determination of the atomic coordinates of the base pairs requires the additional transformation of the coordinate vector of each atom  $\mathbf{v}_s = [x_s, y_s, z_s]^T$  from the standard base-pair reference frame to the global DNA frame. For example, the coordinates of atoms on base-pair  $n + 1$  can be expressed in the frame of base-pair  $n$  by the following transformation:

$$\mathbf{v}_n = \mathbf{T}_{n,n+1} \mathbf{v}_s + \mathbf{r}_n. \quad (5.6)$$

#### 5.2.4 DNA end-to-end distance and contour length

The DNA end-to-end vector  $\mathbf{r}_{1:N}$ , which joins the centers of the first and last base pairs, is accumulated in the global generator matrix  $\mathbf{A}_{1:N}$  described in eqn (5.4):

$$\mathbf{r}_{1:N} = \begin{bmatrix} \mathbf{I}_3 & 0 \end{bmatrix} \mathbf{A}_{1:N} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \quad (5.7)$$

Here  $\mathbf{I}_3$  is the identity matrix of order three and the  $\mathbf{0}$ 's are null matrices of orders necessary to fill the  $3 \times 4$  premultiplication and  $4 \times 1$  postmultiplication vectors. The DNA end-to-end distance  $r_{\text{DNA}}$  is the magnitude of  $\mathbf{r}_{1:N}$ . The variance in the DNA end-to-end distance  $\langle \delta r_{\text{DNA}}^2 \rangle$  is given by the standard difference of averages,  $\langle \delta r_{\text{DNA}}^2 \rangle = \langle r_{\text{DNA}}^2 \rangle - \langle r_{\text{DNA}} \rangle^2$ .

The DNA contour length  $L_{\text{DNA}}$  is the sum of the distances between sequential base pairs:

$$L_{\text{DNA}} = \sum_{n=1}^N |\mathbf{r}_n|, \quad (5.8)$$

where  $\mathbf{r}_n$  is the displacement vector stored in the generator matrix associated with base-pair step  $n$ . The variance in the contour length  $\langle \delta L_{\text{DNA}}^2 \rangle$  is obtained, like that for  $\langle \delta r_{\text{DNA}}^2 \rangle$ , from the mean-square and average values of  $L_{\text{DNA}}$ .

The end-to-end distance and contour length are identical if the DNA is perfectly straight. Twisting and stretching a straight DNA do not affect this equality, but bending and shearing lead to differences between the two measurements. Thus, if gold nanocrystals are tethered to the ends of DNA along the lines discussed below, the distance between the centers of the gold particles  $r_{\text{Au}}$  and the nanocrystal-DNA contour length  $L_{\text{Au}}$  will differ, given that the tether may bend and may not lie along the DNA helical axis in its equilibrium rest state.

### 5.2.5 Tether model

The gold nanocrystals tethered to the 3'-ends of DNA in recent small-angle X-ray-scattering experiments [8, 9] are small spherical constructs ( $\sim 75$  atoms) attached, via sulfur, to a three-carbon thiol that is connected in turn to DNA through a phosphodiester linkage (Fig. 5.1B). The spatial positions of the nanocrystals with respect to the DNA bases thus depend upon the internal coordinates (bond lengths, valence angles, and dihedral angles) of both the tether and the sugar-phosphate backbone.

The Cartesian coordinates of the tether are determined with a simple build-up procedure that starts with the approximate coordinates of three successive sugar atoms (C2', C3', O3') generated in the reconstruction of DNA from base-pair-step parameters [25, 26]. Given these coordinates ( $\mathbf{v}_{n-2}$ ,  $\mathbf{v}_{n-1}$ ,  $\mathbf{v}_n$ ), the spatial position  $\mathbf{v}_{n+1}$  of a fourth atom  $n+1$  can be determined from knowledge of (i) the length  $b$  of the chemical bond that joins atom  $n$  to atom  $n+1$ , (ii) the magnitude of the valence angle  $\theta$  formed by atoms  $n-1$ ,  $n$ , and  $n+1$ , and (iii) the value of the dihedral angle  $\varphi$  described by atoms  $n-2$ ,  $n-1$ ,  $n$ , and  $n+1$ .

The coordinates of successive atoms are obtained by iteration of the following procedure. First, the components of  $\mathbf{v}_{n+1}$  are defined by the expression:

$$\mathbf{v}_{n+1} = \mathbf{v}_n + \mathbf{R}_{n-1,n} \mathbf{b}_{n,n+1}, \quad (5.9)$$

where  $\mathbf{R}_{n-1,n}$  is a  $3 \times 3$  matrix that converts a local reference frame associated with atoms  $n-2$ ,  $n-1$ , and  $n$  into the global frame of the molecule and  $b_{n,n+1}$  is a representation of the bond vector between atoms  $n$  and  $n+1$  in the assumed local frame.

The components of  $\mathbf{R}_{n-1,n}$  are given by:

$$\mathbf{R}_{n-1,n} = \begin{bmatrix} \mathbf{x}_{n-1,n} & \mathbf{y}_{n-1,n} & \mathbf{z}_{n-1,n} \end{bmatrix}, \quad (5.10)$$

where  $\mathbf{z}_{n-1,n}$  is a unit vector along the bond that connects atoms  $n-1$  and  $n$ , i.e.,  $\mathbf{z}_{n-1,n} = (\mathbf{v}_n - \mathbf{v}_{n-1})/|\mathbf{v}_n - \mathbf{v}_{n-1}|$ ,  $\mathbf{y}_{n-1,n}$  is the unit normal to the plane containing atoms  $n-2$ ,  $n-1$ , and  $n$ , i.e.,  $\mathbf{y}_{n-1,n} = (\mathbf{z}_{n-2,n-1} \times \mathbf{z}_{n-1,n})/|\mathbf{z}_{n-2,n-1} \times \mathbf{z}_{n-1,n}|$ , and  $\mathbf{x}_{n-1,n}$  is defined by the right-handed rule, i.e.,  $\mathbf{x}_{n-1,n} = \mathbf{y}_{n-1,n} \times \mathbf{z}_{n-1,n}$ .

The components of  $\mathbf{b}_{n,n+1}$  in the local coordinate frame are given by the product:

$$\mathbf{b}_{n,n+1} = \mathbf{R}_z(\varphi)\mathbf{R}_y(\pi - \theta)\mathbf{b}, \quad (5.11)$$

where  $\mathbf{R}_\mathbf{u}(\zeta)$  is a matrix describing the rotation of a vector through an angle  $\zeta$  about axis  $\mathbf{u} = [u_1, u_2, u_3]$ ,  $\mathbf{z} = [0, 0, 1]$  and  $\mathbf{y} = [0, 1, 0]$  are the chosen axes of rotation, and  $\mathbf{b} = [0, 0, b]$  is the representation of the bond between atoms  $n$  and  $n+1$  in a local frame associated with atoms  $n-1$ ,  $n$ , and  $n+1$ . The elements of  $\mathbf{R}_\mathbf{u}(\zeta)$  in this expression follow the standard definition [29]:

$$r_{\nu\mu} = (1 - \cos \zeta)u_\nu u_\mu - \sin \zeta \sum_{\kappa} \varepsilon_{\nu\mu\kappa} u_\kappa + \cos \zeta \delta_{\nu\mu}, \quad (5.12)$$

where  $\delta_{\nu\mu}$  is the Kronecker delta, i.e.,  $\delta_{\nu\mu} = 1$  when  $\nu = \mu$ ,  $\delta_{\nu\mu} = 0$  when  $\nu \neq \mu$ , and  $\varepsilon_{\nu\mu\kappa} = \pm 1$  when  $\nu, \mu, \kappa$  is an even or odd permutation of 1, 2, 3, respectively, and vanishes otherwise.

### 5.2.6 DNA-tether interactions

We allow the tether to undergo small conformational fluctuations and large structural rearrangements via random and specific variations in backbone dihedral angles. The potential  $V$  associated with these changes is given by a standard summation of torsional and nonbonded terms [30]:

$$V = \sum_{\text{dihedrals}} \frac{K_\phi}{F} [1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[ \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right], \quad (5.13)$$

which is evaluated over all pairwise combinations of movable particles, including the gold nanoassembly.

The local moves of the tether also include fluctuations, consistent with experiment [8], in the virtual distance  $b_{\text{S-Au}}$  between the sulphur atoms on the tether and the centers of the gold nanocrystals. We assume the stretching energy to be quadratic and assign an elastic constant  $k_b$  equal to  $(kT/2)\langle\delta b_{\text{S-Au}}^2\rangle^{-1}$ , where  $\langle\delta b_{\text{S-Au}}^2\rangle^{1/2}$  is the observed root-mean-square deviation in the virtual-bond distance,  $k$  the Boltzmann constant, and  $T$  the temperature in Kelvin.

We compute the total non-bonded interaction  $E$  between the DNA and tethers using a Lennard-Jones potential over all atom pairs on the two fragments, and an electrostatic potential between the gold nanocrystals and DNA phosphate groups.

$$E = \sum_{\{m,n\}} \left[ \frac{A_{mn}}{r_{mn}^{12}} - \frac{B_{mn}}{r_{mn}^6} \right] + \sum_{\{\text{Au,P}\}} \frac{q_{\text{Au}}q_{\text{P}}}{\varepsilon(r_{\text{Au-P}} - r_{\text{Au}})}, \quad (5.14)$$

The net negative charge on the nanocrystal [8] (here taken to be  $-0.2$  esu) is located at the center of the spherical gold assembly (an extended atom of radius  $r_{\text{Au}} = 7\text{\AA}$ ) and that on DNA on the P atoms with a value ( $-0.24$  esu) in accordance with the predictions of counterion condensation theory for a B-DNA polyelectrolyte in monovalent salt solution [31].

The atomic parameters used in eqns (5.13-5.14) to describe the non-bonded interactions of DNA and tether atoms are taken from the AMBER 10 force field [32]. The effects of solvent on electrostatic interactions are treated implicitly with the dielectric constant  $\varepsilon$  assigned a value of 80.

### 5.2.7 Monte-Carlo simulation

We simulate the system in three stages using a multiple ‘time-step’ Monte-Carlo approach [18]. First, we generate a random configuration of DNA using Gaussian sampling at each base-pair step. This is a straightforward process, which allows for fast rearrangement of the base pairs [16]. Second, based on the sampled DNA configuration, we simulate the motions of the tethers, which are rooted in the DNA, using the Metropolis-Monte-Carlo method [17] in combination with the energy term in eqn (5.13). This step

is repeated several times after each move of the first type, so that the tethers undergo sufficient rearrangement. Finally, we accept or reject the configuration generated in the first two stages of computation by comparing the DNA-linker interactions obtained with eqn (5.14) with that present before the simulated move, again using the Metropolis algorithm.

### 5.3 Results and discussion

#### 5.3.1 Global fluctuations of DNA

We start by examining the fluctuations in end-to-end extension associated with the twisted wormlike chain behavior of short double-helical DNA. We investigate a series of unlabeled molecules of the same chain lengths (10, 15, 20, 25, 30, 35bp) considered in recent small-angle X-ray-scattering studies [8, 9]. Interpretation of the solution properties of the chemically labeled duplexes used in these and related experiments [1, 2, 3, 4, 5, 6] requires knowledge of the DNA motions as well as any effects of the tethered labels.

We apply two different models of DNA deformability: the first an ideal, inextensible, twisted wormlike chain with the intrinsic structure and elastic parameters described in Methods and the second a naturally straight, mixed-sequence chain subject to the fluctuations in base-pair steps seen in high-resolution structures and scaled to yield a persistence length of  $\sim 500\text{\AA}$ .

As expected, the average end-to-end distances  $\langle r_{\text{DNA}} \rangle$  (points connected by dashed lines in Fig. 5.2A) are slightly smaller than the mean contour lengths  $\langle L_{\text{DNA}} \rangle$  (points connected by solid lines) of the fluctuating DNA molecules. The differences between  $\langle L_{\text{DNA}} \rangle$  and  $\langle r_{\text{DNA}} \rangle$  are smaller for the inextensible, ideal chain (filled-in circles) compared to the ‘realistic’ knowledge-based model (open squares) that allows for the displacement (primarily shearing) of adjacent base pairs. The nearly identical values of  $\langle r_{\text{DNA}} \rangle$  for the two types of chains reflect the similar persistence lengths in the models.

In contrast to published expectations [8], the variance in DNA end-to-end distance shows a quadratic dependence on chain length, with consistently greater fluctuations in

global structure for the more ‘realistic’ model compared to the ideal chain (Fig. 5.2B). The uptake of radial fluctuations differs in longer chains (see Fig. 5.7). The dependence of  $\langle \delta r_{\text{DNA}}^2 \rangle$  on chain length is roughly linear over the range 500-1600 bp and levels off to a constant value at the very long chain lengths ( $\sim 2500\text{bp}$ ) where the simulated double helix is known to exhibit random-coil behavior [33]. The variance in contour length  $\langle \delta L_{\text{DNA}}^2 \rangle$  shows a linear dependence on chain length, with the desired near-zero slope for the simulated, inextensible model and a slope of  $0.08 \text{Å}^2/\text{bp}$  for the mixed-sequence chain.

### 5.3.2 Effects of rigid tethers

We next consider the contributions of two kinds of rigid tethers to the end-to-end distances between gold nanocrystals attached to a short (15-bp), fluctuating DNA duplex. Here the DNA is modeled as an ideal, inextensible, twisted wormlike chain and the linkers are fixed in one of two different rigid states: an extended form with dihedral angles ( $\varepsilon = 180^\circ$ ,  $\xi = -90^\circ$ ,  $\alpha = 180^\circ$ ,  $\beta = 180^\circ$ ,  $\gamma = 180^\circ$ ,  $\eta = 180^\circ$ ) selected such that the centers of the nanocrystals are close to the DNA helical axis in the equilibrium rest state and a kinked arrangement with dihedral angles ( $\varepsilon = 180^\circ$ ,  $\xi = 180^\circ$ ,  $\alpha = -60^\circ$ ,  $\beta = 180^\circ$ ,  $\gamma = 180^\circ$ ,  $\eta = 180^\circ$ ) chosen so that centers of the nanocrystals are far from the helical axis in the rest state.

Although the DNA undergoes the same motions in both cases, the distributions of end-to-end distances  $W(r_{\text{Au}})$  differ significantly. The separation between nanocrystals is much larger but the range of distances adopted by the extended linker (Figs. 5.3B,E) is much narrower than that adopted by the kinked tether (Figs. 5.3C,F). The distribution of extended tethers resembles that of DNA alone (Figs. 5.3A,D), i.e., similar shapes. Because the projections of the nanocrystal centers on the terminal base-pair planes roughly coincide with the origins of the base-pair frames, the added chain extension tends to widen the arc of sampled points without significantly altering the highly skewed shape of the end-to-end distribution. The roughly sevenfold increase in radial variance with added chain extension, i.e.,  $\langle \delta r_{\text{Au}}^2 \rangle$  vs.  $\langle \delta r_{\text{DNA}}^2 \rangle$ , is not quite as rapid as that illustrated in Fig. 5.2. The extended linkers add  $\sim 15.5 \text{Å}$ , or the equivalent of 5

rigid base-pair steps to the simulated DNA chain. The rearrangement of sampled points associated with the kinked linker has a drastic effect on the end-to-end distribution, increasing the variance by nearly two orders of magnitude compared to that of DNA alone. The shape of the latter distribution more closely resembles a Gaussian distribution with relatively symmetric tails on either side of the most probable end-to-end separation.

In order to gain a better understanding of how the tethered nanocrystals, although rigidly attached to DNA, contribute to the variance of the system as a whole, we studied the interdependence of the DNA end-to-end distances  $r_{\text{DNA}}$  and nanocrystal end-to-end distances  $r_{\text{Au}}$ . The two types of distances are highly correlated when the nanocrystals are bound to extended tethers (Fig. 5.4A) and uncorrelated when bound to the kinked tethers (Fig. 5.4B). The spread of data supports the qualitative rationale presented above.

In contrast to the nanocrystal centers attached to DNA via extended linkers, which build up symmetrically on the two ends of the fluctuating duplex, those attached via kinked linkers accumulate on one side of the molecule (Figs. 5.3B,C). Thus, the distances between chemical probes attached in the former manner will depend primarily on helical displacement, whereas the separation of probes tethered via kinked linkers will also reflect the helical twist. Furthermore, DNA with extended linkers should show a monotonic increase of end-to-end distance with chain length, whereas those with kinked linkers should exhibit non-monotonic variation. Interestingly, the observed distances between gold nanocrystals attached to short DNA duplexes of increasing chain length show minimal deviation from linearity, but the uptake of variance follows the zig-zag behavior expected of a slightly offset linker [8].

We also examined the global bending of DNA measured by the positions of terminal base pairs and nanocrystal centers with respect to the DNA center, i.e., the origin of the reference frame on the central base pair. The bending angles  $\Gamma_{\text{DNA}}$  described by the DNA base-pair points are roughly equivalent to and linearly correlated with those associated with the nanocrystals linked by extended tethers (Fig. 5.4C). The values of  $\Gamma_{\text{Au}}$  span a substantially wider range of values when the nanocrystals are attached to



the kinked tethers (Fig. 5.4D) and show no correlation with the bending of DNA alone. Interestingly, the enhancement in variance brought about by the attachment of rigid tethers varies with DNA chain length (Fig. 5.5). That is, the difference  $\langle \delta r_{\text{Au}}^2 \rangle - \langle \delta r_{\text{DNA}}^2 \rangle$  grows with chain length despite the fixed arrangement of the nanoparticles with respect to terminal base pairs. The tethers thus contribute different levels of intrinsic variance to the distances between gold nanocrystals, contrary to the assumptions [8] that the contribution of the tether to the variance is fixed and that the variance in end-to-end length is directly proportional to chain length. Although the computed variation in  $\langle \delta r_{\text{Au}}^2 \rangle - \langle \delta r_{\text{DNA}}^2 \rangle$  depends in part on the long-range electrostatic interactions between gold centers and DNA included in the simulations, the chain-length-dependent growth in the difference persists in neutral systems. The observed values reflect the dependence of the end-to-end distance of the tethered assembly on numerous factors, including the DNA end-to-end distance, the tether lengths, the angles between tethers and DNA, and the ‘torsion’ of the tethers with respect to the DNA axis.

### 5.3.3 Simulated distributions of end-to-end distances

The introduction of slight flexibility in the extended tethers brings the computed distributions of end-to-end distances  $W(r_{\text{Au}})$  in reasonable agreement with those extracted from the small-angle X-ray scattering of gold nanocrystals (Table 5.1) [8, 9]. The ‘stiff’ tethers used in these simulations have a single degree of conformational freedom: the dihedral angle  $\eta$  immediately preceding the thiol-nanocrystal linkage (Fig. 5.1B), which fluctuates in an energy well about its *trans* rest state. The choice of torsional parameters introduced in eqn (5.13) ( $K_\phi = 1.2$ ,  $F = 1$ ,  $n = 1$ ) restricts the sampled angular states to values in the range  $180 \pm 5^\circ$ . The ‘flexible’ tethers introduced in other calculations allow for fluctuations of the same magnitude in all seven ( $\varepsilon$ ,  $\xi$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\eta$ ) dihedral angles of the tether.

The average distances  $\langle r_{\text{Au}} \rangle$  between nanocrystals linked via ‘stiff’ tethers to either ideal, inextensible or mixed-sequence chains account for the experimentally reported data (Table 5.1). The predicted spread of distances  $\langle \delta r_{\text{Au}}^2 \rangle$ , although consistent with the observed nonlinear increase in range with chain length, slightly overestimates the

fluctuations observed in 10-15-bp chains and somewhat underestimates the measured variation in 30-35-bp chains, i.e., the dependence of variance on chain length. Variation of the persistence length within the wide range of values (450-490 Å) [34, 35] used to account for the solution properties of long DNA and/or modifications of the treatment of the tether can improve the match with experiment.

The predicted range of separation distances increases if the sequence-dependent deformability of DNA base-pair steps is incorporated in the calculations, i.e., each of the dimers in the double helix obeys a characteristic set of force constants [13]. The greater range of local distortions in the base-pair steps incorporated in the model has very little, if any, effect on the average distances between gold nanoparticles. The elastic constants of the dimers in the specific sequences, however, are lower, on average, than those governing the deformations of the mixed-sequence step, where the contributions of all 16 dinucleotides are equally weighted [14]. These differences in local structural mobility underlie the enhanced variance of the simulated duplexes. In particular, the lateral shearing of adjacent base pairs along their long axes, i.e., fluctuations in slide, and the coupling of these motions with the preferential bending of DNA about the same axis (roll), soften the apparent Young's modulus of 'real' sequence-dependent vs. ideal DNA [24]. Moreover, omission of the fluctuations in slide from the 'real' model reduces the variation in the distances between gold nanocenters substantially. The contributions of lateral shearing to the extension of DNA suggest the underlying structural basis of the published rationalization [8] of the distance fluctuations of short end-labeled DNA in terms of enhanced stretching.

Finally, the distributions of distances between nanocrystals with 'stiff' tethers at the ends of short, ideal, inextensible DNA chains are slightly skewed from ideal Gaussian curves (Fig. 5.6). That is, the modified duplexes tend to shorten rather than lengthen with respect to their most probable extension. The skewness becomes more pronounced with increase in chain length in rough correspondence with experiment (where secondary peaks of shorter chain extension appear in plots of relative abundance derived from the scattering data). The distributions associated with mixed-sequence DNA chains containing 'stiff' tethers roughly coincide with those shown for the ideal chains.

The simulated curves, however, widen and shift to lower values of  $r_{Au}$  when sequence-dependent deformability is considered. The incorporation of ‘flexible’ tethers on ideal DNA models similarly broadens and shifts the end-to-end distributions while concomitantly enhancing the propensity of the chains to shorten. (See the plotted curves in Figs. 5.8-5.10.)

## 5.4 Conclusions

The physical properties of DNA depend upon chain length. The dimensions of chains of a few hundred base pairs are typical of a wormlike coil that bends smoothly and gradually into compact forms [7]. Because the deformations in three-dimensional structure used to account for this behavior are quite limited, short chain fragments are often modeled as rigid rods. As demonstrated herein, this oversimplification misses the key contribution of the natural dimeric flexibility of DNA to the end-to-end properties of chains of only a few helical turns. In particular, there is no need to posit enhanced cylindrical stretching fluctuations as the source of the recently measured quadratic dependence of the variance in DNA end-to-end distance on chain length [8]. Mixed-sequence chains with the more restricted levels of stretching deduced from single-molecule experiments [10, 11, 12] and analyses of high-resolution structures [13, 14] also show a quadratic increase in end-to-end variance with chain length (Fig. 5.2). Indeed, even ideal, inextensible DNA chains limited to isotropic bending and twisting fluctuations exhibit such behavior.

As chain length increases, the deformability in the added base-pair steps opens the range of three-dimensional forms available to the DNA helix. If the ends of the chain are separated by a sufficient number of intervening residues, the duplex exhibits ideal Gaussian (random coil) behavior. Thus, the variance in end-to-end distances is linear in longer chains (over the range of chain lengths where the twisted wormlike coil model is normally fitted to DNA properties) and levels off to a constant when the chain is very long.

Given the restrictions on local base-pair structure, the tethers used to attach chemical

probes to short DNA may contribute to the detected dispersion of chain ‘ends’. For example, the distribution of end-to-end distances is narrower (Fig. 5.3) and more closely correlated (Fig. 5.4) with the distances between terminal base-pair centers for probes that are directed along rather than perpendicular to the helical axis. The spin-labeled sugar-phosphate backbones probed in electron paramagnetic resonance studies [4, 5, 6] fall in the latter category. The distributions of intramolecular distances extracted from such experiments thus reflect their relative helical positioning as well as any distortions imposed by chemical modification of the double-helical structure.

Surprisingly, the contributions of even perfectly ‘stiff’ tethers to the end-to-end dispersion of DNA chain ‘ends’ are nonlinear. That is, the difference in the variance in the distances between chemical probes and the variance in the distances between terminal base pairs increases with chain length (Fig. 5.5).

Fluctuations in the tether conformation also affect the distances and dispersion of chain ‘ends’ (Table 5.1). The average distance between chemical probes decreases and the dispersion increases if the probe lies close to the DNA helical axis in its equilibrium rest state. The distance between probes, however, may increase upon tether deformation if the probe lies far from the DNA axis in its rest state. The greater variance in DNA chain extension detected in fluorescence resonance energy transfer experiments [3] compared to small-angle X-ray-scattering [8, 9] studies may thus reflect the enhanced flexibility of the longer tethers attached to fluorescent dyes compared to those used to link gold nanocrystals to DNA. The enhancement in ‘end’-to-‘end’ variance and differences in average ‘end’-to-‘end’ extension may be even greater if the tethers undergo large rearrangements between different conformational forms.

The sequence-dependent deformability of DNA base-pair steps may also contribute to the measured dispersion of chain ‘ends’ (Table 5.1). Like the effects of tether deformations, these effects are nonlinear and become more pronounced at longer chain lengths. Thus, one can account for the observed distances between the ‘ends’ of short DNA chains in terms of the normal physical properties of the double helix and the chemical linkers used to attach various molecular probes (Fig. 5.6). The ranges of measured

distances provide useful benchmarks for all-atom simulations of DNA. The coarse-grained approach used here does not differentiate among the many ways to fit the measured distances. Rather this work highlights the overlooked contribution of small room-temperature fluctuations on the configurational properties of short DNA duplexes and the importance of small conformational deformations in the interpretation of spectroscopic measurements of DNA chain extension.

Finally, a technical comment [36], which addresses the effect of linker offset on the observed distances between gold nanocrystals, and a response from the authors [37], which includes new data supporting the stretching arguments used originally to account for the build-up of the distances between tethered nanocrystals, appeared after this work was completed. The fluctuations in slide incorporated in our ‘realistic’ models of short DNAs suggest the underlying structural basis for this phenomenological interpretation.

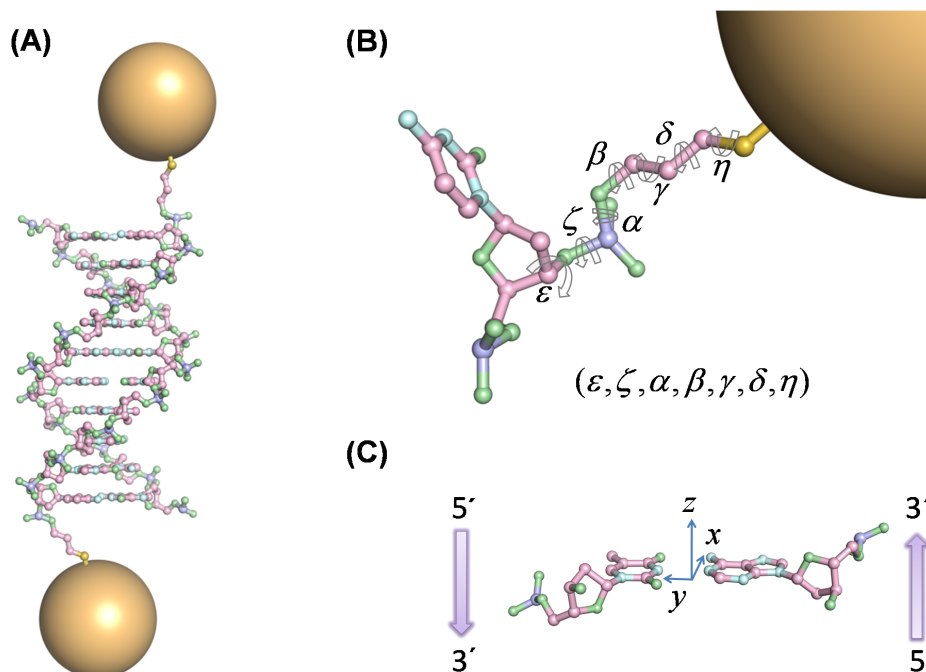


Figure 5.1: (A) Atomic-level representations of a 10-bp DNA duplex with gold nanocrystals (large spheres) attached via short tethers to the 3'-ends of complementary strands. (B) Close-up of the tether highlighting the chemistry and dihedral angles of the linkage. Color-coding denotes atom type: carbon (pink); nitrogen (blue); oxygen (green); sulfur (yellow); phosphorus (violet). (C) Reference frame, called the “middle” frame [25], associated with a DNA base pair. Antiparallel directions of complementary strands are denoted by arrows.

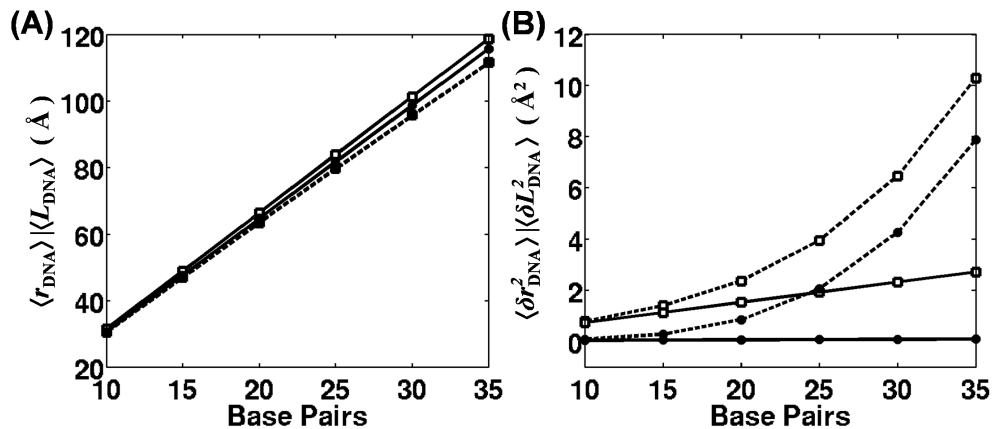


Figure 5.2: Chain-length dependence of (A) the average end-to-end distances and contour lengths (points connected respectively by dashed and solid lines) and (B) the associated variances for two types of DNA — an ideal, inextensible, twisted wormlike chain (filled circles) and a mixed-sequence chain, which is naturally straight in its equilibrium rest state and subject to the deformational properties characteristic of high-resolution structures (open squares). Both models are naturally straight with 10.5 bp/turn and elastic constants scaled to yield a persistence length of  $\sim 500\text{\AA}$  [14].

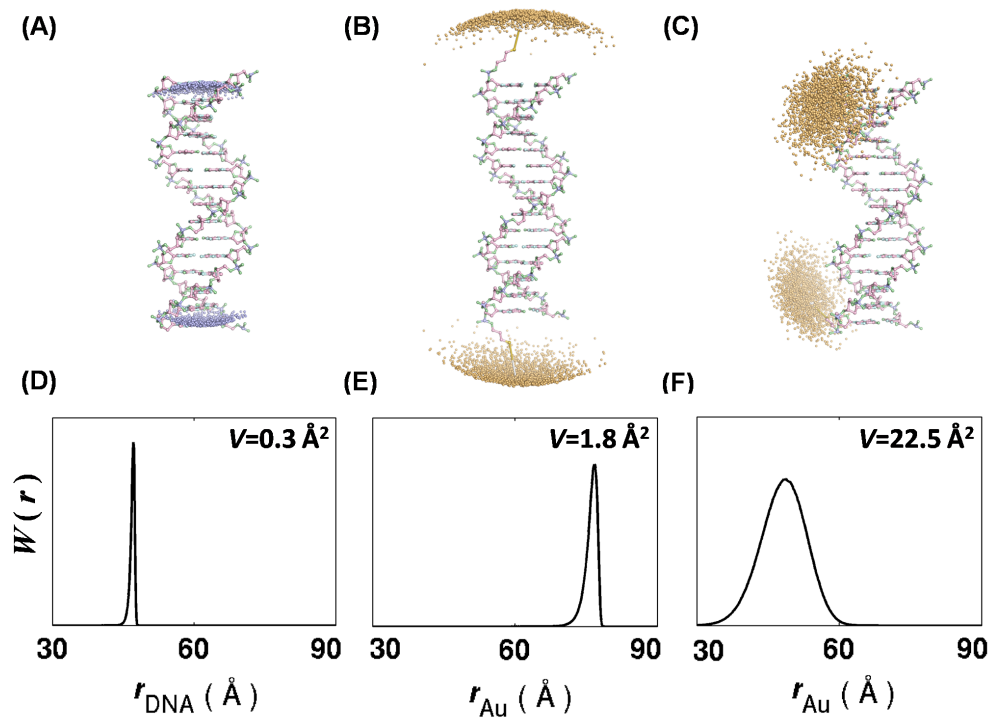


Figure 5.3: Effects of tethers on the end-to-end variance of a 15-bp ideal, inextensible DNA duplex. Scatterplots depict the simulated positions of the centers of (A) terminal base pairs (blue dots) and (B, C) rigidly tethered gold (Au) nanocrystals (orange dots) with respect to the equilibrium structure of DNA. Tethers adopt (B) fully extended and (C) kinked forms. Normalized distributions of the corresponding (D) DNA $\cdots$ DNA and (E, F) Au $\cdots$ Au end-to-end distances ( $r_{\text{DNA}}$  and  $r_{\text{Au}}$ ) that stem from base-pair-step deformations.



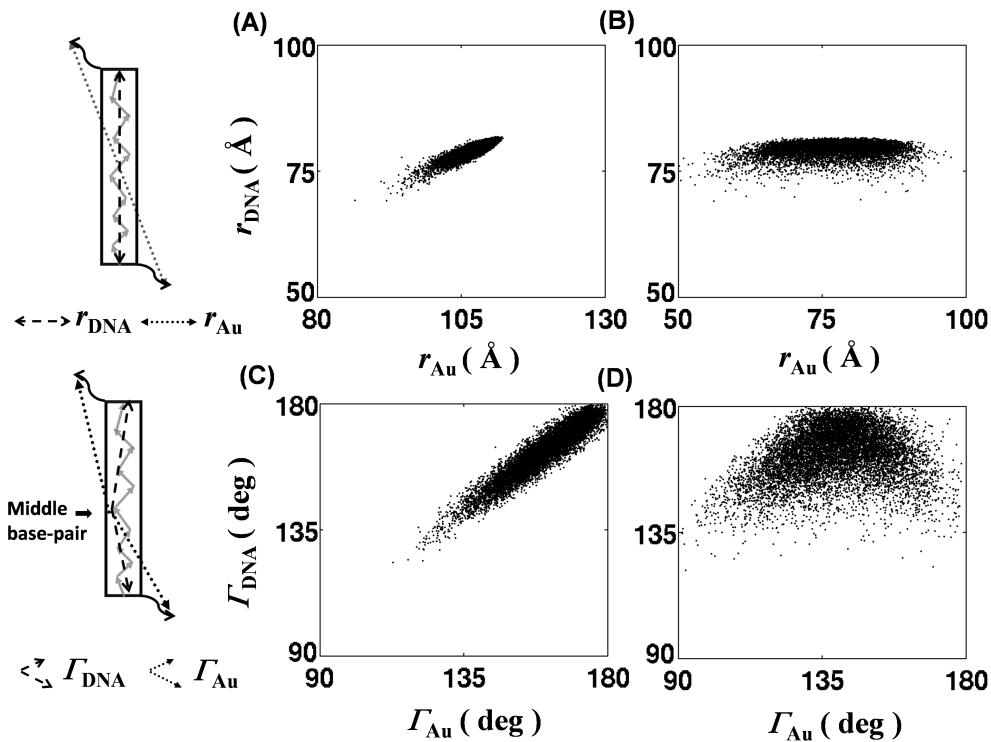


Figure 5.4: Scatter plots of the covariance of DNA and nanocrystal end-to-end distances —  $r_{\text{DNA}}$  and  $r_{\text{Au}}$  — for the (A) extended and (B) kinked tethers considered respectively in Figs. 5.3B/E and C/F and the covariance (C, D) of the corresponding global bending angles,  $\Gamma_{\text{DNA}}$  and  $\Gamma_{\text{Au}}$ , for the same chains.

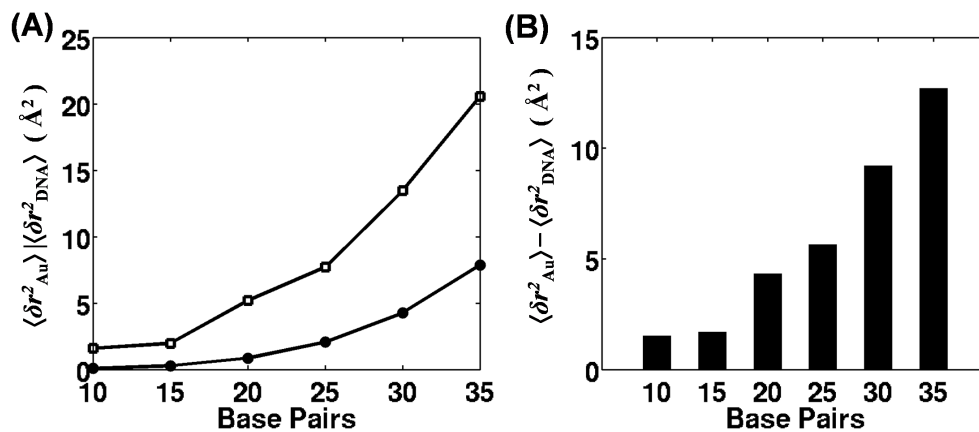


Figure 5.5: Chain-length dependence of (A) the variance of the DNA-DNA and Au-Au end-to-end distances  $\langle \delta r_{\text{DNA}}^2 \rangle$  and  $\langle \delta r_{\text{Au}}^2 \rangle$  (filled and open circles, respectively) of an ideal, inextensible DNA duplex with gold nanocrystals configured along the same lines as Fig. 5.4B and (B) the differences between the two measurements.

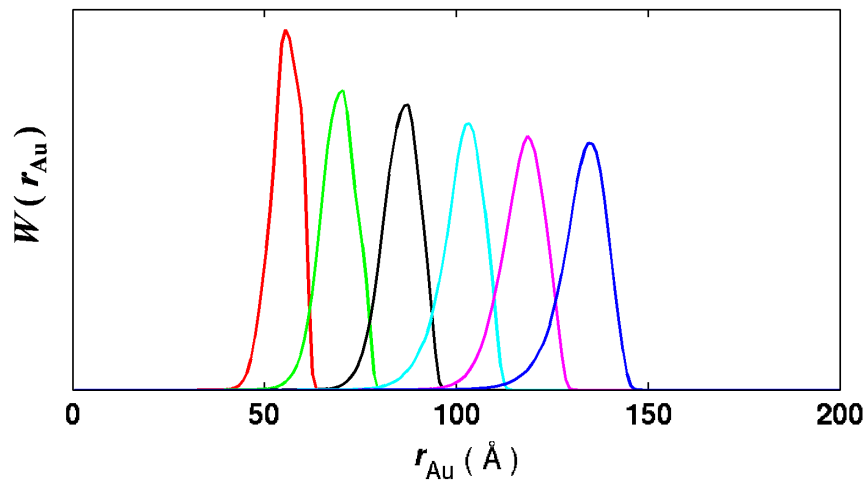


Figure 5.6: Simulated probability density distributions of the distances  $r_{\text{Au}}$  between gold nanocrystals attached via ‘stiff’, extended tethers to the ends of ideal, inextensible DNA duplexes of 10 bp (red), 15 bp (green), 20 bp (black), 25 bp (cyan), 30 bp (magenta), and 35 bp (blue). In contrast to the perfectly rigid tethers considered in Figs. 5.3-5.5, the system modeled here incorporates small fluctuations in the dihedral angle  $\eta = 180 \pm 50^\circ$  and the length  $b_{\text{S-Au}} = 7 \pm 1 \text{ \AA}$  of the virtual bond between sulphur and the center of the gold nanocrystal. See Table 5.1 for the means and variances of these normalized profiles.

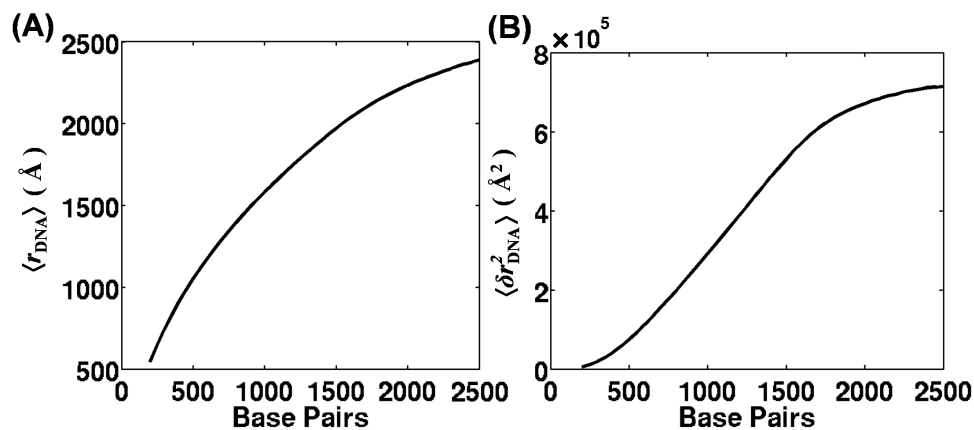


Figure 5.7: Chain-length dependence of (A) the average and (b) the associated variance of the end-to-end distance of long, ideal, inextensible, twisted wormlike chains (more than 200 bp). The DNA model is naturally straight with 10.5 bp/turn and elastic constants scaled to yield a persistence length of  $\sim 500\text{\AA}$ .

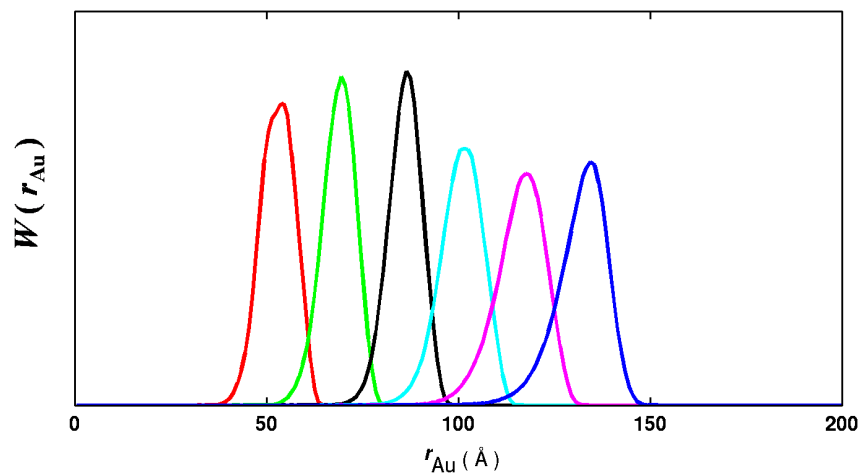


Figure 5.8: Simulated probability density distributions of the end-to-end distance  $r_{\text{Au}}$  between gold nanocrystals attached via ‘stiff’, extended tethers to the ends of mixed-sequence DNA duplexes of 10 bp (red), 15 bp (green), 20 bp (black), 25 bp (cyan), 30 bp (magenta), and 35 bp (blue). See Table 5.1 for the means and variances of these normalized profiles.

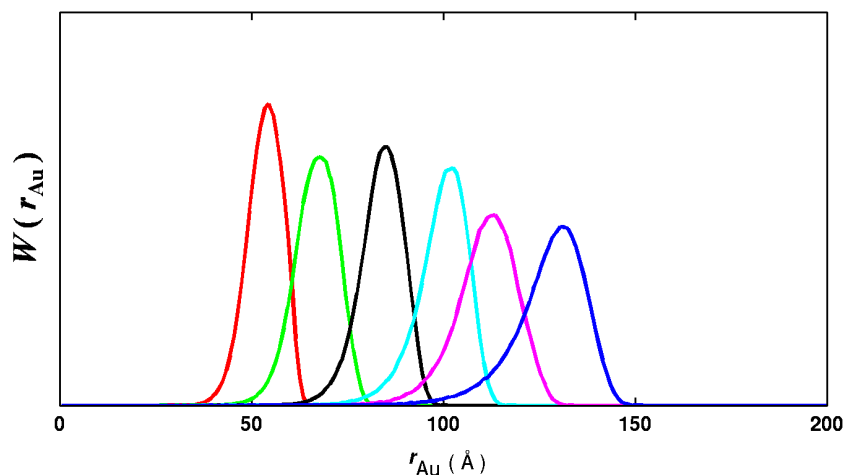


Figure 5.9: Simulated probability density distributions of the end-to-end distance  $r_{Au}$  between gold nanocrystals attached via ‘stiff’, extended tethers to the ends of sequence-dependent DNA duplexes of 10 bp (red), 15 bp (green), 20 bp (black), 25 bp (cyan), 30 bp (magenta), and 35 bp (blue). See Table 5.1 for the means and variances of these normalized profiles.

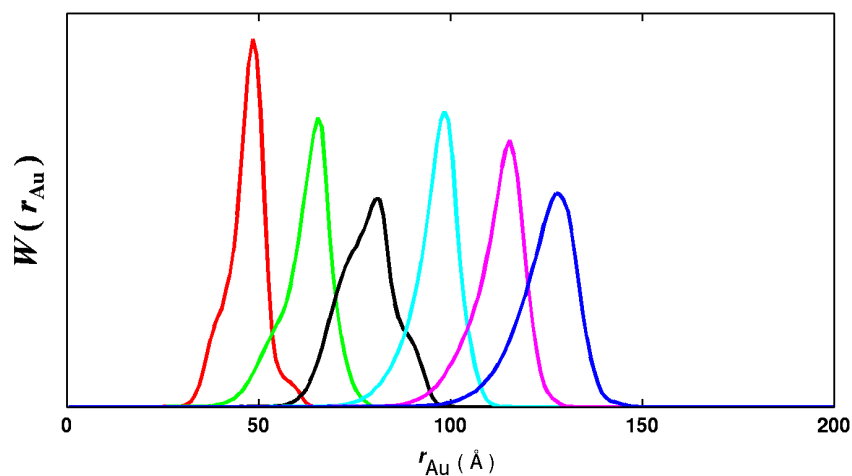


Figure 5.10: Simulated probability density distributions of the end-to-end distance  $r_{Au}$  between gold nanocrystals attached via ‘flexible’, extended tethers to the ends of ideal, inextensible DNA duplexes of 10 bp (red), 15 bp (green), 20 bp (black), 25 bp (cyan), 30 bp (magenta), and 35 bp (blue). See Table 5.1 for the means and variances of these normalized profiles.

Table 5.1: Observed vs. computed distances and fluctuations between gold nanocrystals attached to DNA chains. Abbreviations:  $N$ : DNA chain length (bps); Exp: experimental observation [8]; I+s: ideal, inextensible DNA with ‘stiff’ tethers; M+s: mixed-sequence DNA with ‘stiff’ tethers; S+s: sequence-dependent DNA with ‘stiff’ tethers; I+f: ideal, inextensible DNA with ‘flexible’ tethers.

$N$	Exp	I+s	M+s	S+s	I+f
$\langle r_{\text{Au}} \rangle$					
10	$55.7 \pm 0.3$	$53.9 \pm 0.5$	$53.3 \pm 0.6$	$52.5 \pm 0.6$	$48.5 \pm 0.8$
15	$69.7 \pm 0.4$	$69.6 \pm 0.6$	$69.5 \pm 0.4$	$67.4 \pm 0.5$	$60.5 \pm 2.7$
20	$86.0 \pm 0.4$	$85.2 \pm 0.4$	$85.7 \pm 0.4$	$83.8 \pm 0.6$	$78.5 \pm 1.9$
25	$101.0 \pm 0.5$	$101.1 \pm 1.0$	$101.2 \pm 0.5$	$98.3 \pm 0.6$	$92.6 \pm 2.5$
30	$119.1 \pm 0.6$	$117.5 \pm 0.3$	$117.1 \pm 0.6$	$113.0 \pm 0.7$	$110.8 \pm 1.7$
35	$131.3 \pm 0.7$	$132.3 \pm 0.6$	$133.0 \pm 0.6$	$127.5 \pm 0.7$	$124.6 \pm 1.9$
$\langle \delta r_{\text{Au}}^2 \rangle$					
10	$8.5 \pm 0.6$	$17.5 \pm 2.4$	$20.4 \pm 2.9$	$22.8 \pm 3.7$	$27.4 \pm 6.4$
15	$16.5 \pm 1.1$	$22.1 \pm 1.5$	$21.8 \pm 1.0$	$27.4 \pm 1.3$	$54.0 \pm 21.5$
20	$21.6 \pm 1.4$	$22.9 \pm 3.1$	$24.2 \pm 2.1$	$32.0 \pm 3.3$	$41.8 \pm 7.6$
25	$30.0 \pm 2.0$	$28.4 \pm 3.6$	$30.6 \pm 3.1$	$51.2 \pm 5.4$	$57.7 \pm 16.8$
30	$41.1 \pm 2.7$	$31.8 \pm 1.3$	$33.7 \pm 2.4$	$59.7 \pm 3.0$	$44.0 \pm 8.2$
35	$50.9 \pm 3.4$	$42.2 \pm 3.7$	$41.6 \pm 3.5$	$86.9 \pm 4.3$	$70.6 \pm 16.3$

## References

- [1] Cooper, J. P. and Hagerman, P. J. (1990) Analysis of fluorescence energy transfer in duplex and branched DNA molecules. *Biochemistry*, **29**, 9261–9268.
- [2] Clegg, R. M., Murchie, A. I. H., Zechel, A., and Lilley, D. M. J. (1993) Observing the helical geometry of double-stranded DNA in solution by fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci., U.S.A.*, **90**, 2994–2998.
- [3] Laurence, T. A., Kong, X., Jager, M., and Weiss, S. (2005) Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. *Proc. Natl. Acad. Sci., U.S.A.*, **102**, 17348–17353.
- [4] Schiemann, O., Piton, N., Mu, Y., Stock, G., Engels, J. W., and Prisner, T. F. (2004) A PELDOR based nanometer distance ruler for oligonucleotides. *J. Am. Chem. Soc.*, **126**, 5722–5729.
- [5] Cai, Q., Kusnetzow, A. K., Hubbell, W. L., Haworth, I. S., Gacho, G. P. C., Eps, N. V., Hideg, K., Chambers, E. J., and Qin, P. Z. (2006) Site-directed spin labeling measurements of nanometer distances in nucleic acids using a sequence-independent nitroxide probe. *Nucleic Acids Res.*, **34**, 4722–4730.
- [6] Ward, R., Keeble, D. J., El-Mkami, H., and Norman, D. G. (2007) Distance determination in heterogeneous DNA model systems by pulsed EPR. *ChemBioChem*, **8**, 1957–1964.
- [7] Shimada, J. and Yamakawa, H. (1984) Ring-closure probabilities for twisted worm-like chains. application to DNA. *Macromolecules*, **17**, 689–698.
- [8] Mathew-Fenn, R. S., Das, R., and Harbury, P. A. B. (2008) Remeasuring the double helix. *Science*, **322**, 446–449.
- [9] Mathew-Fenn, R. S., Das, R., Silverman, J. A., Walker, P. A., and Harbury, P. A. B. (2008) A molecular ruler for measuring quantitative distance distributions. *PLoS One*, **3**, e3229.
- [10] Smith, S. B., Cui, Y., and Bustamante, C. (1996) Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science*, **271**, 795–799.
- [11] Wang, M. D., Yin, H., Landick, R., Gelles, J., and Block, S. M. (1997) Stretching DNA with optical tweezers. *Biophys. J.*, **72**, 1335–1346.
- [12] Gore, J., Bryant, Z., Nollmann, M., Le, M. U., Cozzarelli, N. R., and Bustamante, C. (2006) DNA overwinds when stretched. *Nature Chemical Biology*, **442**, 836–839.

- [13] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci., U.S.A.*, **95**, 11163–11168.
- [14] Olson, W. K., Colasanti, A. V., Czapla, L., and Zheng, G. (2008) Insights into the sequence-dependent macromolecular properties of DNA from base-pair level modeling. In G. A. Voth, Editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Page 205-223, Taylor and Francis Group, Boca Raton, FL.
- [15] Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Interscience Publishers, New York.
- [16] Czapla, L., Swigon, D., and Olson, W. K. (2006) Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theor. Comp.*, **2**, 685–695.
- [17] Metropolis, N. A., Rosenbluth, A. W., Rosenbluth, M. N., Teller, H., and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
- [18] Hetenyi, B., Bernacki, K., and Berne, B. J. (2002) Multiple "time step" Monte Carlo. *J. Chem. Phys.*, **117**, 8203–8207.
- [19] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. H., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., and Olson, W. K., et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 781–791.
- [20] Horowitz, D. S. and Wang, J. C. (1984) Torsional rigidity of DNA and length dependence of the free energy of DNA supercoiling. *J. Mol. Biol.*, **173**, 75–91.
- [21] Heath, P. J., Clendenning, J. B., Fujimoto, B. S., and Schurr, J. M. (1996) Effect of bending strain on the torsion elastic constant of DNA. *J. Mol. Biol.*, **260**, 718–730.
- [22] Zhurkin, V. B., Lysov, Y. P., and Ivanov, V. I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, **6**, 1081–1096.
- [23] Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K., and Zhurkin, V. B. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**(3), 725–738.
- [24] Matsumoto, A. and Olson, W. K. (2002) Sequence-dependent motions of DNA: A normal mode analysis at the base-pair level. *Biophys. J.*, **83**, 22–41.
- [25] Lu, X.-J. and Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
- [26] Lu, X.-J. and Olson, W. K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protoc.*, **3**, 1213–1227.



- [27] Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, R. C., Heinemann, U., Lu, X.-J., Neidle, S., and Shakked, Z., et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
- [28] El Hassan, M. A. and Calladine, C. R. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA: a new local calculation scheme. *J. Mol. Biol.*, **251**, 648–664.
- [29] Jeffreys, H. and Jeffreys, B. S. (1946) *Methods of Mathematical Physics*, Cambridge University Press, Cambridge, UK.
- [30] Fox, T. and Kollman, P. A. (1998) Application of the RESP methodology in the parameterization of organic solvents. *J. Phys. Chem. B*, **102**, 8070–8079.
- [31] Manning, G. S. (1978) The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quart. Rev. Biophys.*, **11**(2), 179–246.
- [32] Case, D. A. and Darden, T. A., et al. (2008) AMBER 10. *University of California, San Francisco, CA*,.
- [33] Olson, W. K. (1978) The flexible DNA double helix. I. Average dimensions and distribution functions.
- [34] Hagerman, P. J. (1981) Investigation of the flexibility of DNA using transient electric birefringence. *Biopolymers*, **20**, 1503–1535.
- [35] Vologodskaya, M. and Vologodskii, A. (2002) Contribution of the intrinsic curvature to measured DNA persistence length. *J. Mol. Biol.*, **317**, 205–213.
- [36] Becker, N. B. and Everaers, R. (2009) Comment on "Remeasuring the Double Helix". *Science*, **325**(5940), 538–b.
- [37] Mathew-Fenn, R. S., Das, R., Fenn, T. D., Schneiders, M., and Harbury, P. A. B. (2009) Response to comment on "Remeasuring the Double Helix". *Science*, **325**(5490), 538–c.

## Chapter 6

### Cylindrical view of the nucleosome core particle

We have developed a novel representation of the nucleosome core particle (NCP) in a cylindrical reference frame. This chapter explains the mathematical details of the methodology, and applies it to characterize the NCP architecture. We herewith also examine the NCP in terms of (i) the atom coordinates, (ii) the DNA-histone and histone-histone contacts, and (iii) the distribution of charges. In the context of this chapter, the crystal structure of NCP 147 {Protein Data Bank (PDB) ID: 1KX5} [1], the currently available best-resolution NCP diffraction model, is used for study.

#### 6.1 Introduction

The nucleosome core particle is a large-size cylindrical assembly of protein and DNA molecules with a radius of  $\sim 10$  nm and height of  $\sim 6.5$  nm and consisting of more than 13,000 heavy atoms (not including hydrogen). The system contains a total 10 different chains including 8 proteins and 2 DNA strands. Given this complexity, it is, nevertheless, feasible to examine the atomic-level interactions and organization inside the nucleosome core particle, if the set of molecules is treated appropriately.

Currently available molecular visualization tools, such as Pymol [2], Rasmol [3], and Chimera [4], cannot quantify the locations of the constituent atoms with respect to the NCP as a whole, although these programs are very useful for displaying molecular structures in many perspectives. Besides, these software tools cannot describe the relative positions between atoms or subunits within the scope of the overall shape of the nucleosome core particle. In this regard, we have developed a realistic shape-based method to represent the nucleosome and to quantify the atomic organization and interactions inside the NCP. This method is based on a cylindrical reference frame resting on

the NCP and arising from the observation that the nucleosome core particle resembles a cylinder. In such a reference frame, the coordinates of atoms can be easily linked to the overall structure, and therefore the atomic-level protein-protein and protein-DNA interactions can be easily visualized and quantified.

## 6.2 Methods

### 6.2.1 NCP reference frame

The reference frame on the nucleosome core particle is key to chromatin simulations, providing a convenient way to track the location and orientation of each protein-DNA assembly as well as to describe the spatial arrangement of the constituent subunits. We employ a shape-based model — a cylindrical reference frame, arising from the observation that the nucleosome core particle resembles a cylinder — based on the pathway of DNA in the currently best-resolved nucleosome core-particle structure (PDB ID: 1KX5) [1]. The 117 base pairs of the nucleosomal DNA centered about the dyad position form an approximate circle located on the surface of the NCP cylinder. The geometric centers of these base pairs are used to determine the reference frame as detailed below.

**First of all**, we find the orientation of the cylindrical axis by applying principal component analysis to the spatial distributions of these base-pair centers,  $\mathbf{w}_i = (bx_i, by_i, bz_i)$ ,  $i = 1, 2, \dots, n$  ( $n = 117$ ), where  $b$  stands for base-pair. The analysis starts with the calculation of the covariance matrix of the  $n \times 3$  coordinate array  $\mathbf{W}$ , composed of the  $n$  coordinate vectors  $\mathbf{w}_i$ ,

$$\mathbf{c} = \frac{1}{n-1} [\mathbf{W}^T \mathbf{W} - \frac{1}{n} \mathbf{W}^T \mathbf{u} \mathbf{u}^T \mathbf{W}], \quad (6.1)$$

where  $T$  stands for array/matrix transpose,  $\mathbf{u}$  is an  $n \times 1$  column with elements of 1, and  $\mathbf{c}$  is the  $3 \times 3$  covariance matrix. We then calculate the eigen-system of the covariance matrix and obtain the three eigenvectors and the corresponding eigenvalues. The eigenvector  $\mathbf{e}$  having the smallest eigenvalue, is used to define the orientation of the cylindrical axis, arising from the observation that the distribution of atoms along the cylindrical axis is narrower than that along the radial direction of the cylindrical NCP.

**Secondly**, we find the origin of the reference frame. In the following process, we keep

the orientation of the cylindrical axis unchanged and move it in space until the DNA radii (the distances from the DNA base-pair centers to the cylindrical axis) are approximately evenly distributed, namely the radial distribution has a minimum variance. We start the process with the calculation of the average location  $\mathbf{w}_0$  of the selected DNA base pairs,

$$\mathbf{w}_0 = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i. \quad (6.2)$$

Next we compute the projection point of the dyad position on the vector  $\mathbf{L}$ , which is a virtual axis passing through the averaged location  $\mathbf{w}_0$  and parallel to the orientation  $\mathbf{e}$  derived above. We then search for the origin of the desired reference frame using the above projection position as a starting heuristic searching point  $\mathbf{s}_0$  and following the procedure: (i) find the distance  $d_i$  of each DNA base-pair center  $\mathbf{w}_i$  to the vector  $\mathbf{L}$ ; (ii) calculate the standard deviation  $\sigma$  of the aforementioned distances; (iii) move the search point  $\mathbf{s}$  with a small displacement in the plane that is perpendicular to the orientation of the cylindrical axis  $\mathbf{e}$  and that includes the averaged location  $\mathbf{w}_0$ ; (iv) redefine the virtual axis  $\mathbf{L}$ , allowing it to go through the new search point  $\mathbf{s}_n$  while keeping the same orientation; (v) perform steps (i) and (ii); (vi) iterate steps (iii) to (v); and (vii) set the search point  $\mathbf{s}_m$ , which gives the minimum value of the standard deviation  $\sigma$ , as the origin  $\mathbf{o}$  of the reference frame. The cylindrical axis is defined as the vector  $\mathbf{L}$ , which passes through the determined point  $\mathbf{s}_m$  along the direction of the eigenvector  $\mathbf{e}$  having the smallest eigenvalue.

**Thirdly**, we set the cylindrical axis as the  $\mathbf{Z}$ -axis. The vector pointing from the origin  $\mathbf{o}$  to the dyad position is defined as the  $\mathbf{X}$ -axis of the NCP reference frame. The  $\mathbf{Y}$ -axis can be determined following the right-hand rule of orthogonality.

### 6.2.2 Coordinate systems

The atomic coordinates of the NCP can be expressed in the above reference frame in two ways, as Cartesian or cylindrical coordinates. The transformation of coordinates begins by constructing a  $3 \times 3$  rotation matrix  $\mathbf{R}$ , using the above determined  $\mathbf{X}$ -,  $\mathbf{Y}$ -, and  $\mathbf{Z}$ - vectors:

$$\mathbf{R} = [\mathbf{X}, \mathbf{Y}, \mathbf{Z}]. \quad (6.3)$$

The Cartesian coordinates of atom  $\mathbf{v}_i = (x_i, y_i, z_i)$  can then be obtained with the following relation,

$$v_i = \mathbf{R}[\mathbf{a}_i - \mathbf{o}], \quad (6.4)$$

where  $\mathbf{a}_i$  is the absolute coordinate vector of the  $i$ th atom in the arbitrary reference frame given in the PDB file, and  $\mathbf{o}$  is the origin of the reference frame that has been determined above.

The cylindrical coordinate system can be easily constructed from the Cartesian coordinate system determined above. It is straightforward to calculate the radius and height components of each atom, which yield

$$R_i = \sqrt{x_i^2 + y_i^2}, Z_i = z_i \quad (6.5)$$

The determination of the phase component ( $\theta_i$ ) must be achieved in several steps. We set the phase angle in the range from  $-360$  to  $360$  degree, arising from the observation that the DNA wraps onto the histone core about 1.65 turns (close to 2 turns).

$$\theta_i = \tilde{\theta}_i + \varphi_i, \quad (6.6)$$

where  $\tilde{\theta}_i$  is defined as

$$\tilde{\theta}_i = \begin{cases} \tan^{-1}(\frac{y_i}{x_i}), & x_i > 0; \\ \tan^{-1}(\frac{y_i}{x_i}) + \pi, & x_i < 0 \text{ \& } y_i > 0; \\ \tan^{-1}(\frac{y_i}{x_i}) - \pi, & x_i < 0 \text{ \& } y_i < 0. \end{cases}$$

The quantity  $\varphi_i$  is a correction term for fitting the phase angle in the above range and has two possible values,  $2\pi$  or  $-2\pi$  radians. For the first half of DNA (base pairs 1 to 74) and the associated histones (H3 $\alpha$ , H4 $\alpha$ , H2A $\alpha$ , and H2B $\alpha$ ), if  $\tilde{\theta}_i$  is a positive value then  $\varphi_i = -2\pi$  else  $\varphi_i = 0$ . For the other part of DNA and other histones, if  $\tilde{\theta}_i$  is a negative value then  $\varphi_i = 2\pi$  else  $\varphi_i = 0$ . All phase angles are finally converted from units of radians to degrees.

## 6.3 Results

### 6.3.1 Molecular organization

We report the three cylindrical coordinate components for each of the NCP atoms in Fig. 6.1-6.4, respectively. Atoms are displayed in the scatter plots for individual chains, in the order of peptide or nucleotide sequence, and in different colors. For each histone subunit, the atom numbers are ordered from the N-terminus to the C-terminus. It is revealed in Fig. 6.1 that most N-terminal atoms of each histone possess a large radius, intruding beyond the radial boundary of the nucleosomal DNA. The C-terminal atoms of H2A and H2B also have large radii. This is consistent with the observation that all the eight core histones have N-terminal tails of variable length, and that besides N-terminal tails, histones H2A and H2B also have C-terminal tails. Among those tails, H3 has the longest N-terminal tail in terms of its radial coordinate, while H2A has the longest C-terminal tail according to this definition. Fig. 6.1 also shows that, compared to the H2A and H2B subunits, the H3 and H4 histones are relatively far away from the cylindrical central axis of the NCP, favoring more interactions of H3 and H4 with the DNA that is located on the outer surface of the molecular assembly.

The scatter plot of the phase angle of each atom ( $\theta$ ) in Fig. 6.2 reveals that the two copies of both the H3-H4 and the H2A-H2B heterodimers are organized symmetrically with respect to the line  $\theta = 0^\circ$ , which represents the  $(X, Z)$  plane of the NCP reference frame and cuts the core particle in two halves. The H3-H4 dimers lie tangent to the above line with the H3 histones intersecting it, while, in contrast, the H2A-H2B dimers lie far away from the line. This mirrors the three-dimensional observation that the  $(\text{H3-H4})_2$  tetramer is formed symmetrically with respect to the NCP dyad axis with a four-helix bundle from the two H3 histones closely interacting at the dyad position. The image also represents the three-dimensional organization of the H2A-H2B subunits, which lie across the symmetry axis in the half cylinder apical to the dyad position. Also, each histone protein shows a “Z” form pattern in the phase angle scatter plot (Fig. 6.2). This pattern maps the histone folding motif — histone fold — where three  $\alpha$ -helices fold in such a manner that the two terminal helices turn across the central

one perpendicularly resembling a “Z” topology, as described in Chapter 1.

The scatter plot of the cylindrical height ( $Z$ ) component of each NCP atom is shown in Fig. 6.3. The two H2A-H2B heterodimers lie respectively on the top and bottom faces of the NCP cylinder with the (H3-H4)<sub>2</sub> tetramer placed in the middle. Nevertheless, these proteins are not associated in a “sandwich”. The (H3-H4)<sub>2</sub> tetramer has little overlap with the H2A-H2B dimer on the ( $X, Y$ ) plane when viewed down along the cylindrical helix.

Projection of the atomic coordinates of the NCP onto the ( $R, \theta$ ) plane, mimics the top-down view along the cylindrical axis of the NCP (Fig. 6.4). The plot of atoms in the ( $Z, \theta$ ) plane corresponds to the side surface of the cylinder with the radial dimension compressed (Fig. 6.5). The histone atoms are uniquely displayed in two panels in both figures in order to avoid overlaps between the (H3-H4)<sub>2</sub> tetramer and the H2A-H2B dimers and to present the full scope of histone atoms. As above in Figs. 6.1–6.3, DNA is colored in cyan, H3 in magenta, H4 in green, H2A in red, and H2B in blue. The three-dimensional organization of molecular chains can, to some extent, be retrieved from the combined information in Fig. 6.4 and Fig. 6.5. It is revealed that histone proteins spread along the track of the nucleosomal DNA where each subunit is tightly associated with a part of the DNA. The H3 and H4 histones mainly interact with the central part of DNA on either side of the dyad position, while H2A and H2B proteins associate with the ends of the DNA helix. The organization of histones in such way provides a helical ramp to wrap DNA onto it and allows DNA to contact intensively with each histone.

### 6.3.2 Molecular contacts

Atomic contacts between molecular chains play a key role in the assembly of the nucleosome core particle. Interactions between DNA and histones, by means of atomic contacts, can directly mediate the deformation and packaging of DNA onto the surface of the histone core and stabilize such wrapping. Meanwhile, associations between individual histone proteins through contacts are related to the arrangement and formation of the histone octamer core, e.g. the (H3-H4)<sub>2</sub> tetramer is organized by close atomic

contacts between two H3 histones. Within this text, a pair of atoms are said to be in contact if they are within a distance no more than 4 Ångstrom.

Histone atoms in contact with the DNA are displayed on the  $(R, \theta)$  and  $(Z, \theta)$  planes in Fig. 6.6. These contacts, many of which are histone tail-DNA interactions, occur primarily around the minor grooves of the DNA. There are about 160 to 240 atoms from each of the histone chains contacting the DNA. Furthermore, these atoms mostly touch DNA on the inside face of the double helix with respect to the radial direction of the cylinder. This statement is justified by the observation that the contacted histone atoms have approximately the same heights as the DNA atoms in the  $(R, \theta)$  plot of Fig. 6.6. We also find atomic contacts between the  $(\text{H3-H4})_2$  tetramer and H2A-H2B heterodimers (Fig. 6.7), which are located in the interior of the NCP and away from the DNA with respect to the cylindrical radius. These histone contacts deviate from the DNA along the cylindrical axis, which can be seen in the  $(Z, R)$  plot of Fig. 6.7.

Thus, a cylindrical representation of the nucleosome, in a few two-dimensional plots, is very convenient for displaying the contacts of atoms in the overall landscape of the NCP. Every histone protein is tightly associated not only with the DNA but also with other histone counterparts. Such short-distance interactions can be of various types, such as hydrogen bonding, salt bridges, etc. One can further look into detailed contacts inside the NCP and take advantage of the cylindrical representation developed here-with, for the purpose of understanding the dynamics of the nucleosome and nucleosome positioning.

### 6.3.3 Distribution of charges

Histone-tail regions are identified based on the three-dimensional folding of histones in the crystal structure of NCP 147 [1]. Every core histone possesses an N-terminal tail region covering the following amino-acid residues: (i) 1-35 of H3; (ii) 1-24 of H4; (iii) 1-15 of H2A; and (iv) 1-29 of H2B. The C-terminal tail, found only in the H2A histone, span residues 121-128. Fig. 6.8 displays the sequences of the four core histones, with the tail regions underscored. Polar amino acids are highlighted in the illustrated histone sequences by color coding. Positively charged amino acids (arginine, lysine,



and histidine) are color-coded respectively in red at the tail regions and in blue in the histone-fold regions. Negatively charged amino acids (aspartic acid and glutamic acid) are color-coded respectively in orange at the tail regions and in green at histone-fold regions. The locations of these charges are then visualized with the cylindrical representations (Fig. 6.9). Many protein charges lie on the NCP surfaces in contact with DNA, including the cylindrical wall and faces. The hydrophilic properties of polar amino acids may partly account for this phenomenon. Other positive charges lie deep inside the protein assembly, and many of these charges are closely associated with negative charges, which suggests the possible occurrence of salt bridges. The association of negative and positive charges from different histone units may contribute to bringing and holding the histone folds together.

The distribution of charges with respect to the chains of histones is summarized in Table 1. Among the total of 980 amino-acid residues in the nucleosome core particle, about 23 percents (224) are located in the tail regions. These tails contain a high density of positive charges (about 37%). In contrast, about 24% of all residues are positive. Negatively charged amino-acid residues account for about 8% of all amino acids and about 4% in the tail regions. The four core histones, have similar numbers of positive amino-acid residues in each tail region: (1) 11 on H3; (2) 10 on H4; (3) 9 on H2A; and (4) 11 on H2B. From Fig. 6.9, we can see that some of the positive charges of the histone tails protrude from the NCP into the surrounding space and the others stay close to the DNA, possibly through electrostatic bonds to the negatively charged phosphate backbones of the DNA helix. The location of the tails further suggests that in a compressed chromatin fiber, protruding tail charges can easily interact with encountered parts, such as histone tails or phosphate backbones, of other nucleosomes. In contrast to the minimal distribution of negative charges on histone tails (only 8), the histone folds making of the inner core of the NCP possess a total of 66 negative amino-acid residues. These negative charges occur in such a way that they closely couple with positive charges and form ion pairs. Ion pairs in proteins are usually identified as salt bridges, defined as the electrostatic interactions between the nitrogen atoms of basic residues (arginine, lysine, and histidine) and the carboxylate oxygen atoms of

acidic residues (aspartatic and glutamatic acid) within a close distance cutoff [5, 6]. Such interactions contribute to the stability in thermophilic proteins. We count the total number of amino acid charges found within a given radius from the center of the nucleosome core particle. Fig. 6.10 presents the charge number for positive charges in blue, negative charges in red, and net charges in black. It is clear that within a radius range of about 23 Ångstrom, the number of positive and negative charges has a very high correlation, leading to a net charge close to zero. Beyond this radial point, the number of negative charges stays at the same level, because there are only few negative charges distributed close to DNA, while positive charges keep growing given that there is a rich distribution of positive charges around nucleosomal DNA. Obviously, the number of net charges grows almost parallel to the number of positive charges at larger radii, due to the fact that positive charges become the dominant contributor to the net charge.

#### **6.3.4 Electrostatic potential**

### **6.4 Concluding remarks**

How the histone proteins and the DNA double helix associate and interact inside the nucleosome core particle is of paramount importance to understanding both the packing of DNA and the dynamics of the nucleosome. Given the very large size of the molecule, it is challenging to visualize the detailed local arrangements of individual atoms in the nucleosome. Our newly developed cylindrical representation of the NCP provides a fresh look at the nucleosome core-particle structural assembly, and a novel way to uncover the overwhelming number of interactions of atoms in a simple, straightforward, and apparent way.

The quantitative representations of the NCP can be further applied to compare X-ray structures of different nucleosome core particles. The cylindrical representation of the nucleosome describes the atoms in such a manner that, each dimension gives straightforward information about the relative locations of atoms with respect to the entire

shape. Thus, under such a specific system of representation, the deviations of corresponding atoms from different structures can be directly expressed in terms of relative movement compared to the whole nucleosome. In this regard, we have developed a database with a graphic web-based interface to explore the internal organization of all currently available nucleosome core particles determined in various X-ray crystallography experiments. Details of the database will be presented elsewhere.

The reference frame defined for the cylindrical representation also plays a key role in the computational modeling of the chromatin. A reference frame must be added to the nucleosome core-particle so that spatial relationships between nucleosomes, including DNA, can be described. The NCP reference frame is defined in such a way that one can easily calculate (i) the relative rotations and translations between two nucleosomes for the evaluation of internucleosome interactions, (ii) the location and orientation of nucleosomal DNA base pairs for tracking the packaging of DNA, and (iii) the global spatial arrangements of nucleosomes. With this advantage, in the course of a chromatin simulation (Chapter 8), a record of the NCP reference frames is sufficient for many basic numerical characterizations of chromatin. More description about the usage of the NCP reference in the modeling of chromatin is given in Chapter 7.

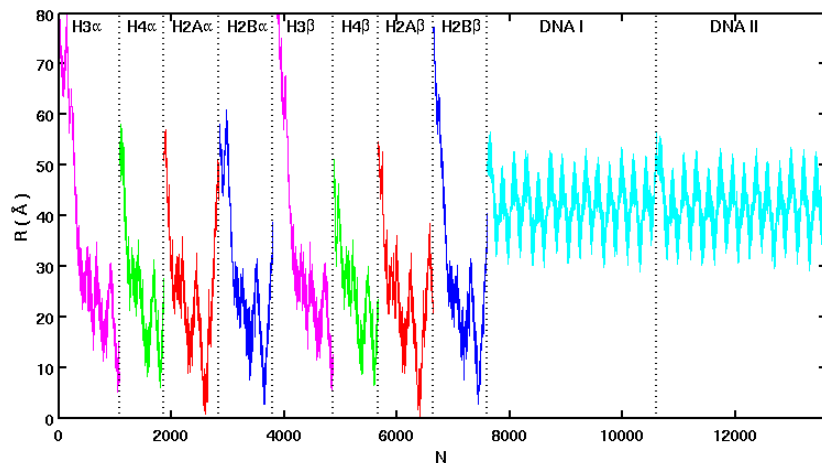


Figure 6.1: Scatter plot of the cylindrical radii of NCP atoms. Atoms are displayed for individual chains of the core histones and DNA. For each histone subunit, atom numbers are ordered from the N-terminus to the C-terminus. The first strand of DNA (DNA I) is plotted from left to right in ascending order of base-pair numbers, while the second strand (DNA II) is illustrated in descending order of the base-pair numbers from left to right. Each histone protein possesses an N-terminal tail which crosses over the radial boundary of DNA. The H2A and H2B histones also have C-terminal tails which present in a similar sense.

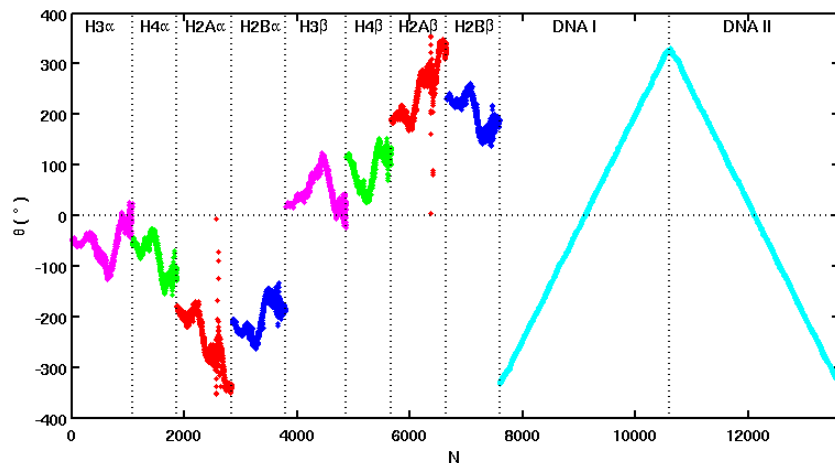


Figure 6.2: Scatter plot of the cylindrical phase angle  $\theta$  of nucleosome atoms. The H3-H4 dimers lie tangent to the  $\theta = 0^\circ$  line, with the H3 histones intersecting with it. In contrast, the H2A-H2B dimers lie far away from the line. Each histone protein shows a “Z” form pattern mapping the histone-fold motif.

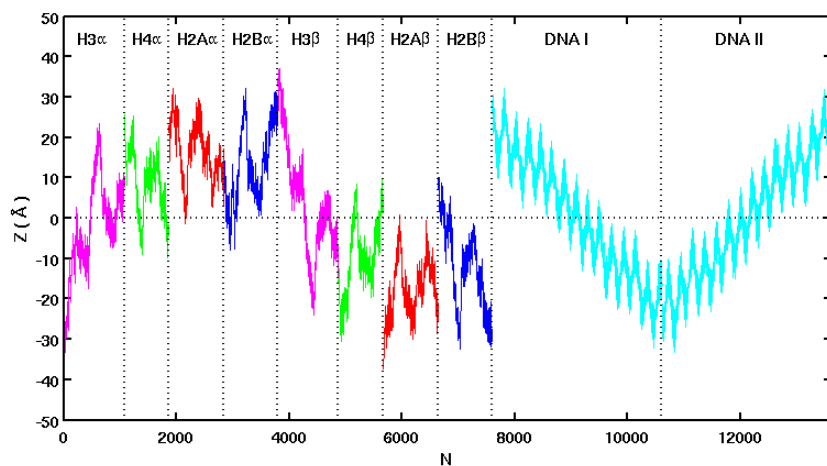


Figure 6.3: Scatter plot of the cylindrical height of nucleosome atoms. The two copies of the H2A-H2B dimers distribute on the top and bottom halves of the nucleosome, with the  $(\text{H3-H4})_2$  tetramer in between.

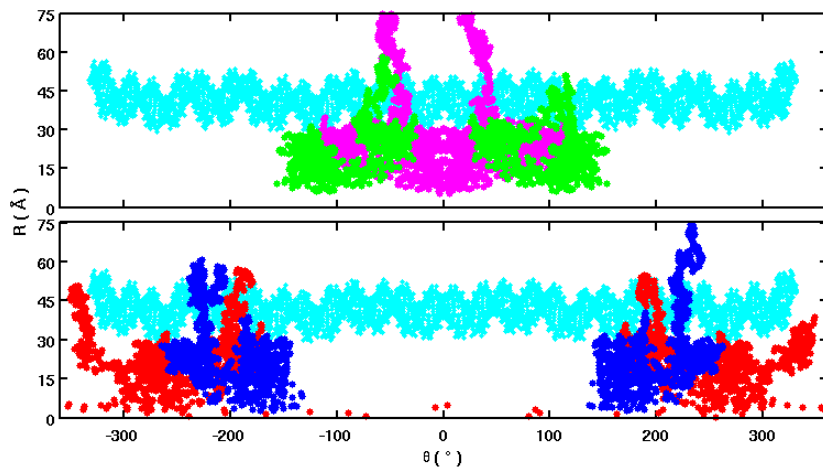


Figure 6.4: Scatter plots of nucleosome atoms on the cylindrical  $(R, \theta)$  plane. Molecular chains are color coded with DNA in cyan, H3 in magenta, H4 in green, H2A in red, and H2B in blue. This combined images, the top depicting the  $(H3-H4)_2$  tetramer and the bottom for the H2A-H2B dimers, mirror the top-down view of the NCP along the cylindrical axis.

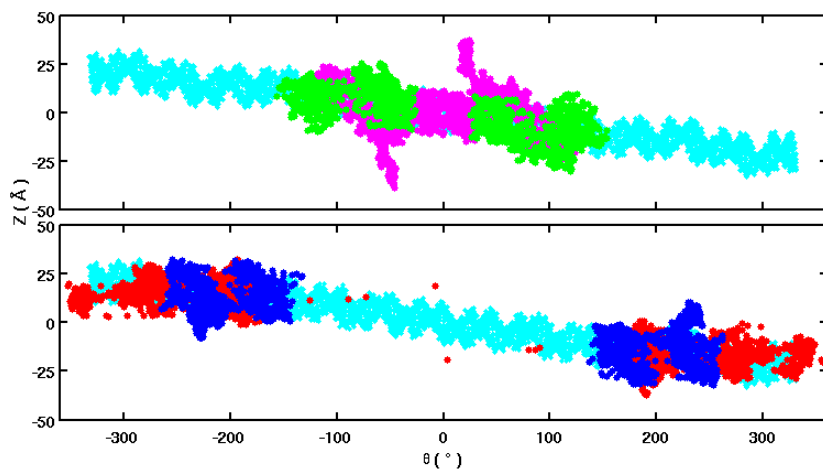


Figure 6.5: Scatter plots of nucleosome atoms on the cylindrical  $(Z, \theta)$  plane. Molecular chains are color-coded and plotted in two graphs as in Fig. 6.4. The combined images mirror the side-surface of the NCP.

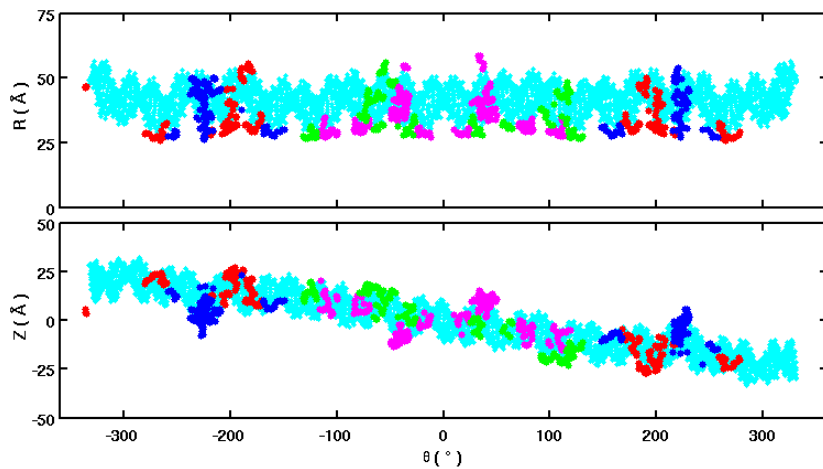


Figure 6.6: Atomic contacts between histone proteins and the nucleosomal DNA. DNA atoms are fully shown and provide a baseline for displaying histone atoms. The distance criterion for contacts is 4 Ångstrom

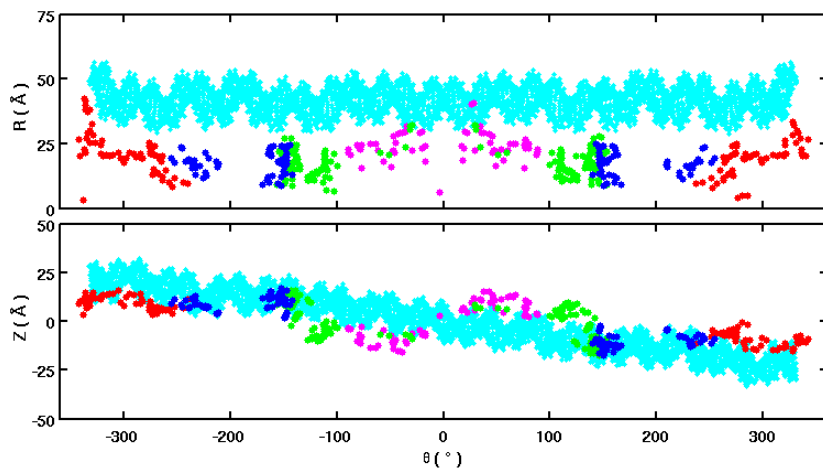


Figure 6.7: Atoms of the (H3-H4) tetramer and the H2A-H2B dimers in contact with other protein atoms in the NCP. The distance criterion for contacts is 4 Ångstrom.

**H2A (128):** SGRGKQGGKT RAKAKTRSSR AGLQFPVGRV HRLLRKGNYA ERVGAGAPVY LAAVLE~~YL~~TA EILELAGNAA RDNKKTRIIP  
RHLQLAVRND EELNKLLGRV TIAQGGVLPN IQSVLLPKT ESSKSKSK

**H2B (125):** PEPAKSAPAP KKGSKKAVTK TQKDGKKRR KTRKESYAIY VYKVLKQVHP DTGISSKAMS IMNSFVNDVF ERIAGEASRL  
AHYNKRSTIT SREIQTAVRL LLPGELAKHA VSEGTKAVTK YTSAK

**H3 (135):** ARTKQTARKS TGGKAPRKQL ATKAARKSAP ATGGVKKPHR YRPGTVALRE IRRYQKSTEL LIRKLFPQRL VREIAQDFKT  
DLRFQSSAVM ALQEASEAYL VALFEDTNLC AIHAKRVTIM PKDIQLARRI RGERA

**H4 (102):** SGRGKGGKGL GKGGAKRHRK VLRDNIGGIT KPAIRRLARR GGVKRISGLI YEETRGVLKV FLENVRDAV TYTEHAKRKT  
 VTAMDVVYAL KRQGRTLYGF GG

Figure 6.8: Sequences of core histones with tail regions underscored. Positively charged amino acids are color coded in red at tail regions and blue at histone-fold regions. Negatively charged amino acids are color coded in orange at tail regions and green at histone-fold regions.

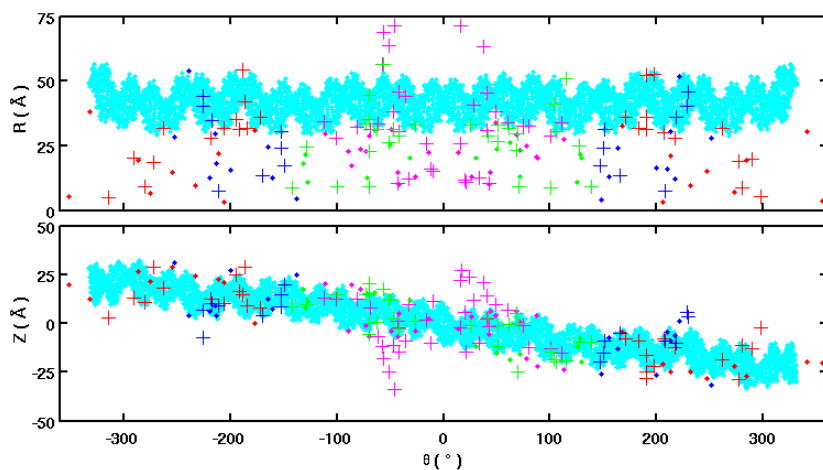


Figure 6.9: Cylindrical view of histone charges. DNA is represented at the all-atom level to provide a base line. Representative atoms thought to carry the charge of polar amino acids include: NH1 for arginine; ND for lysine; NZ1 for histidine; OD1 for aspartic acid; and OE1 for glutamic acid. Positive charges are marked by a '+' sign, while negative charges are plotted as dots



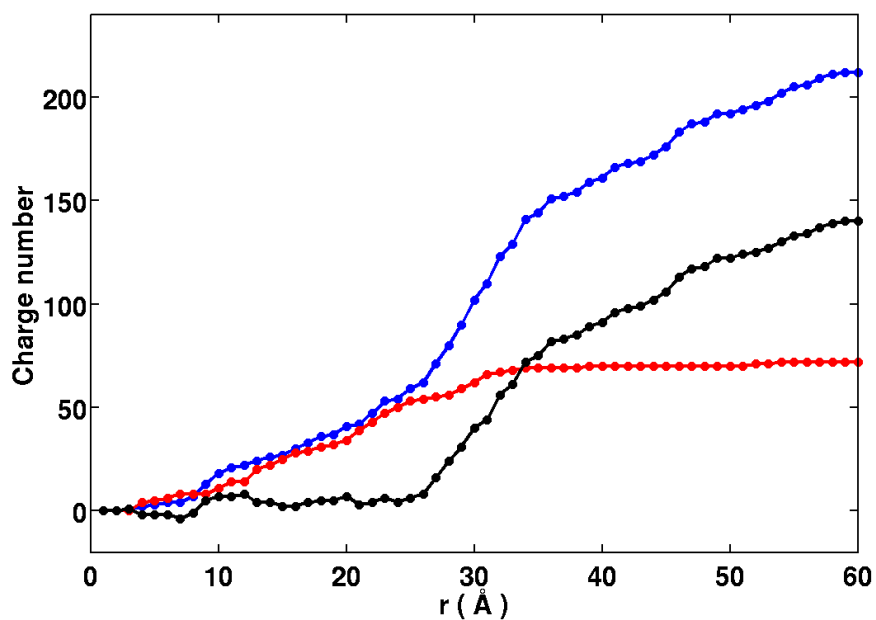


Figure 6.10: Number of charges within a given radial cutoff. Positive charge numbers are marked in blue, negative charge numbers in red, and net charges in black.

Histone	Region	Total res.	Positive res. (ratio)	Negative res. (ratio)
H3	Entire	135	33 (0.24)	11 (0.081)
	Tail	36	11 (0.31)	0 (0)
H4	Entire	102	27 (0.26)	7 (0.069)
	Tail	24	10 (0.42)	1 (0.042)
H2A	Entire	128	28 (0.22)	9 (0.07)
	Tail	23	9 (0.39)	1 (0.043)
H2B	Entire	125	31 (0.25)	10 (0.08)
	Tail	29	11 (0.38)	2 (0.069)
Total (2 copies)	Entire	980	238 (0.243)	74 (0.076)
	Tails	224	82 (0.366)	8 (0.036)

Figure 6.11: Distribution of charges on histones

## References

- [1] Davey, C. A., Sargent, D. F., Luger, K., Mader, A. W., and Richmond, T. J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, **319**, 1097–1113.
- [2] Delano, W. (2002) The PyMOL molecular graphics system, <http://www.pymol.org>.
- [3] Sayle, R. (1995) RasMol, <http://www.umass.edu/microbio/rasmol/index2.htm>.
- [4] Pattersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., and Ferrin, T. E. (2004) CSF Chimera - a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**(13), 1605–1612.
- [5] Barlow, D. J. and Thornton, J. M. (1983) Ion-pairs in proteins. *J. Mol. Biol.*, **168**, 867–885.
- [6] Kumar, S. and Nussinov, R. (1999) Salt bridge stability in monomeric proteins. *J. Mol. Biol.*, **293**, 1241–1255.

## Chapter 7

### Coarse-grained modeling of the chromatin

We have developed a novel coarse-grained method for the simulation of chromatin fibers that incorporates base-pair level modeling. The treatment of the nucleosome core particle includes details of (i) the nucleosomal DNA, (ii) the histone tails, and (iii) the histone folds. The histone folds, forming the core of the NCP, and the nucleosomal DNA on the surface of the NCP, are approximated as rigid bodies, while the histone tails are allowed to flex with respect to the core of the NCP. DNA linkers, connecting successive NCPs, are treated by an elastic model at the base-pair level. Furthermore, the distributions of charges on the protein and DNA atoms in the NCP are simplified by a clustering analysis, and electrostatic interactions among the representative points are evaluated by the Debye-Huckel equation. The chromatin system is sampled by a ‘multi-step’ Monte-Carlo simulation incorporated within the above molecular treatment.

#### 7.1 Introduction

An appropriate simplified model of the nucleosome core particle is necessary in the computer simulation of chromatin fibers, which are too large in size for all-atom calculations. For instance, molecular dynamic simulation with the Amber 10 force field [1] and implicit solvent takes about one week on a powerful computer cluster to obtain a 4-ns trajectory of two-NCP mini-chromatin fiber with a 30-bp DNA linker [Unpublished findings of Dr. Thomas Gaillard]. In practice, all-atom-model simulation is usually an  $N^2$  computing problem for a molecule of  $N$  atoms. That is, a doubling of the system size will require about four fold more computer time to perform a trajectory simulation similar to that of the unit system.

The molecular shape and charge distribution are among the most important factors for

modeling the nucleosome core particle at a coarse-grained level. On the surface of the NCP is 1.65 turns of DNA, which is tightly associated with the proteins inside. The DNA alone sketches the shape and volume of a cylinder, of diameter  $\sim 10$  nm and height  $\sim 6.5$  nm, that approximates the shape of the NCP. Experiments [2, 3, 4] have shown that the nucleosomal DNA in a mononucleosome is subject to an equilibrium of fast unwrapping and wrapping at relatively high salt concentration [5, 6]. It is also observed that histone folding and histone-histone associations in the central NCP octamer are very conservative, even across different species [7, 8]. Thus, it is reasonable to approximate the octamer core of the NCP as a rigid body but to allow the nucleosome DNA to wrap and unwrap. Although we keep the nucleosomal DNA fixed in our preliminary results reported here, our future model will take into account the fluctuation of nucleosomal DNA.

Electrostatic interactions between highly-negative-charged DNA and the polar amino acids of the histone proteins play an important role in nucleosome-nucleosome interactions. There are two major sources of positive charges in the histone proteins. One is those distributed on the surface of the histone octamer. These charges are related to the recruitment and wrapping of DNA onto the histone octamer. The wrapping is enhanced by the insertion of arginines or lysines into the DNA minor grooves about every 10 bp. The second source of positive charges come from the amino acids on the histone tails. The tail residues contribute not only to the stabilization of DNA binding in the nucleosome core particle but also to the geometric compression of nucleosome arrays. In the following context, we show our mathematical designs for modeling the DNA linkers, the charge distributions of the NCP, and the dynamics of histone tails.

The chromatin fiber is a complicated biological system subject to multi-scale molecular interactions and dynamics. The histone tails can flex relatively fast with respect to the other part of the nucleosome core particle and adopt random-coiled configurations. The DNA linkers fluctuate through bending, twisting, shearing, and stretching, and thus can directly determine the spatial relationship between nucleosome core-particles. In turn, the flexibility of the histone tails and DNA linkers are influenced by long-range electrostatic interactions with other molecular components, such as the histone folds

and nucleosomal DNA on the same and different NCPs. Therefore, we employ a ‘multi-step’ Monte-Carlo simulation to sample the chromatin fiber. The simulation is also enhanced by a multi-threading replica-exchange method.

## 7.2 Methods

### 7.2.1 Nucleosome core-particle model

The reference frame on the nucleosome core particle is key to the chromatin simulation, providing a convenient way to track the location and orientation of each protein-DNA assembly as well as to describe the spatial arrangement of the constituent subunits. We employ a shape-based model — a cylindrical reference frame, arising from the observation that the nucleosome core particle resembles a cylinder — based on the currently best-resolved nucleosome core-particle structure (Protein Data Bank ID: 1KX5) [9]. The cylindrical or  $Z$ -axis of the complex coincides with the superhelical axis of the DNA, i.e., the line from which the 147 base-pair centers are minimally displaced [10]. The  $X$ -axis lies along the two-fold symmetry axis of the crystal structure, passing through the center of the central base pair (# 74). The  $Y$ -axis is defined by the right-handed rule, and the origin is placed at the intersection of the two structure-based ( $X$ ,  $Z$ ) axes. A detailed description of this model treatment is given in Chapter 6.

### 7.2.2 Charge distribution model

Nearly a quarter of the amino-acid residues on the histone proteins (238/980) carry a positive charge. Although the proportion of negatively charged amino-acid residues is small ( $\sim 8\%$ ), they compensate many of the positively charged residues in the protein interior. Thus, our coarse-grained model ignores all amino-acid charges within  $23 \text{ \AA}$  of the cylindrical radius, most of which associate as ion pairs (where cationic nitrogen and anionic oxygen atoms are separated by distances of  $4 \text{ \AA}$  or less). The 503 charges retained after truncation (90 on histone tails, 119 on histone folds, 294 on DNA) are separated into 83 clusters based on their spatial locations in the core-particle structure and/or chemical identities. Except for H2A, where the charges on the N- and C-terminal

tails are directly arranged into four clusters, the charges on the histone tails are grouped into two subsets using K-means clustering [11]. The same technique is used to divide the ‘surface’ charges on the histone folds into 18 groups, each containing approximately six charges. The negative charges on DNA are placed into clusters of the same size, i.e., 49 clusters, each made up of six negative charges from three consecutive base-paired residues. Each cluster of protein or DNA charges (Fig. 7.6) is then reduced for structural simulations to a point charge located at its geometric center with magnitude equal to the net charge of the cluster.

### 7.2.3 Linker DNA model

The linker DNA is modeled at the level of base-pair steps in terms of six rigid-body parameters: three angular variables termed tilt, roll, and twist and three variables called shift, slide, and rise with dimensions of distance [12]. A configuration of DNA is defined by the set of parameters at each base-pair step and is said to be relaxed when all parameters adopt their preferred equilibrium values. The potential governing the fluctuations in base-pair steps is assumed to follow a quadratic expression of the form:

$$\Psi = \frac{1}{2} \sum_{i=1}^6 \sum_{j=1}^6 f_{ij} \Delta\theta_i \Delta\theta_j, \quad (7.1)$$

where the  $\Delta\theta_i$  are deviations of the base-pair-step parameters  $\theta_i$  from their intrinsic value  $\theta_i^0$ , and the  $f_{ij}$  are ‘stiffness’ constants. Local sequence-dependent structure and deformability in DNA can be incorporated in the  $\theta_i^0$  and  $f_{ij}$  [13].

If the  $\Delta\theta_i$  at base-pair step  $n$  are collected in the  $6 \times 1$  vector  $\Delta\Theta_n$  and the  $f_{ij}$  in the  $6 \times 6$  force-constant matrix  $\mathbf{F}_n$ , eqn (7.1) takes the form:

$$\Psi_n = \frac{1}{2} \Delta\Theta_n^T \mathbf{F}_n \Delta\Theta_n, \quad (7.2)$$

with the total deformation energy  $U$  of DNA equal to the sum of the  $\Psi_n$  over all  $N$  base-pair steps:

$$U = \sum_{n=1}^N \Psi_n. \quad (7.3)$$

Charges on the linker DNA are treated in the same way as those on the nucleosome core particle, i.e., groups of six negative charges from three base-paired nucleotides,

modeled as point charges located at the centers of the central base pairs.

#### 7.2.4 Gaussian sampling

We take advantage of the quadratic form of the energy in eqn (7.1) and the assumption that the base-pair steps fluctuate independently of one another to collect a Boltzmann distribution of dimeric states. We achieve this, as described elsewhere [14], by diagonalizing  $\mathbf{F}$  and sampling linear combinations of base-pair-step parameters along the principal axes of dimeric deformation. Here we treat the DNA as an inextensible, naturally straight molecule with an intrinsic helical repeat of 10.5 bp/turn. The tilt and roll angles are according null and the twist is  $\sim 34.3^\circ$  in the rest state ( $\theta_1^0 = \theta_2^0 = 0$ ;  $\theta_3^0 = 34.3^\circ$ ). The translational parameters are ‘fixed’ at their intrinsic values ( $\theta_4^0 = \theta_5^0 = 0$ ;  $\theta_6^0 = 3.4 \text{ \AA}$ ) by the assignment of large force constants. The root-mean-square fluctuations in tilt are equated to those in roll, i.e.,  $\langle \Delta\theta_1^2 \rangle^{1/2} = \langle \Delta\theta_2^2 \rangle^{1/2}$ , so that bending is isotropic, and assigned values of  $4.84^\circ$  corresponding to a persistence length  $a = 2\Delta s / (\langle \Delta\theta_1^2 \rangle + \langle \Delta\theta_2^2 \rangle)$  of nearly  $500 \text{ \AA}$  (if  $\Delta s$ , the per residue base-pair displacement, is taken as  $3.4 \text{ \AA}$ ). The fluctuations in twist are assumed to be independent of the bending deformations so that the model corresponds to the classic twisted wormlike chain representation of DNA [15]. The assumed fluctuations in twist  $\langle \Delta\theta_3^2 \rangle^{1/2} = 4.09^\circ$  correspond to a global twisting constant  $C = k_B T / \langle \Delta\theta_3^2 \rangle$  somewhat larger in magnitude than the global bending constant  $A$ , i.e.,  $C/A = 1.4$ , where  $A = ak_B T$ . This choice of  $C$  is compatible with measurements of the equilibrium topoisomer distributions of DNA minicircles and the fluorescence depolarization anisotropy of ethidium bromide molecules intercalated in DNA minicircles [16, 17].

We also model mixed-sequence chain subject to the fluctuations in base-pair steps seen in high-resolution structures and scaled to yield a persistence length of  $\sim 500 \text{ \AA}$  [13, 18]. The latter model allows for well-known features of DNA deformability such as anisotropic bending [19], the coupling of bending and shearing deformations [10], chain extensibility [20], etc. More details about this model are presented in Chapter 4.

### 7.2.5 Electrostatic interactions

Inter-nucleosomal electrostatic interactions play a key role in the folding of chromatin fibers. The present computations take account of all pairwise interactions between point charges on the nucleosome core particles and DNA linker segments along a given nucleosome-decorated DNA sequence. The electrostatic interactions are divided into three groups (nucleosome-nucleosome, nucleosome-linker, and linker-linker) and evaluated with a Debye-Huckel potential:

$$\begin{aligned}
 E_e = & \sum_{i=1}^{N_n-1} \sum_{j=i+1}^{N_n} \left[ \sum_{m=1}^{M_n} \sum_{n=1}^{M_n} V_{im,jn} \right] \\
 & + \sum_{i=1}^{N_n} \sum_{j=i+1}^{N_l} \left[ \sum_{m=1}^{M_n} \sum_{n=1}^{M_l} V_{im,jn} \right] \\
 & + \sum_{i=1}^{N_l} \sum_{j=i+1}^{N_l} \left[ \sum_{m=1}^{M_l} \sum_{n=1}^{M_l} V_{im,jn} \right]
 \end{aligned} \tag{7.4}$$

$$V_{im,jn} = \frac{q_{im}q_{jn}}{4\pi\epsilon_0\epsilon r_{im,jn}} \exp(-\kappa r_{im,jn}) \tag{7.5}$$

where the three terms on the right side of eqn (7.4) correspond respectively to the three types of electrostatic interactions listed above. The number of nucleosomes in the chromatin system is denoted as  $N_n$ , and the number of DNA linkers is denoted as  $N_l$ . The notations  $M_n$  and  $M_l$  refer respectively to the number of charge clusters on a nucleosome core particle and a DNA linker. The distance between charge cluster  $m$  in object  $i$  and charge cluster  $n$  in object  $j$  is marked as  $r_{im,jn}$ . The object can be the center of cluster charges on a nucleosome core particle or a DNA linker.

### 7.2.6 Rigid-body representations

Both DNA base pairs and nucleosome core particles (except for the histone tails) are treated as rigid bodies, with the relative positions of the constituent atoms held fixed. We consider two levels of molecular space: (i) a local frame resting on a rigid body and (ii) a global frame in which the simulated chromatin system is described. A generator matrix containing a  $3 \times 1$  displacement vector  $\mathbf{r}$  and a  $3 \times 3$  rotation matrix  $\mathbf{T}$  is used



to express the locations and orientations of all objects in these spaces:

$$\mathbf{A} = \begin{bmatrix} \mathbf{T} & \mathbf{r} \\ \mathbf{0} & 1 \end{bmatrix}. \quad (7.6)$$

The local frame of a nucleosome core particle is the cylindrical reference frame described above. The coordinates of all atoms in the core particle, including the relative positions of the charged-atom clusters, are transformed from the original crystallographic reference frame stored in the Protein Data Bank to the local cylindrical space of the nucleosome from knowledge of the spatial relationship between the two frames. The precise arrangement of individual base pairs within the core particle, i.e., the  $\mathbf{r}$  and  $\mathbf{T}$  that describe the spatial disposition of local base-pair frames with respect to the cylindrical frame, are extracted from the transformed coordinates using the 3DNA suite of programs [21, 22]. Using this strategy, every object associated with a nucleosome core particle is expressed in the local cylindrical space and saved for later usage in the simulation.

The global structure of chromatin is evaluated using a series of generator matrices, where the matrix product that describes the position of the  $n$ th nucleosome is given by:

$$\mathbf{A}_n = \mathbf{A}_{n-1} \mathbf{A}_{D \rightarrow N} \mathbf{A}_{D1} \cdots \mathbf{A}_{Di} \cdots \mathbf{A}_{DL} \mathbf{A}_{N \rightarrow D} \quad (7.7)$$

Here  $\mathbf{A}_{n-1}$  is the generator matrix that describes the position of the preceding core particle,  $\mathbf{A}_{D \rightarrow N}$  is the generator matrix that expresses the coordinate frame on the exiting base-pair of nucleosome  $n - 1$  in the frame of the core particle,  $\mathbf{A}_{D \rightarrow N}$  is the generator matrix that expresses the frame of core particle  $n$  in that of the entering base pair, and  $\mathbf{A}_{Di}$  is the generator matrix that relates the frame of base pair  $i + 1$  to that of base-pair  $i$  on the linker DNA between nucleosome core particles  $n - 1$  and  $n$ .

The location of a charge cluster in the  $n$ th core particle is obtained from the generator matrix  $\mathbf{A}_g$  expressed as:

$$\mathbf{A}_g = \mathbf{A}_n \mathbf{A}_b, \quad (7.8)$$

where  $\mathbf{A}_b$  is the generator matrix that describes the charge cluster in the cylindrical frame of the corresponding core particle.

### 7.2.7 Excluded volume

The excluded volumes between nucleosome core particles and DNA linkers are identified and eliminated in the course of the chromatin simulations. Given that the pathway of DNA alone roughly defines both the volume of the core particle and the space occupied by linker DNA, the centers of the 3-bp DNA segments on the core particle and the linker DNA are used as test points. The overlap between core particles is estimated from positions of the centers of the DNA segments on one nucleosome with respect to coordinate frame of the other core particle. Steric clashes between a DNA linker and a core particle are similarly followed from the locations of the DNA linker segments with respect to the coordinate frame on the target core particle. A clash is identified if the radial location  $r_T$  and the  $z$ -component  $z_T$  of the test point, expressed in the cylindrical frame of the core particle, lie within the volume of an idealized nucleosomal cylinder of radius  $R_N$  and height  $H_N$ :

$$r_T < R_N, |z_T| < \frac{H_N}{2}. \quad (7.9)$$

Overlaps between DNA linkers are identified from the pairwise distances between all DNA-segment centers in the two linkers. An overlap occurs if any such distance is smaller than the diameter of the DNA, which is about 2 nm.

## 7.3 Model details

### 7.3.1 Ion pairs

Charges on the NCP are plotted on the cylindrical planes (Fig. 7.1) with different color coding and grouping criteria, from those in Fig. 6.9 in Chapter 6. The nucleosomal DNA is represented by its backbone phosphorus atoms in cyan, which are approximated as carriers of negative charges. Polar amino acids are represented respectively by their charged groups: NH1 for arginine, ND for lysine, and NZ1 for histidine, OD1 for aspartic acid, and OE1 for glutamic acid. Each charge in the map, marked as a dot, carries either one positive or one negative electronic charge. Positively charged amino acids are color-coded respectively in red in the tail regions and blue in the histone-fold regions. Negatively charged amino acids are color-coded respectively in orange in the

tail regions and green in the histone-fold regions.

Within the boundary of the nucleosomal DNA ( $R < 50 \text{ \AA}$ ), the negative amino acid charges populate in such a way that their positions are closely coupled with those of positively charged atoms and form ion pairs. Ion pairs in proteins are usually identified as salt bridges, defined as the electrostatic interactions between the nitrogen atoms of basic residues (arginine, lysine, and histidine) and the carboxylate oxygen atoms of acidic residues (aspartic acid and glutamic acid) within a distance cutoff. Such interactions contribute to the stability in thermophilic proteins [23, 24]. The cutoff distance used to define salt bridges varies in the literature, ranging from  $3.5 \text{ \AA}$  to  $4.5 \text{ \AA}$ . The coupling of ion pairs is also meaningful in long-distance electrostatic effects. The electrostatic potential of a charge can be partially or totally offset by its paired counterpart depending on the position at which the potential is measured. Figure 7.2 shows that the minimal offset of the electrostatic potential of one charge from the other charge of opposite sign in a pair can be more than 70% if the distance from the point of measurement to the center of the pair is larger than  $20 \text{ \AA}$  along the line connecting the two ions. If the experimental points lie on the symmetric axis of the ion pair, which is the line that passes through the middle of the two ions and perpendicular to the line linking them, the potential from the two charges cancel each other. This suggests that beyond a certain distance range the electrostatic potential of the ion pair can be ignored. This is very helpful when we consider the effect of amino acid charges inside the histone core on the surrounding of the nucleosome core particle, which has a radius about  $50 \text{ \AA}$ . On the other hand, if a charge is altered in a pair of ions, the surrounding electrostatic potential pattern could be changed. Such changes in potential could be useful for the interpretation of how histone mutation affects nucleosomal DNA wrapping.

For the purpose of the study of long-distance electrostatic potential effects, we use a distance cutoff of  $7 \text{ \AA}$  for the identification of ion pairs. This distance criterion is more generous than classical values of  $3.5 - 4.5 \text{ \AA}$  in general, but is acceptable for our coarse-grained modeling. Figure 7.3 is a ‘salt bridge’ map, which lists all histone positive charges on the horizontal axis and all histone negative charges on the vertical axis. The

charges, which are organized by their radial values, follow a major descending trend from the upper left to lower right, although minor violations may occur. Most negative charges of small radii are paired with a positive charge, and most positive charges of small radii are paired with a negative charge. This leads to the conclusion that a great degree of ion pairing occurs inside the nucleosome core particle.

It has been shown in Fig. 6.10 that within a radial range of about  $23 \text{ \AA}$ , the number of positive and negative charges are roughly equal, leading to a net charge close to zero. Beyond this radial point, the number of negative charges stays at the same level, because there are only few negative charges distributed close to DNA, while the number of positive charges keeps growing, given that there is a rich distribution of positive charges around the nucleosomal DNA. Obviously, the net charges grows almost in parallel with the number of positive charges at large radii, due to the fact that the positive charges are the dominant contributor of charge.

### 7.3.2 Reduction of charges

The coarse-grained model, ignores amino acid charges within a radial range of  $23 \text{ \AA}$ , due to the fact that this region has a net charge closed to zero and that charges within this region mostly occur in the form of ion pairs. In contrast to Figure 7.1, Figure 7.4 displays only the amino acid charges within a radius of greater than  $23 \text{ \AA}$ . The charges include the charged amino acids on the histone tails and DNA backbone, that is, 92 charges on the tails (86 positive, 6 negative), and 292 negative charges on the DNA backbone. Of the 218 charges on the histone folds, 119 charges remain after truncation, among which 100 are positive and 19 are negative. These charges lie on the surface of the histone octamer and closely interact with the nucleosomal DNA. The reduced charges are involved in two potentially important interactions: (i) the histone fold amino acids may interact with the nucleosomal DNA, and (ii) the histone tails may interact with the nucleosomal DNA as well as the nucleosomal surroundings.

### 7.3.3 Clustering analysis

After charge reduction, there are still 503 charges on a single nucleosome core particle. If we consider pairwise atomic interactions between two nucleosomes, the electrostatic energy calculation must be computed 126253 times. This complexity precludes the treatment of many nucleosome core particles. One way to reduce this complexity is to cluster with the charges into groups and represent the whole group of charges by a single point charge. This is the style of a mean-field approximation, in which an  $n$ -body problem is simplified as a one-body field.

Clustering analysis is performed in three domains: (i) the histone tails, (ii) the nucleosomal DNA, and (iii) the histone fold. Charges on each histone tail are grouped into two subsets using K-means method, except for those on H2A. The two copies of the H2A histone have four distinct tails, including two C-terminii. Each H2A tail accounts for one group of charges, since it is short compared to other histone tails. In total, there are 16 groups of charges on the histone tails. The K-means clustering method is also applied to cluster the surface charges of the histone folds into 18 groups. The number of clusters is chosen such that each cluster contains about 6 charges on average. For the nucleosomal DNA, every three base pairs are segmented as a group, which contains 6 negative charges centered at the center of the middle base pair. The nucleosomal DNA thus contributes a total of 49 groups.

The Cartesian coordinates of the centroid positions, the number of charged carriers, and the net charges of the clusters are listed respectively in Table 7.1 and Table 7.2 for the histone tails and folds. Figure 7.5 illustrates these clusters in the cylindrical planes. Clusters on the histone tails are color-coded in red, and those on the histone folds in blue. The DNA centers are marked in magenta. The DNA phosphorus atoms in cyan are also shown for the convenience of displaying the relative locations of the clusters. The number of clusters is distributed evenly over positive and negative values of the phase  $\theta$ . The surface charges of the histone folds (in blue) mostly cluster near the minor grooves of the nucleosomal DNA. This is consistent with the fact that DNA on the nucleosome core particle is mediated by the insertion of arginines and lysines into its minor grooves. In the three-dimensional model, each of the clusters is represented as

a sphere located at the centroid of the cluster. The sphere carries the same net charge as the corresponding cluster. Cluster information along with the nucleosomal DNA are also plotted in three-dimensional space (Fig. 7.6). The color code is same as in Fig. 7.5. The histone proteins are shown in gray, to distinguish the distribution of histone tails with their charge clusters.

### 7.3.4 Electrostatic potential

One NCP can influence its surroundings, including other NCPs and linker DNAs, via long-range electrostatic interactions between charges on the histone proteins, nucleosomal DNA, and linker DNA. Such a potential plays a key role in the determination of chromatin folding, and is supported by experimental evidence [25, 26, 27] showing that chromatin folding is significantly altered by varying salt concentration. A sound computer model for chromatin folding should be able to account for an appropriate representation of the potential of an NCP. To make our model more accurate, we use the aforementioned all-charge model for the calculation of the electrostatic potential between an NCP with other charges, under the assumption that the NCP is a rigid body. For calculation of the interaction between an NCP and a surrounding target charge, one straightforward way is to sum the point-point interactions between all charges of the NCP with the target charge. However, this would lead to a heavy burden to obtain the electrostatic potential between two NCPs, which requires 126253 such calculations, given that an NCP possesses 503 charges. We propose the following method to significantly reduce this computational cost.

We pre-calculate the potential field of an NCP in its body reference frame, which has been defined previously. The potential field is obtained by the summation of the potential of all charges on the NCP with a unit charge placed in pre-defined bins. These bins are chosen in the dimensions of the cylindrical coordinate system, namely each bin is represented by  $(r, \theta, z)$ . All the three axes are grided evenly, such that the volume of a bin grows with the radius. This makes more sense than using a Cartesian reference frame, because the electrostatic potential damps very fast along the radial direction, and allows one to reduce the fineness of the grid and thereby save much memory. With

such a pre-calculated potential field, a target charge in space is first expressed in the body reference frame of the NCP, then assigned a bin based on its cylindrical coordinates, and finally used to obtain the potential energy by multiplying the field in the bin and its charge. This calculation only needs a one-time calculation of the potential field and storage in a library. The electrostatic calculation between two NCPs, thus requires only 503 calculations with the potential field. It is emphasized that this is a calculation at the level of the all-charge model, which may be argued to be more accurate than the mean-field model. Use of this approach requires that we assume the whole NCP to be a rigid body. If we allow the DNA or histone tails to flex, this method may not be appropriate but the mean-field model can be used.

Fig. 7.7 presents the electrostatic potential of an NCP on its  $(Z, \theta)$  plane for different radii. The potential is mapped with contours with levels ranging from  $-10 kT$  to  $10 kT$ . The color bar on the right side of each panel shows the potential value scale and color code. Potentials are reported for 5 different radii between  $50 \text{ \AA}$  and  $70 \text{ \AA}$  at increments of  $5 \text{ \AA}$ . The panel for radius  $50 \text{ \AA}$  displays the electrostatic potential on the surface of the NCP (with a radius of  $50 \text{ \AA}$ ), effected by all charges from amino acids and DNA, including histone tails, with a radius over  $23 \text{ \AA}$ , as listed in Fig. 7.1. The left panels in Fig. 7.7 are potentials calculated with histone tails, while the rights ones are evaluated without histone tails. Around the NCP surface (small radii), the negative potentials show a distinct pattern along the DNA track, which is due to the high density of negative charges on the P atoms of DNA. Every five base-pairs of the DNA exhibit a potential motif with three deep wells (minima in dark blue), the latter value corresponding to the P atoms most distant from the NCP center. With increase of radius, the negative potentials become weaker. If the histone tails are absent, the electrostatic potential becomes neutral at locations far away from the NCP; if the histone tails are present, islands of positive potential associated with the cationic atoms on the histone tails appear. These islands of positive potential can extend to distances over  $20 \text{ \AA}$  away from the surface of the NCP and contribute to interactions over with other nucleosome core particles at long distances. Fig. 7.8 is another presentation of the electrostatic potential of an NCP in the  $(r, \theta)$  plane at different heights  $Z$ . The

left panels are potentials calculated with histone tails, and the right panels are cases without histone tails. For  $-30 < Z < 30$ , only potentials with  $r \geq 50$  are shown, as we are currently interested in the interactions between an NCP with other molecules other than those inside the NCP. The  $50 \text{ \AA}$  cutoff is selected because the NCP has a radius of  $\sim 50 \text{ \AA}$ . It is shown that the existence of histone tails can affect a location as far as  $50 \text{ \AA}$  from the NCP surface, in the sense of electrostatic interactions, by comparing left and right panels in Figure 7.8.



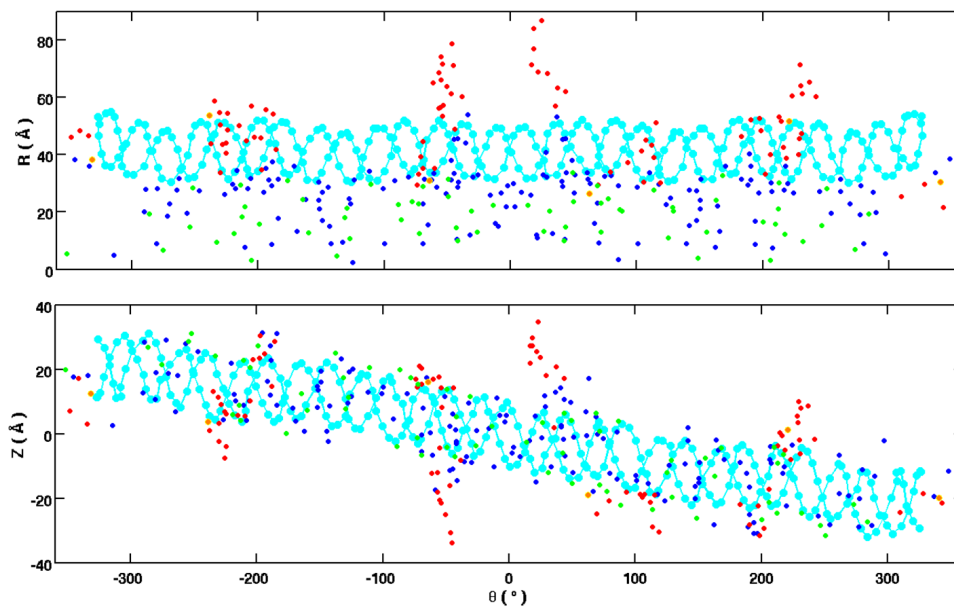


Figure 7.1: Two-dimensional map of the charge distribution on NCP-147 [9]. DNA is represented by its phosphorus atoms in cyan, arginine by NH1, lysine by ND, histidine by NZ1, aspartic acid by OD1, and glutamic acid by OE1. These representative protein atoms are approximated as the charge carriers of the specific amino-acid residues. Positively charged amino-acid atoms (arginine, lysine, histidine) are color-coded in red in the tail regions and blue in the histone-fold regions. Negatively charged amino-acid atoms (aspartic acid and glutamic acid) are color-coded in orange in the tail regions and green in the histone-fold regions.

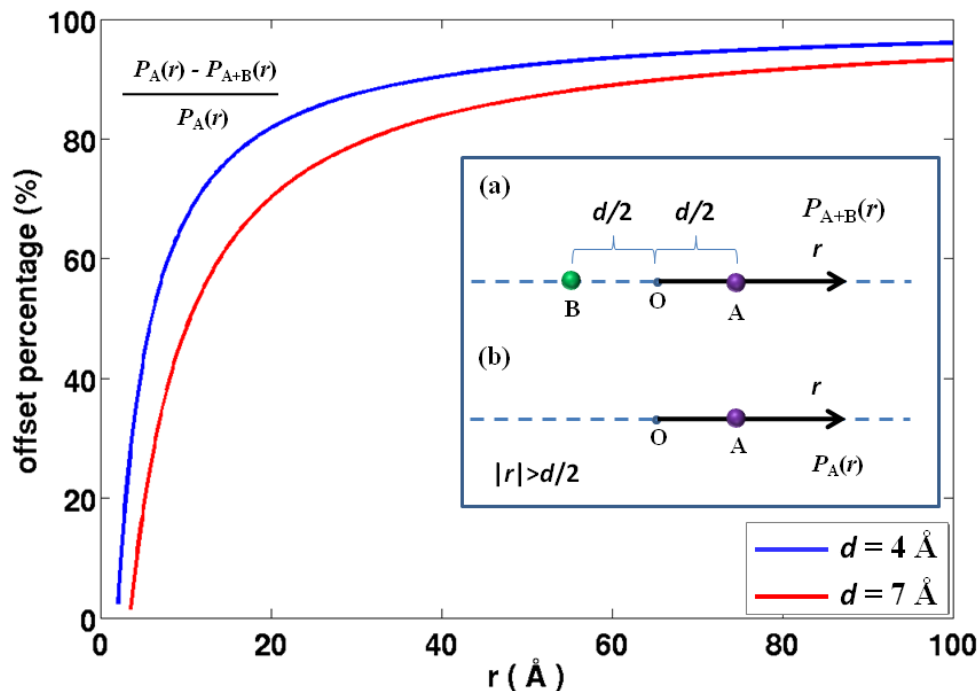


Figure 7.2: Long-distance electrostatic effect of an ion pair. The embedded image presents two potentials: (a) the potential  $P_{A+B}(\mathbf{r})$  of the system of Ion A and Ion B, at the point  $\mathbf{r}$  with respect to the middle point  $\mathbf{O}$  along the direction  $A \rightarrow B$ ; (b) the potential  $P_A$  of Ion A alone at the same location as that in (a). Ions A and B have charges of opposite sign. The two curves plot the percentage of minimal potential offset  $\frac{P_A(\mathbf{r}) - P_{A+B}(\mathbf{r})}{P_A(\mathbf{r})}$  against the distance  $r$  with respect to the middle point of the line connecting the pair. The blue line is for an ion pair separated by  $d = 4 \text{ Å}$ ; the red line for a pair with  $d = 7 \text{ Å}$ .



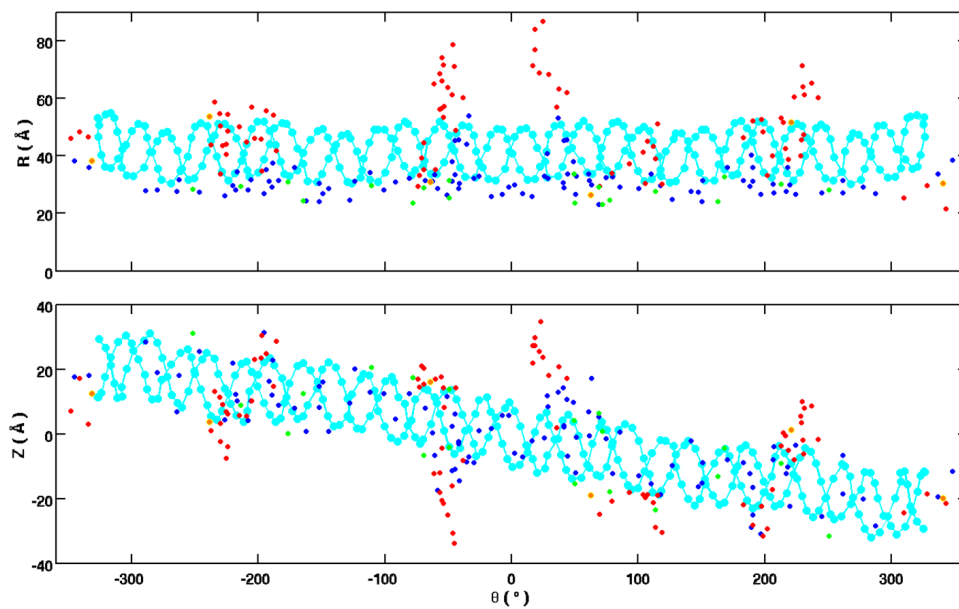


Figure 7.4: Two-dimensional cylindrical map of the charged atoms of NCP-147 [9] with a radius over 23 Ångstrom. This map has the same atomic representation and color coding as in Figure 7.1.

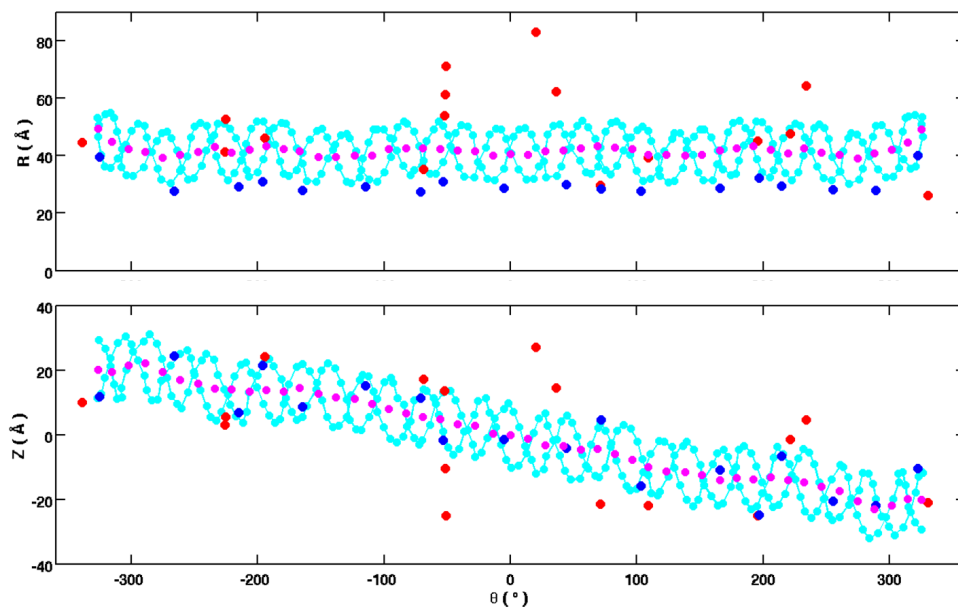


Figure 7.5: Two-dimensional cylindrical map of charge clusters. Centers of clustered groups are displayed: (1) 16 groups of tail charges shown in red, (2) 18 groups of histone fold charges shown in blue, and (3) 49 centers of nucleosomal DNA shown in magenta. The DNA phosphorus atoms are also displayed as a reference for viewing convenience.

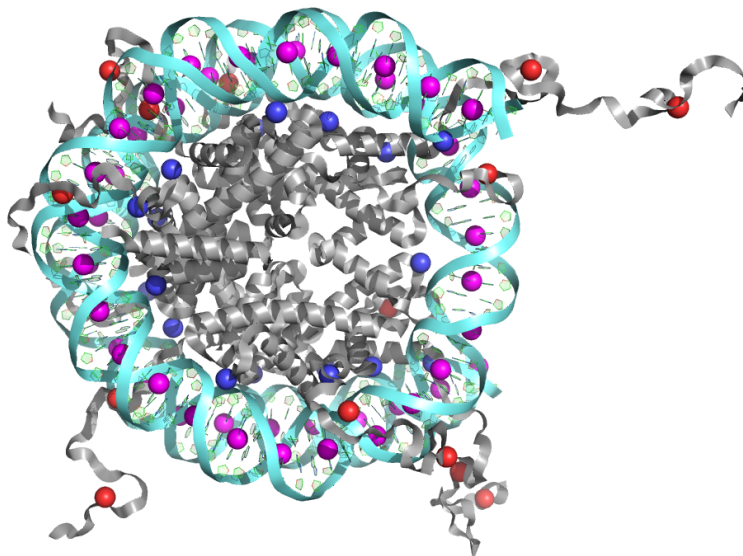


Figure 7.6: Centers of clustered charges within NCP-147 [9] in 3D. The DNA backbones are displayed as ribbons and color-coded in cyan. The histone proteins are shown as ribbons and color-coded in grey. Cluster centers representing charges on the histone tails are color-coded in red, those on DNA charge centers in magenta, and those on histone-fold cluster centers in blue.

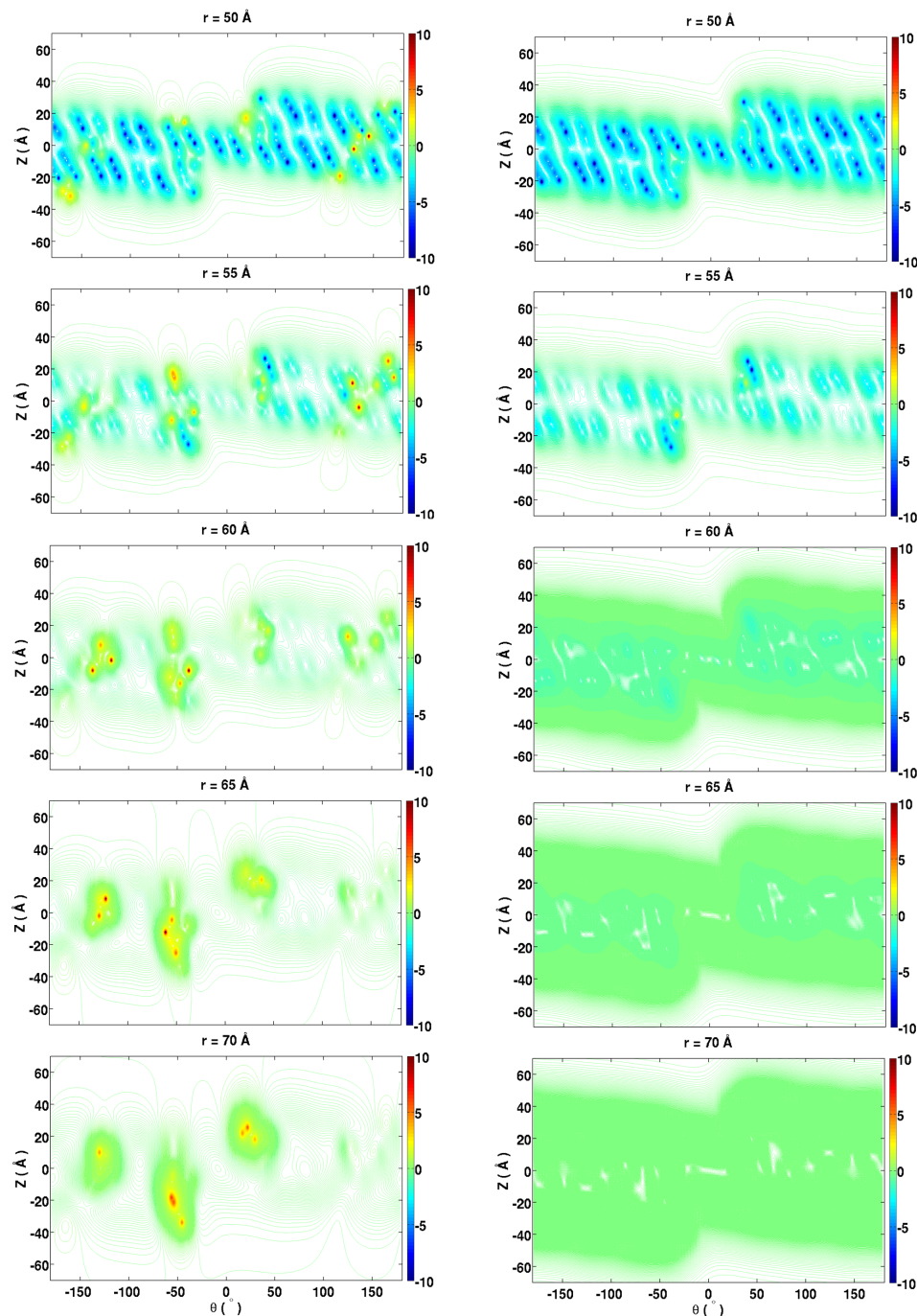


Figure 7.7: Potential of the NCP in the  $(Z, \theta)$  plane, effected by charges from amino acids and DNA with a radius over  $23 \text{ \AA}$ , as listed in Fig. 7.1. Left panels are potentials calculated with histone tails; right panels are cases without histone tails. The potential is color-coded by many contour levels. The space within a contour represents the same level of values as the color-coded contour line. White spaces in the figure are generally bounded by contour lines of a level close to zero. That is, these white areas have a potential close to zero.



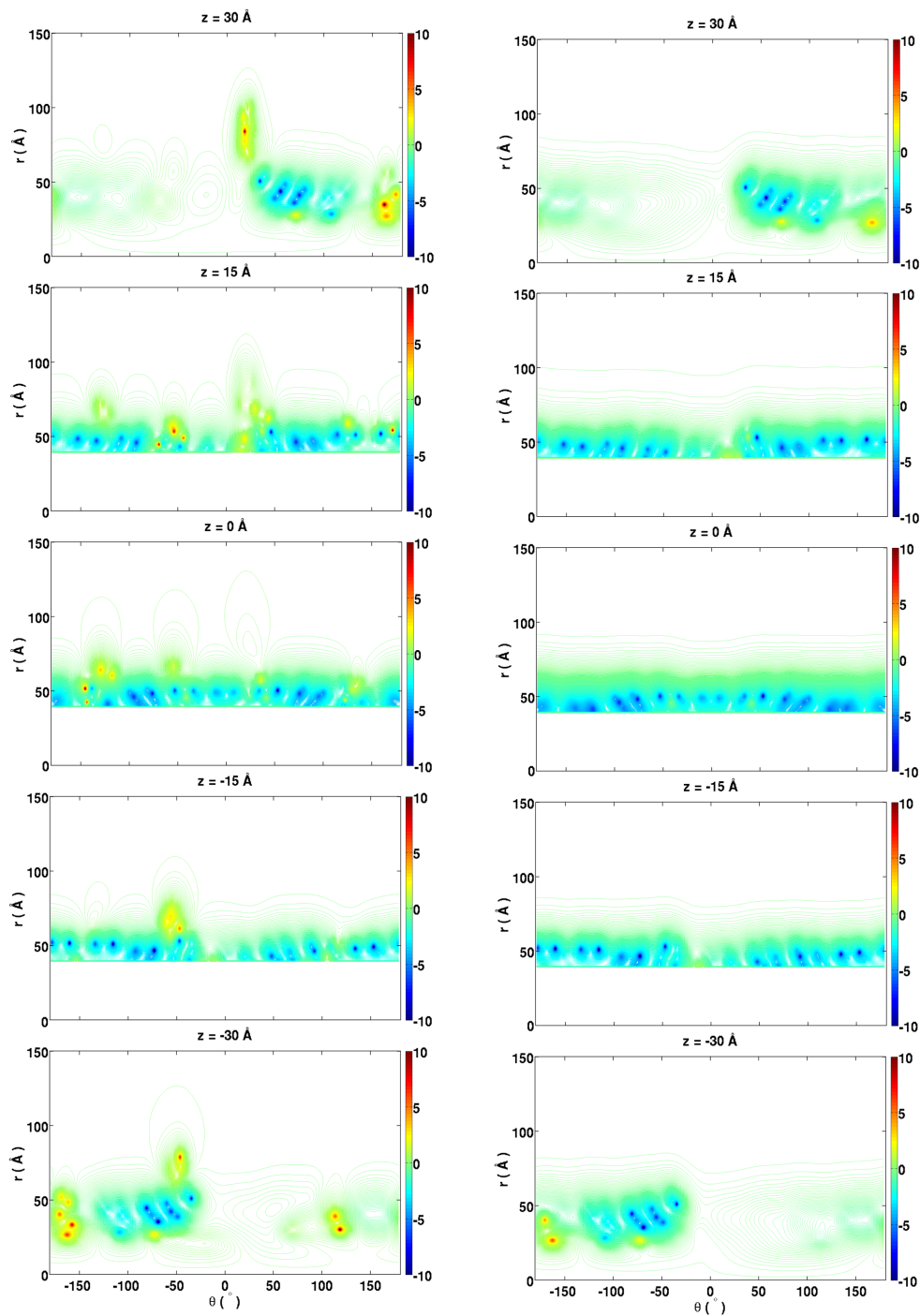


Figure 7.8: Potential of the NCP in the  $(r, \theta)$  plane, effected by charges from amino acids and DNA with a radius over  $23 \text{ \AA}$ , as listed in Fig. 7.1. Left panels are potentials calculated with histone tails; right panels are cases without histone tails. For  $-30 < Z < 30$ , only potentials with  $r \geq 50$  are shown.



Table 7.1: Clusters of histone tail charges

Cluster ID	Charge carriers	Net charges	R	$\theta$	Z	Chain ID
			(Å)	(°)	(Å)	
1	4	2	44.6	-339.3	9.9	H2A $\alpha$ *
2	6	6	41.3	-225.8	3.1	H2B $\alpha$
3	8	6	52.7	-225.5	5.5	H2B $\alpha$
4	6	6	46.2	-194.3	24.2	H2A $\alpha$
5	7	5	35.3	-68.8	17.3	H4 $\alpha$
6	4	4	53.9	-52.2	13.6	H4 $\alpha$
7	5	5	61.2	-51.9	-10.5	H3 $\alpha$
8	6	6	71.2	-51.3	-24.9	H3 $\alpha$
9	7	7	83.0	20.0	27.2	H3 $\beta$
10	4	4	62.4	36.3	14.4	H3 $\beta$
11	3	1	29.7	71.0	-21.5	H4 $\beta$
12	8	8	39.4	108.8	-21.9	H4 $\beta$
13	6	6	45.1	195.6	-25.1	H2A $\beta$
14	9	7	47.7	221.4	-1.5	H2B $\beta$
15	5	5	64.3	233.8	4.7	H2B $\beta$
16	4	2	26.1	330.5	-21.1	H2A $\beta$ *

\* C-terminal tail

Table 7.2: Clusters of histone core charges. Each cluster may contain charges from various histone chains.

Cluster ID	Charge carriers	Net charges	R	$\theta$	Z
			(Å)	(°)	(Å)
1	9	9	39.5	-325.3	11.8
2	5	3	27.7	-266.0	24.4
3	7	5	29.2	-215.2	6.8
4	6	6	31.1	-196.2	21.6
5	8	4	28.0	-164.9	8.6
6	5	3	29.2	-114.8	15.1
7	6	2	27.4	-70.9	11.3
8	7	3	30.8	-53.5	-1.7
9	8	8	28.8	-5.4	-1.6
10	7	3	30.0	44.3	-4.2
11	9	5	28.4	71.7	4.7
12	7	3	21.6	103.5	-15.8
13	8	4	28.7	165.9	-10.9
14	5	5	32.2	196.5	-24.7
15	7	5	29.4	214.5	-6.6
16	4	2	28.3	255.0	-20.5
17	3	3	27.8	288.9	-22.0
18	8	8	40.1	322.6	-10.4

## References

- [1] Case, D. A., Darden, T. A., III, T. E. C., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Hayik, B. S., Roitberg, A., Seabra, G., Kolossvy, I., K.F.Wong, Paesani, F., Vanicek, J., X.Wu, Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., and Kollman, P. A. (2008) AMBER 10. *University of California, San Francisco, CA*.
- [2] Li, G. and Widom, J. (2004) Nucleosomes facilitate their own invasion. *Nature Struct. Mol. Biol.*, **11**(8), 763–769.
- [3] Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005) Rapid spontaneous accessibility of nucleosomal DNA. *Nature Struct. Mol. Biol.*, **12**(1), 46–53.
- [4] Koopmans, W., Brehm, A., Logie, C., Schmidt, T., and van Noort, J. (2007) Single-pair FRET microscopy reveals mononucleosome dynamics. *J. Fluoresc.*, **17**, 785–795.
- [5] Anderson, J. D. and Widom, J. (2000) Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *J. Mol. Biol.*, **296**, 979–987.
- [6] Widom, J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, **34**(3), 269–324.
- [7] Leffak, I. M. (1983) Stability of the conservative mode of nucleosome assembly. *Nucleic Acids Res.*, **11**, 2717–2732.
- [8] Muthurajan, U. M., Bao, Y., Forsberg, L. J., Edayathumangalam, R. S., Dyer, P. N., White, C. L., and Luger, K. (2004) Crystal structures of histone Sin mutant nucleosomes reveal altered protein-DNA interactions. *EMBO J.*, **23**, 260–271.
- [9] Davey, C. A., Sargent, D. F., Luger, K., Mader, A. W., and Richmond, T. J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, **319**, 1097–1113.
- [10] Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K., and Zhurkin, V. B. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**(3), 725–738.
- [11] McQueen, J. (1967) Some methods for classification and analysis of multivariate observations, In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Page 281-297, University of California Press, Berkeley and Los Angeles.

- [12] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. H., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., and Olson, W. K., et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 781–791.
- [13] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci., U.S.A.*, **95**, 11163–11168.
- [14] Czapla, L., Swigon, D., and Olson, W. (2006) Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theor. Comp.*, **2**, 685–695.
- [15] Shimada, J. and Yamakawa, H. (1984) Ring-closure probabilities for twisted worm-like chains: application to DNA. *Macromol.*, **17**, 689.
- [16] Horowitz, D. S. and Wang, J. C. (1984) Torsional rigidity of DNA and length dependence of the free energy of DNA supercoiling. *J. Mol. Biol.*, **173**, 75–91.
- [17] Heath, P. J., Clendenning, J. B., Fujimoto, B. S., and Schurr, J. M. (1996) Effect of bending strain on the torsion elastic constant of DNA. *J. Mol. Biol.*, **260**, 718–730.
- [18] Olson, W. K., Colasanti, A. V., Czapla, L., and Zheng, G. (2008) Insights into the sequence-dependent macromolecular properties of DNA from base-pair level modeling. In G. A. Voth, Editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Page 205–223, Taylor and Francis Group, Boca Raton, FL.
- [19] Zhurkin, V. B., Lysov, Y. P., and Ivanov, V. I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, **6**, 1081–1096.
- [20] Matsumoto, A. and Olson, W. K. (2002) Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophys. J.*, **83**, 22–41.
- [21] Lu, X.-J. and Olson, W. K. (2003) DNA: a software package for the analysis, rebuilding, and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
- [22] Lu, X.-J. and Olson, W. K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding, and visualization of three-dimensional nucleic-acid structures. *Nature Protoc.*, **3**, 1213–1227.
- [23] Barlow, D. J. and Thornton, J. M. (1983) Ion-pairs in proteins. *J. Mol. Biol.*, **168**, 867–885.
- [24] Kumar, S. and Nussinov, R. (1999) Salt bridge stability in monomeric proteins. *J. Mol. Biol.*, **293**, 1241–1255.
- [25] Bednar, J., Horowitz, R. A., Dubochet, J., and Woodcock, C. L. (1995) Chromatin conformation and salt-induced compaction: three-dimensional structural information from cryoelectron microscopy. *J. Cell Bio.*, **131**, 1365–1376.
- [26] Carruthers, L. M., Bednar, J., Woodcock, C. L., and Hansen, J. C. (1998) Linker histones stabilize the intrinsic salt-dependent folding of nucleosomal arrays: mechanistic ramifications for higher-order chromatin folding. *Biochemistry*, **37**(42), 14776–14787.

- [27] Hansen, J. C. (2002) Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 361–392.

## Chapter 8

### Histone tails enhance distant communication in chromatin

#### 8.1 Introduction

Enhancer action over a large distance clearly requires use of special facilitating mechanisms — DNA sequences separated by more than 1 kb do not communicate efficiently on linear DNA in *vitro* [1, 2]. These studies define a “short distance” as 0.1-1 kb. Efficient communication between regulatory elements within this range does not necessarily require special “facilitating” mechanisms. In contrast, the vast majority of eukaryotic genes is regulated by transcription enhancers (TEs) — short DNA sequences that after binding of proteins can activate transcription over variable distances (up to hundreds kb). Action of eukaryotic TEs involves direct interaction between proteins bound at an enhancer and its target (promoter) with accompanying formation of a large loop, including the intervening chromatin-covered DNA [3, 4].

While both kinetic and equilibrium aspects of communication over various distances on histone-free DNA have been modeled and extensively studied experimentally [5, 6, 7, 8, 9], the mechanisms of communication in chromatin remain poorly understood. Recently the Studitsky group has developed experimental approaches allowing quantitative analysis of the rate of enhancer-promoter communication (EPC) in chromatin in vitro [10, 11] (See Fig. 8.1 for a sketch of the experimental system). They have demonstrated that the assembly of relaxed or linear DNA templates into chromatin that is sub-saturated with nucleosomes strongly facilitates distant enhancer-promoter communication and that the observed effect cannot be explained only by DNA compaction [12].

Understanding long-range communication requires knowledge of both the equilibrium

properties and the kinetics of the enhancer-promoter search. Earlier studies have focused on a few of the many facets of this structural complexity, for example, the interplay between the length of linker DNA and the entrance-exit angle of the DNA wrapped on the histone octamer, in predicting the degree of compaction of chromatin [13]. However, determining the role of other facets, for example, histone tails (effective electrostatic charge, tail modifications, etc.) on the equilibrium configuration of chromatin and the kinetics of enhancer-promoter interaction warrants systematic exploration. To develop quantitative understanding of this system, a multi-scale approach is indispensable owing to computational complexity. Effective coarse-grained models of chromatin, built by extracting the essentials from detailed modeling, could surmount the problem of computational complexity [14].

In this study, in order to identify the most important large-scale chromatin features, we performed Monte Carlo simulations of several different coarse-grained representations of chromatin and explored the relative likelihood of EPC in short nucleosome-bound and unbound chains. Our computational studies suggest that transient electrostatic internucleosomal interactions mediated by the N-terminal “tails” of core histones could strongly facilitate EPC communication over a large distance and provide a distant-independent component during enhancer action. These predictions were experimentally evaluated using a recently developed experimental system that allows for quantitative analysis of distant EPC on physiologically relevant, saturated arrays of regularly spaced, precisely positioned nucleosomes [11]. In agreement with the predictions of our computational studies, fully saturated arrays support highly efficient, distance-independent EPC over distances from 0.7 to at least 4.5 kb. We find that the histone N-terminal tails of the core histones are essential for efficient distant EPC in chromatin. Taken together, the data suggest that transient, electrostatic internucleosomal interactions mediated by histone tails are essential for highly efficient, distance-independent, long-range communication between regulatory elements and their targets in eukaryotic chromatin.

## 8.2 Methods

### 8.2.1 Long-range enhancer-promoter (E-P) interactions

Enhancer-promoter communication is a variant of the classic DNA cyclization problem (Fig. 8.2). The ease of long-range communication is proportional to the likelihood that the two protein-bound segments of DNA come into appropriate contact. Given the detailed structure of the enhancer-promoter-DNA assembly, one can define base-pair step parameters that relate the ends of the enhancer- and promoter-bound DNA fragments and a ring-closure probability term analogous to the classic Jacobson-Stockmayer cyclization factor [15]. The latter quantity depends upon the fraction of configurations that meet the selected chain-closure criteria [16], namely the product of probability densities that (i) the vector  $\mathbf{r}$  between the bound DNA fragments terminates in the desired location, (ii) the base-pair normals are correctly aligned, given that  $\mathbf{r}$  adopts the requisite value, and (iii) the overall twist between enhancer and promoter sites assumes the desired value, given that the normals are aligned and the components of  $\mathbf{r}$  are right. In the absence of detailed structural knowledge of the enhancer-promoter assembly, we approximate the proteins as geometric objects consistent with currently available information. The activator protein, nitrogen regulatory protein C (NtrC), which is bound at the enhancer site (Fig. 8.2), forms a heptameric ring-like assembly,  $\sim 124 \text{ \AA}$  in diameter by  $\sim 40 \text{ \AA}$  high [17], which is treated as a cylinder of the same proportions. Similarly, given that RNA polymerase is a claw-shaped object,  $\sim 150 \text{ \AA}$  long by  $110 \text{ \AA}$  wide and  $115 \text{ \AA}$  tall [18], the protein is modeled as an prolate spheroid of similar dimensions.

### 8.2.2 E-P distance

The global structure of chromatin is evaluated with products of generator matrices, as introduced in Chapter 7 (Eqn. 7.7). The E-P distance is measured by the magnitude of displacement from the centers of the enhancer- and promoter-binding proteins mediated by the structure of chromatin. A generator matrix is assigned to express the spatial arrangement of each protein with respect to the DNA to which it is bound:  $\mathbf{A}_{D \rightarrow P}$  relates the first base-pair of DNA to the frame of the promoter-binding protein,

whose local coordinate system is defined so that the  $Z$ -axis lies along the longest axis of the prolate spheroid that mimics the shape of the RNA polymerase assembly;  $\mathbf{A}_{E \rightarrow D}$  expresses the coordinate frame of the NtrC enhancer-binding protein in the local reference frame of the last base-pair of DNA. The local coordinate reference frame on the enhancer is defined in such a way that the  $Z$ -axis lies along the cylindrical axis of the activator protein assembly. Thus, the product of generator matrices needed to find the spatial arrangement of the enhancer and promoter proteins is,

$$\mathbf{A}_{E \rightarrow P} = \mathbf{A}_{D \rightarrow P} \mathbf{A}_c \mathbf{A}_{E \rightarrow D}, \quad (8.1)$$

where  $\mathbf{A}_{E \rightarrow P}$  is the generator matrix of the enhancer-binding protein expressed in the reference frame of the promoter-binding protein;  $\mathbf{A}_c$  is the overall generator matrix of the chromatin, including DNA, that links the enhancer and promoter binding sites.

The center-to-center distance between enhancer- and promoter- proteins can then be obtained by

$$r_{E-P} = |(\mathbf{A}_{E \rightarrow P})_{\mathbf{r}}|, \quad (8.2)$$

where, the subscript  $\mathbf{r}$  in the right term refers to the displacement vector of the generator matrix.

### 8.2.3 Radius of gyration

The radius of gyration is a quantity used to characterize the size of an object, a surface, or an ensemble of points. There are many variants of the definition of the radius of gyration. We use the definition of Flory [19]:

$$R^2 = \frac{1}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N L_{ij}^2, \quad (8.3)$$

where,  $L_{ij}$  is the distance between the centers of the  $i$ th and  $j$ th NCPs, which can be obtained by

$$L_{ij} = |(\mathbf{A}_{ni})_{\mathbf{r}} - (\mathbf{A}_{nj})_{\mathbf{r}}|. \quad (8.4)$$

There  $\mathbf{A}_{ni}$  represents the generator matrix of the  $i$ th NCP in the global coordinate frame used to obtain the displacement vector  $\mathbf{r}$ .



### 8.2.4 Neighbor density

A straightforward measurement of compaction is the density of NCPs within a confined space. We count the number of NCPs found in a spherical region of 15-nm radius centered about the origin of a particular nucleosome core particle. The term  $N_m$  denotes the number  $N$  of other NCPs around the  $m$ th NCP. An NCP is said to fall in the region of the  $m$ th NCP if its center lies within the spherical bounds.

## 8.3 Results and discussion

The coarse-grained model described in Chapter 7 is used to obtain information about the spatial distribution of nucleosome-bound DNA from the sizes of Monte-Carlo simulations described in Table 8.1. Fig. 8.3 displays two snapshots of the configuration of a 25-nucleosome chromatin fragment with and without tails on the histone proteins, obtained from the Monte-Carlo simulations. Calculations were conducted for three systems of enhancer- and promoter-bound DNA mediated by various molecules: (1) a tail-containing chromatin fiber; (2) a tailless chromatin fiber; and (3) a free DNA. Chromatin fibers introduced herewith have a fixed length of linker DNA — 30 bp. That is, the nucleosomes are evenly placed along the DNA. For every calculation for a chromatin fiber, we also estimate the distribution of E-P distances mediated by a free DNA, which has exactly the same nucleotide sequence as that in the chromatin fiber or can be hypothetically obtained by removing all histone proteins from the chromatin fiber (See systems in Fig. 8.1).

### 8.3.1 Distribution of E-P distances

Fig. 8.4 displays normalized distributions (the areas under the curves are unity) of chromatin-mediated E-P distances. Enhancer and promoters are mediated by a series of chromatin fibers of various sizes and with different histone-tail contents — containing 4, 7, 13, and 25 nucleosomes with or without histone tails. The Monte-Carlo simulation begins with an initial rest state where the DNA linkers are ideally straight. The E-P distances of these reference systems are also shown in the figure. The separation of

chain ends and the width of the distributions vary with the number of bound nucleosomes. Each distribution profile yields approximately a Gaussian curve, except for that generated by a chain with 4 tail-containing nucleosomes, which shows a bimodal distribution. Other than the 4-nucleosome cases, the peaks of the distance distributions in the tail-containing chromatin chains are shifted towards smaller values than those with the same number of nucleosomes but without histone tails. The displacement of the distribution peaks between the tailless and tail-containing cases with the same fiber length also increases with the number of bound nucleosomes. Fig. 8.5 presents normalized distributions of the distances between enhancer and promoter on free DNA of the same lengths as the chromatin fibers with 4, 7, 13, and 25 nucleosomes, as introduced above. The separation of chain ends increases with the length of DNA, and the variance of the distributions grows very rapidly with the length of DNA. The locations of the peaks in the distribution function shown in Figs 8.4 and 8.5 are summarized in Table 8.2.

### 8.3.2 Simulated enhancement of E-P communication

We assume that an enhancer and promoter pair has a contact probability purely dependent upon the center-to-center distance between the two proteins. The enhancer-promoter communication probability  $W(\text{E-P})$  is then given by,

$$W(\text{E-P}) = cW(r_0), \quad (8.5)$$

where  $c$  is a constant and  $r_0 = 150 \text{ \AA}$  is the maximal distance between the two proteins that allows for possible contacts to occur, given the dimensions of the modeled structures. In other words, if the enhancer and promoter are separated by a distance greater than  $r_0$ , the contact probability will drop to zero. We treat enhancer-promoter communication on free DNA in the same way. The constant  $c$  vanishes when we calculate the ratio of the communication probability in chromatin compared to free DNA. The ratio data are presented in Fig. 8.6.

The smallest system has the higher communication frequency between enhancers and promoters. The level of communication decreases with the number of nucleosomes in

the system. The histone tails apparently enhance the communication probability, with the extent of enhancement growing with the number of nucleosomes. These trends seem to be correlated with the peak displacements in Fig. 8.5, which lead to higher communication rates when the distributions move towards smaller values. This result is also consistent with experimental observations, which show similar trends and are summarized in Table 8.3. The simulated enhancement, however, is greater than the observed values by a factor of 2-4. Consideration of the orientation and positions of promoter and enhancer proteins, may lead to better agreement of simulation with experiment.

### 8.3.3 Distribution of the radius of gyration

We use the radius of gyration to quantify the compaction of chromatin. Give the same DNA sequence and number of bound nucleosomes, a smaller value of the radius of gyration indicates tighter compaction. Fig. 8.7 presents the distributions of the radii of gyration for simulated chromatin fibers of various sizes and histone-tail contents. The average size of the molecular systems, implicated by the peak position of the distributions, grows with the number of nucleosomes. For systems having the same number of nucleosomes, the molecules are more compressed in the tail-containing fibers than their tailless counterparts, as the average radii of gyration are smaller for the systems with tails. The variance (width) of the distributions increases with the number of bound nucleosomes, and appears to be independent of the presence or absence of histone tails. That is, the distributions for tail-containing and tailless systems with the same number of nucleosomes have similar heights and widths of spread. Combined with the result in Fig.8.4, this suggests that the presence of histone tails does not affect the degree of fluctuation of the chromatin fiber. The additional information about the distributions of the radii of gyration given in Table 8.4, shows that the relative shift of the maxima in corresponding distributions (Fig. 8.7) increases with the number of bound nucleosomes. Table 8.5 lists crossing positions of the distributions of the tail-containing and tailless chains with the same number of nucleosomes. The crossing point increases in value with increasing number of bound nucleosomes. Most of the tail-containing chains terminate at distances less than the crossing points and most of

the tailless chains extend to values greater than the crossing points.

### 8.3.4 Nucleosome-nucleosome contacts

We have examined the distributions of E-P distances and radii of gyration. Both these two quantities are based on a measurement at the global level. This section and the following one evaluate the simulated chromatin fiber at the local level. We examine local interactions of nucleosomes by counting how many other nucleosomes fall within a region of radius 15 nm centered on a single nucleosome. We refer to this number, denoted by  $N$ , as the contact-number density. By collecting the contact number density for all individual nucleosomes within every simulated chromatin sample, we obtain the histogram distribution shown in Fig. 8.8 by normalizing over the total number of samples. The tail-containing systems (cyan) have greater populations with higher contact-number densities than the tailless fibers (violet). The dominant contact-number density of the tail-containing chromatin is  $N = 2$ , while that of the tailless system is  $N = 1$ . This observation doesn't take into account the 4-nucleosome system, since in this system there is no way to find more than 4 nucleosomes falling in the defined region of a nucleosome and thus the system might not be appropriate for statistical inference of chromatin compaction. It is interesting to note that the maximum number of other nucleosomes within the 30 nm (diameter) region of a specific nucleosome is about  $N = 6$ . This maxima may be related to the number of nucleosomes in a turn of the chromatin fiber. The contact numbers in the aforementioned histograms (Fig. 8.8) are broken down to a contact map (Fig. 8.9), where the values of the contact numbers for each nucleosome are color coded. In Fig. 8.9, every system displays 4000 representative configurations of chromatin. It is clear, from the yellow, orange, and red colors, that the tail-containing systems can recruit more nucleosomes to the neighborhood of a nucleosome and lead to higher compaction of the system (higher contact numbers).

We also examined contacts between pairs of nucleosomes along the DNA sequence. We ran Monte-Carlo simulations of a 50-nucleosome fiber with a sample size of about  $10^5$ , under both tail-containing and tailless conditions. For every simulated sample, we collected the following data: (1) the number of each nucleosome along the DNA sequence

as  $n = 1, 2, \dots, 50$ ; (2) the identity of a contact between any two nucleosomes,  $n$  and  $n + m$ ,  $m = 1, 2, \dots, 49$ , by checking if the center-to-center distance is less than 15 nm. Then for all the simulated chromatin samples, we calculated the frequency of sequential contacts  $w(n, n+m)$  for every specific pair of nucleosomes, i.e.,  $n$  and  $n+m$  for particular values of  $n$  and  $m$ . Fig. 8.10 displays the frequency of sequential contacts respectively for tail-containing and tailless system. The tail-containing system has more contacts between nucleosomes spaced by larger sequential intervals than the tailless system, which indicates that the former system is more compacted. We also found the total frequency of sequential contacts by fixing the interval  $m$  and varying the nucleosome index  $n$ . That is, for any  $m = 1, 2, \dots, 49$ , a total frequency  $W(m)$  is calculated between any pair of nucleosomes with a sequential interval  $m$ :  $W(m) = \sum_{n=1}^{50-m} f(n, n+m)$ . Fig. 8.11 displays the total frequency of sequential contacts  $W(m)$  against the interval  $m$ . The tail-containing system exhibits a bimodal distribution, with peaks at  $m = 1$  and  $m = 3$ . The first peak ( $m = 1$ ) suggests that the linker DNA is greatly bent in the fiber such that two successive nucleosomes are brought closer. The second peak ( $m = 3$ ) highlights the observation that a pair of nucleosomes separated by another 3 nucleosome often come close together in the simulated fiber, as shown in the inset in Fig. 8.11. For the tailless system, the curve has a peak at  $m = 2$ , which is consistent with the corresponding typical configuration as shown in the figure.

### 8.3.5 Nucleosome-nucleosome flexibility

Since we have treated nucleosome core particles as rigid bodies and have also defined a reference frame resting on each core particle, it becomes straightforward to express the spatial relationship of two NCPs using six rigid-body parameters: two angles of bending, two in shearing displacement, one angle of twisting, and one stretching displacement. [See Appendix ?? for mathematics and definitions.] Such parameters can be used to characterize localized fluctuations of nucleosome pairs. We calculated the rigid-body parameters for all consecutive pairs of NCPs ( $n$ th and  $n + 1$ th nucleosomes) across all computationally generated chromatin samples.

Fig. ?? - 8.13 respectively present the distributions of stretching, shearing, and bending between two consecutive nucleosomes. The tailless systems show unimodal distributions, while the tail-containing systems exhibit bimodal distributions in the translational parameters (shearing and stretching). Bending is greater for systems with histone tails than those without, as evident from the peak values of the distributions. The fluctuations of adjacent nucleosome pairs are not apparently affected by the number of nucleosomes in the system. The distributions of stretching, shearing, and bending appear similar for various numbers of bound nucleosomes with other conditions held the same.

## 8.4 Conclusion

Action over a distance is a hallmark of eukaryotic transcriptional regulation; however the mechanism of communication between widely spaced DNA modules in chromatin remains a mystery. Our molecular modeling studies suggest that transient binary intranucleosomal interactions could mediate distance-independent communication in chromatin. Detailed modeling indicates that electrostatic interactions between the positively charged N-terminal “tails” of core histones and DNA could strongly increase the probability of juxtaposition of DNA regions that are distantly spaced within the 30-nm chromatin fiber. Experimental analysis of the rates of communication in chromatin confirm that communication over a large distance occurs efficiently only on tail-containing, but not on tailless chromatin. Our studies suggest that electrostatic internucleosomal tail-DNA interactions are essential for highly efficient, distance-independent, long-range communication between regulatory elements and their targets in eukaryotic genome.

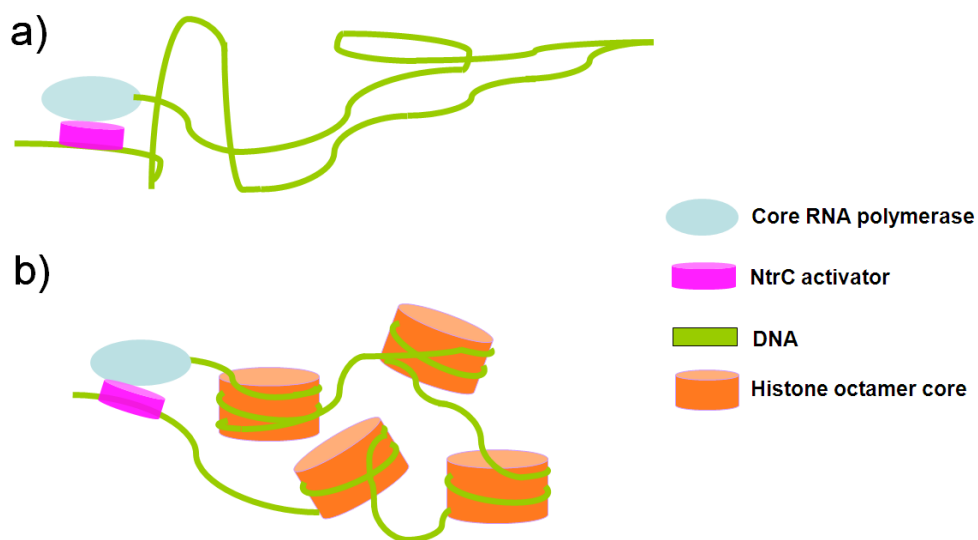


Figure 8.1: The experimental systems used for the measurement of transcription-activation levels in both chromatin and naked DNA systems, with various lengths of DNA and numbers of uniformly spaced nucleosome core particles.

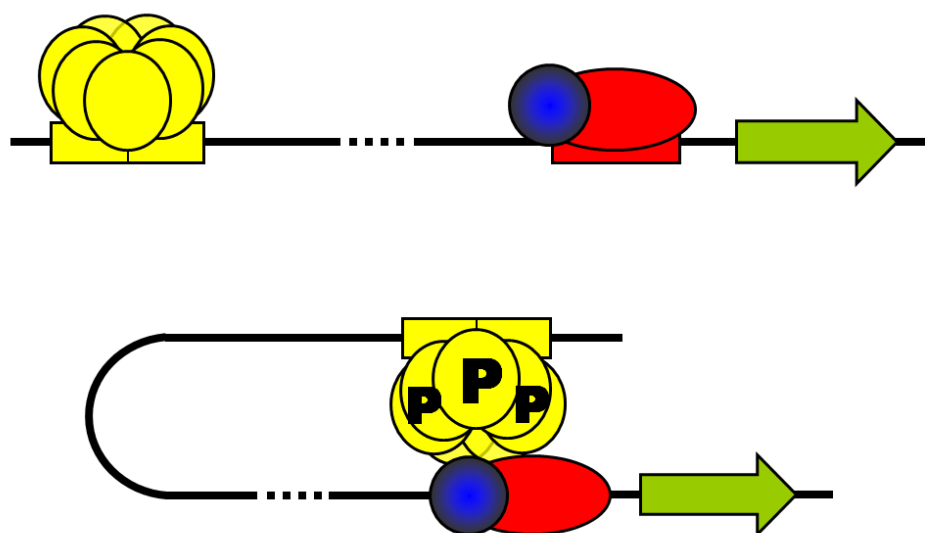


Figure 8.2: Schematic of the looping of DNA mediated by long-range interactions of the heptameric NtrC activator protein assembly (yellow) bound at the enhancer site (also yellow) and the  $\sigma^{54}$  domain (blue circle) of the RNA polymerase (red ellipsoid) complex bound at the *glnAp2* promoter site (also red). The green arrow denotes the gene activated upon complex formation.

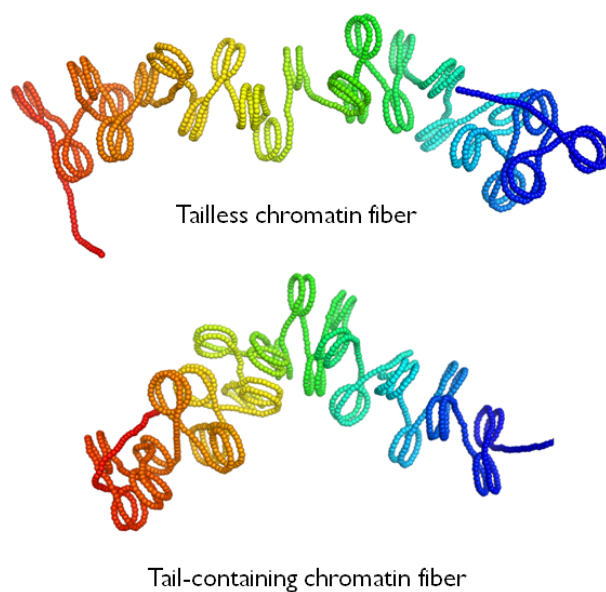


Figure 8.3: Configurations of two 25-nucleosome chromatin fibers. The top image is a chromatin fiber without histone tails, and the bottom one is an assembly with histone tails. For simplicity, only DNA is shown in the snapshots without. The missing histone proteins would lie inside the superhelical segments of DNA. The DNA is represented by spheres, each of which covers three base pairs, and is color-coded using the spectrum of colors so that the trajectory of the sequence can be easily tracked.



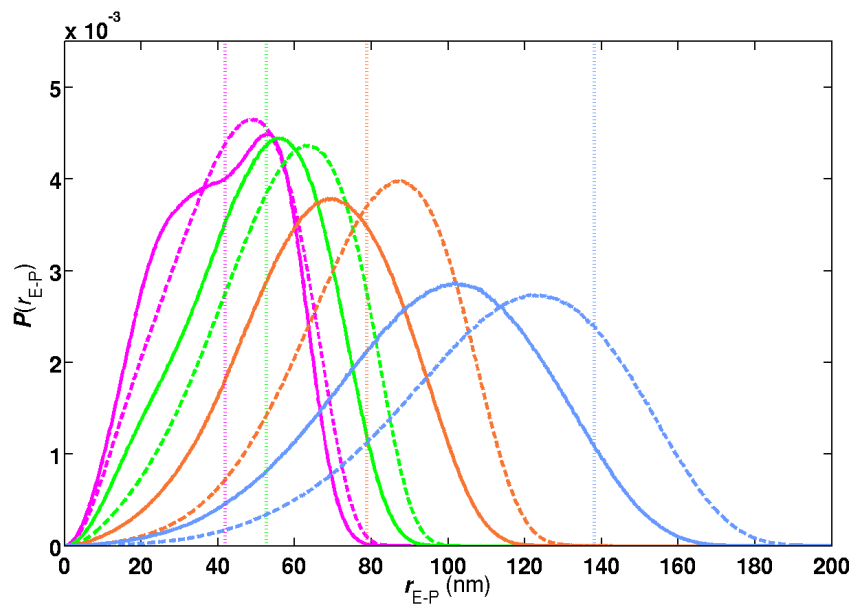


Figure 8.4: Distributions of E-P distances mediated by simulated chromatin fibers containing 4, 7, 13, and 25 evenly spaced nucleosomes with or without histone tails: dashed lines denote tailless systems, and solid lines tail-containing systems. The size of the chromatin fiber is color coded: magenta for 4-nucleosomes; green for 7-nucleosomes; orange for 13-nucleosomes; and blue for 25-nucleosomes. The vertical dotted lines with the corresponding colors mark the E-P distances in the equilibrium rest states where all linker DNAs are ideally straight.

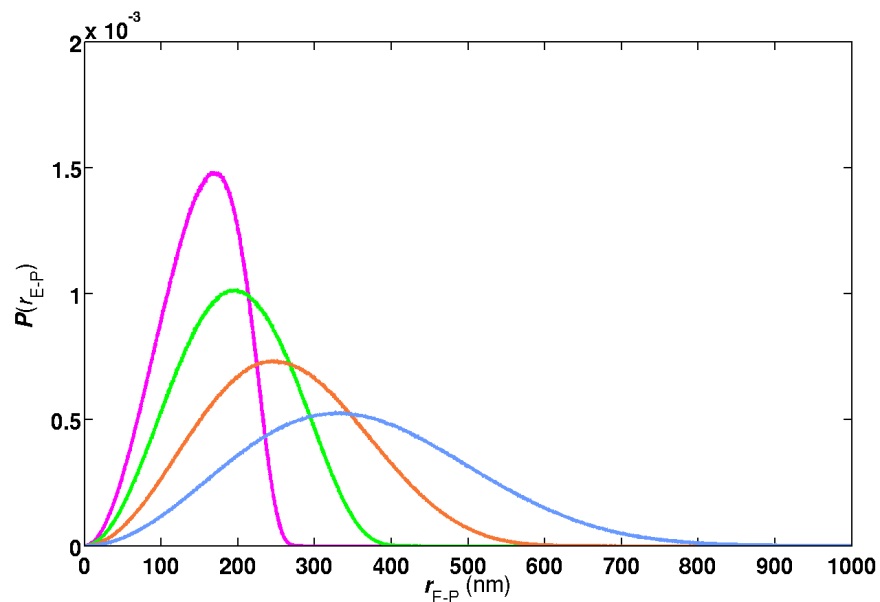


Figure 8.5: Distributions of E-P distances mediated by simulated protein-free DNA chains with the same sequences and lengths found in the corresponding chromatin fibers. The free DNA has exactly the same nucleotide sequence as that in one of the chromatin fibers or can be hypothetically obtained by removing all histone proteins from the chromatin fiber. Color codes are as follow: magenta for free DNA corresponding to the 4-nucleosome chromatin; green for free DNA corresponding to the 7-nucleosome chromatin; orange for free DNA corresponding to the 13-nucleosome chromatin; and blue for free DNA corresponding to the 25-nucleosome chromatin

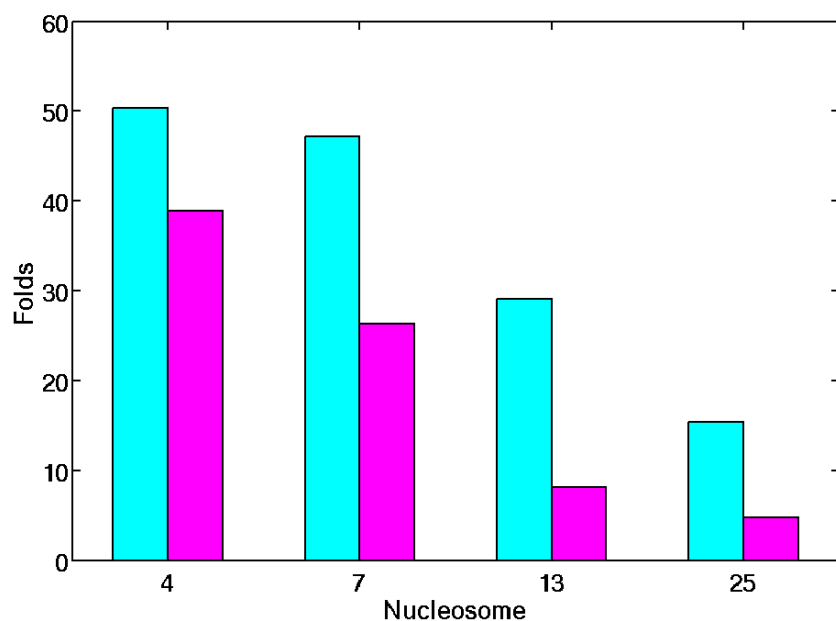


Figure 8.6: Relative likelihood of looping probability for chromatin- vs. free-DNA chains. The cyan bars correspond to tail-containing systems, and the violet ones to tailless systems.

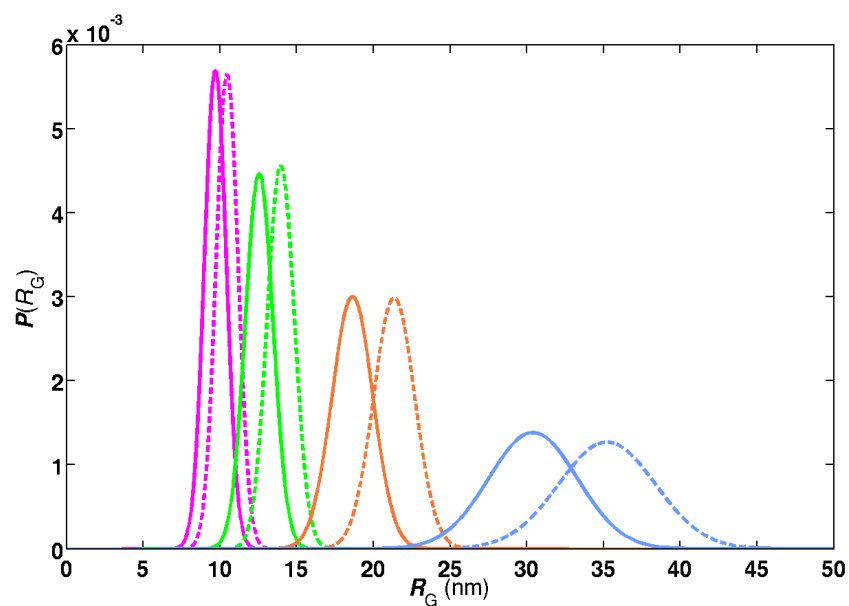


Figure 8.7: Distributions of radius of gyration. The line styles and color codes are the same as those in Fig. 8.4.

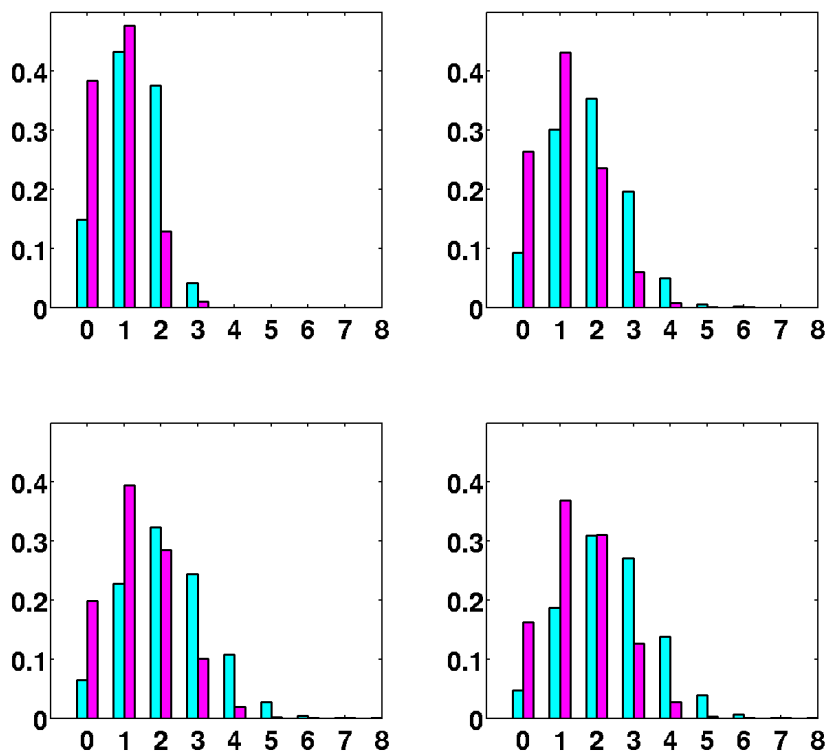


Figure 8.8: Contact number density histograms. The cyan bars correspond to the tail-containing system, and the violet ones to the tailless system. Top-left: 4-nucleosome system; top-right: 7-nucleosome system; bottom-left: 13-nucleosome system; bottom-right: 25-nucleosome system. The  $X$ -axis is the number  $N$  of other nucleosomes falling within the 15-nm region around a nucleosome. The  $Y$ -axis is the frequency of the number density observed for every nucleosome in all simulated chromatin samples.

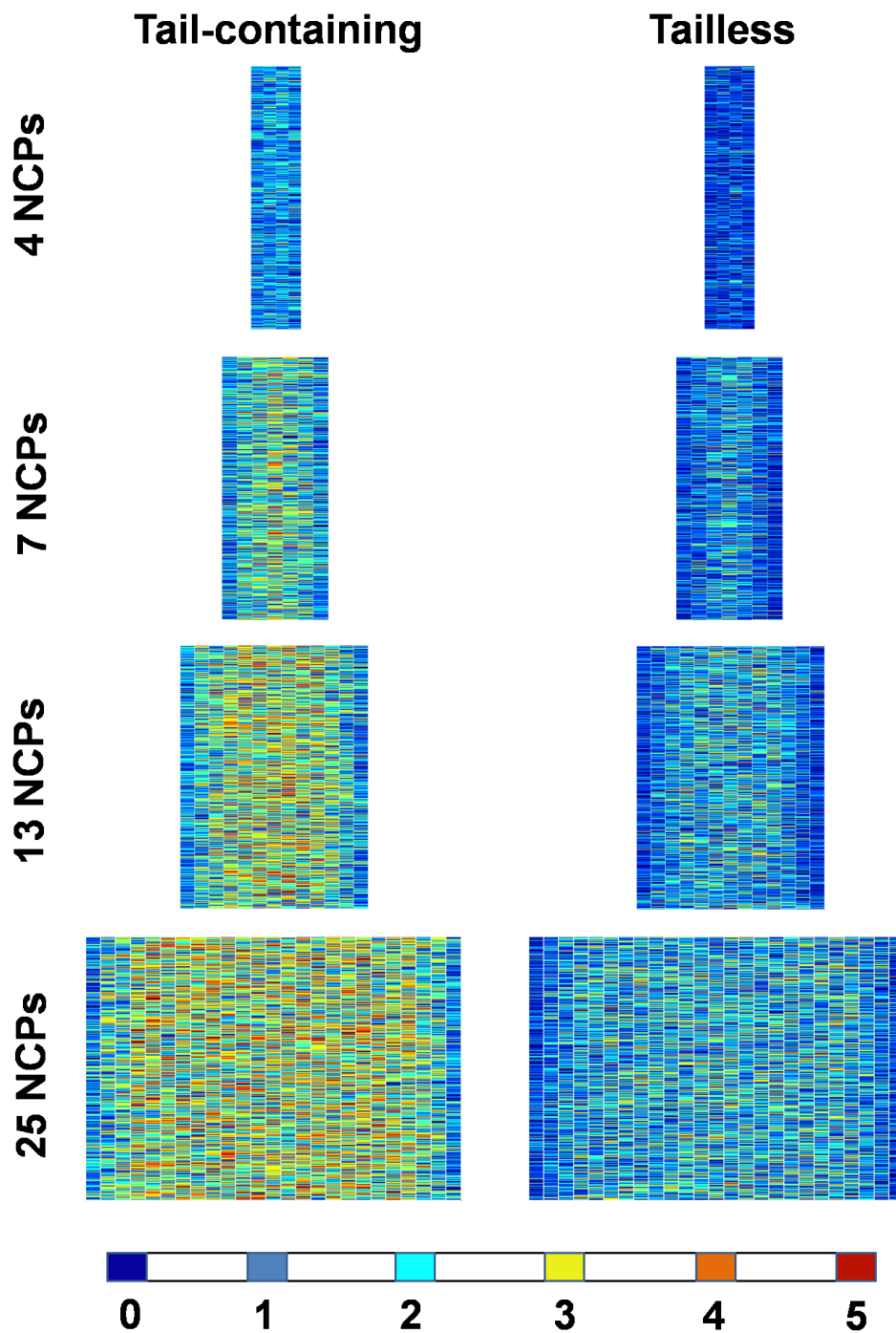


Figure 8.9: Break-down of the contact-number densities for each nucleosome in a selected ensemble of samples (4000 entries). Each column corresponds to the nucleosome in the same sequential position along DNA, i.e., the  $n$ th column contains values for the  $n$ th nucleosome in the simulated chromatin. Each row presents the outcome of a simulated chromatin sample. The color coding is discrete as all the contact numbers are integers.

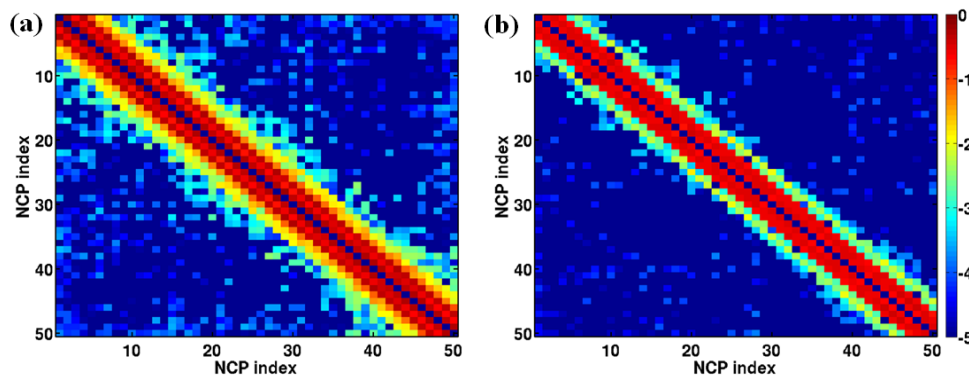


Figure 8.10: Frequency maps of sequential contacts between nucleosomes within simulated 50-nucleosome chromatin fibers. The maps are plotted with the data collected as follows: (1) determine the number of each nucleosome along the DNA sequence as  $n = 1, 2, \dots, 50$ ; (2) identify the contacts between any two nucleosomes,  $n$  and  $n + m$ ,  $m = 1, 2, \dots, 49$ , by checking if the center-to-center distance is less than 15 nm; (3) calculate the frequency  $w(n, n + m)$  of contact among all simulated samples for every specific pair of nucleosomes; (4) transform the frequency to a logarithm scale and map the data. The left panel is for the tail-containing system, the right one for the tailless system.

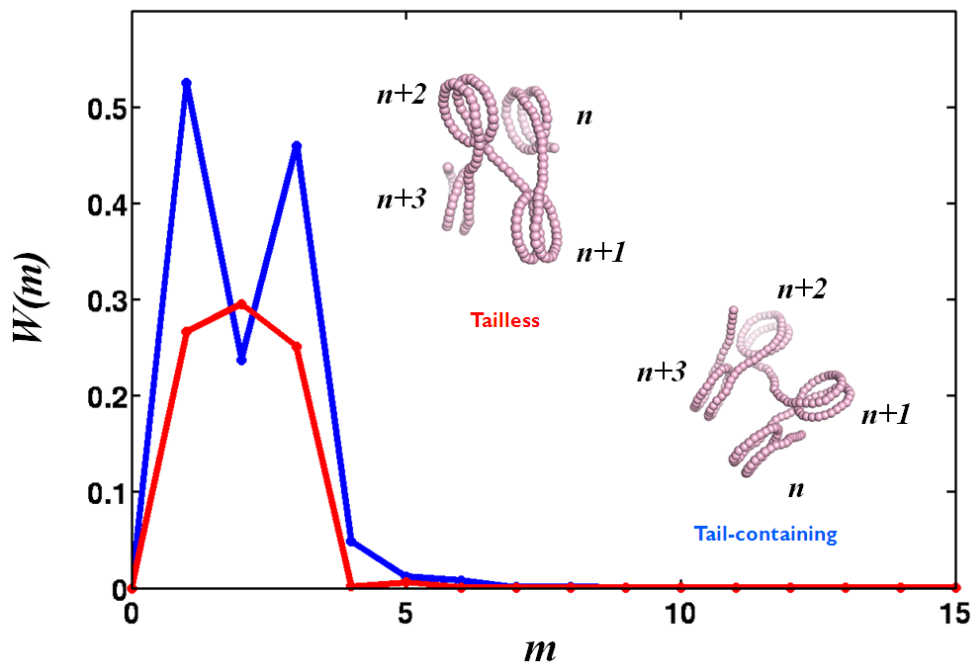


Figure 8.11: Frequency of sequential contacts plotted against the sequential separation interval  $m$ . The curves are plotted using the data described in Fig. 8.10. For any  $m = 1, 2, \dots, 49$ , a total frequency  $W(m)$  between any pair of nucleosomes with a sequential interval  $m$  is calculated:  $W(m) = \sum_{n=1}^{50-m} f(n, n+m)$ . The blue solid line is for the tail-containing system, the red one for the tailless system. The embedded images present typical configurations for a segment of 4 nucleosomes truncated from the simulated 50-nucleosome fibers, respectively for the tail-containing and tailless systems.

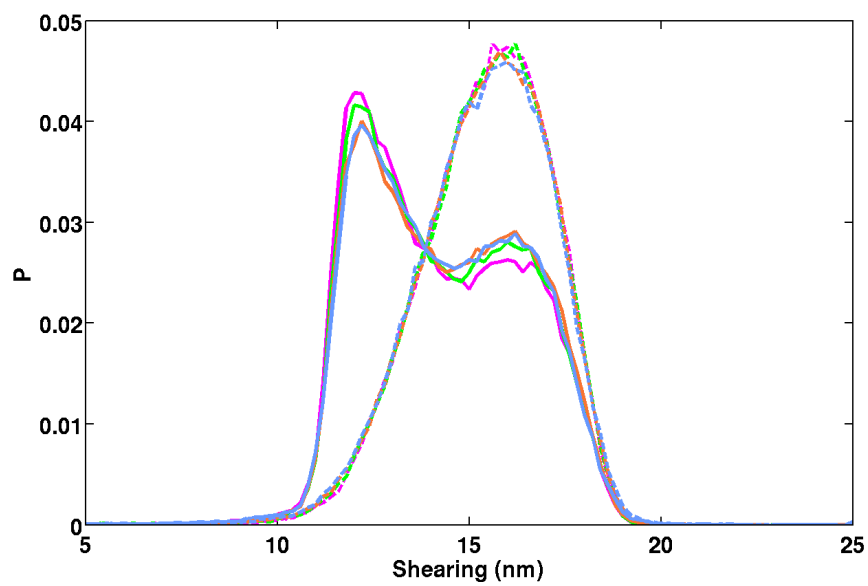


Figure 8.12: Distribution of the shearing between consecutive nucleosomes. The line styles and color codes are the same as those in Fig. ??.

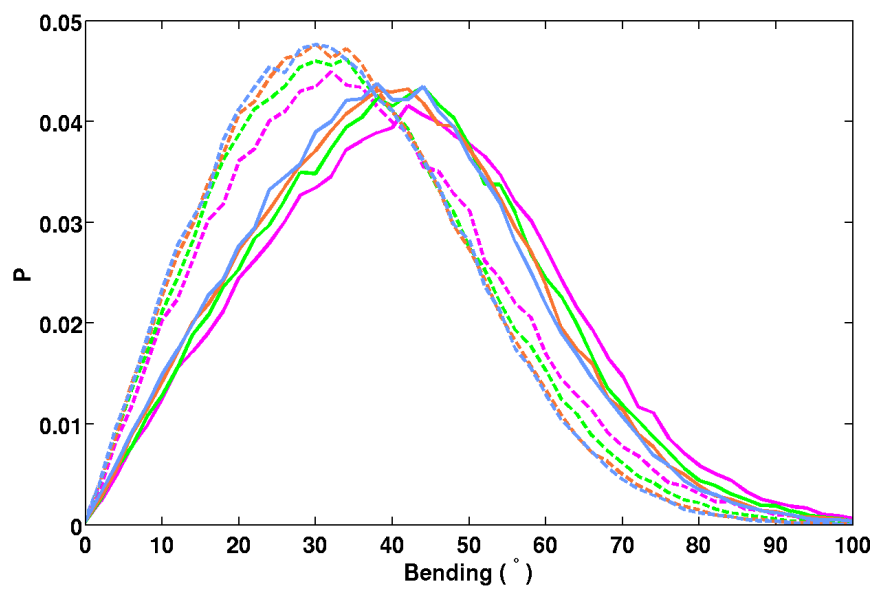


Figure 8.13: Distribution of the bending of consecutive nucleosomes. The line styles and color codes are the same as those in Fig. ??.



Table 8.1: Size of Monte-Carlo simulations of chromatin-mediated systems

System	Accepted chromatin configurations	Total E-P samples
4-nsm +tail	4,235,301	318,654,734
4-nsm -tail	4,417,759	621,151,744
7-nsm +tail	2,752,754	214,935,728
7-nsm -tail	2,785,156	389,892,742
13-nsm +tail	1,598,205	124,668,492
13-nsm -tail	1,574,171	220,206,197
25-nsm +tail	1,246,000	63,597,876
25-nsm -tail	777,877	108,861,130

Table 8.2: The positions of the peaks in the distribution of E-P distances in Figs. 8.4 and 8.5. Peaks are measured in units of nm.

System	+tail chromatin	-tail chromatin	free DNA
4-nsm	53.3	48.5	165.5
7-nsm	55.7	62.9	197.5
13-nsm	68.3	87.5	241.3
25-nsm	101.5	123.9	332.5

Table 8.3: Comparison of the simulated enhancement of the E-P communication shown in Fig. 8.6, with the experimentally observed values.

System	+tail chromatin		-tail chromatin	
	Simulated	Experiment	Simulated	Experiment
4-nsm	52	13	41	10.5
7-nsm	48	12.5	28	7
13-nsm	31	9.5	9	3
25-nsm	17	9	5.5	2

Table 8.4: Peak positions and relative shifts in the distribution curves shown in Fig.

8.7. Values are in units of nm.

System	+tail peak	-tail peak	shift
4-nsm	9.7	10.5	0.8
7-nsm	12.6	14	1.4
13-nsm	18.6	21.4	2.8
25-nsm	30.4	35.2	4.8

Table 8.5: The partition of the distribution function of the radii of gyration of tail-containing and tailless chromatin chains on either side of the crossing points between the two curves for each length of chromatin.

System	Crossing point (nm)	+tail chromatin		-tail chromatin	
		Left	Right	Left	Right
4-nsm	10.1	71%	29%	29.8%	70.2%
7-nsm	13.3	78.5%	21.5%	21%	79%
13-nsm	20	84.7%	15.3%	15.5%	84.5%
25-nsm	32.9	80.4%	19.6%	22.7%	77.3%

## References

- [1] Liu, Y., Bondarenko, V., Ninfa, A., and Studitsky, V. M. (2001) DNA supercoiling allows enhancer action over a large distance. *Proc. Natl. Acad. Sci., U.S.A.*, **98**(26), 14883–14888.
- [2] Bellomy, G. R. and Record, Jr., M. T. (1990) Stable DNA loops *in vivo* and *in vitro*: roles in gene regulation at a distance and in biophysical characterization of DNA. *Prog. Nucleic Acid Res. Mol. Biol.*, **39**, 81–128.
- [3] Bondarenko, V. A., Jiang, Y. I., and Studitsky, V. M. (2003) Rationally designed insulator-like elements can block enhancer action in vitro. *EMBO J.*, **22**(18), 4728–4737.
- [4] de Laat, W., Klous, P., Kooren, J., Noordermeer, D., Palstra, R. J., Simonis, M., Splinter, E., and Grosveld, F. (2008) Three-dimensional organization of gene expression in erythroid cells. *Dev. Biol.*, **82**, 117–139.
- [5] Shore, D., Langowski, J., and Baldwin, R. L. (1981) DNA flexibility studied by covalent closure of short fragments into circles. *Proc. Natl. Acad. Sci., U.S.A.*, **78**(8), 4833–4837.
- [6] Vologodskii, A. V. and Cozzarelli, N. R. (1996) Effect of supercoiling on the juxtaposition and relative orientation of DNA sites. *Biophys. J.*, **70**, 2548–2556.
- [7] Huang, J., Schlick, T., and Vologodskii, A. (2001) Dynamics of site juxtaposition in supercoiled DNA. *Proc. Natl. Acad. Sci., U.S.A.*, **98**(3), 968–973.
- [8] Polikanov, Y. S., Bondarenko, V. A., Tchernachenko, V., Jiang, Y. I., Lutter, L. C., Vologodski, A., and Studitsky, V. M. (2007) Probability of the site juxtaposition determines the rate of protein-mediated DNA looping. *Biophys. J.*, **93**(8), 2726–2731.
- [9] Czaplá, L., Swigon, D., and Olson, W. K. (2008) Effects of the nucleoid protein HU on the structure, flexibility, and ring-closure properties of DNA deduced from Monte Carlo simulations. *J. Mol. Biol.*, **382**, 353–370.
- [10] Polikanov, Y. S., Rubtsov, M. A., and Studitsky, V. M. (2007) Biochemical analysis of enhancer-promoter communication in chromatin. *Methods*, **41**(3), 250–258.
- [11] Polikanov, Y. S. and Studitsky, V. M. (2009) Analysis of distant communication on defined chromatin templates *in vitro*. *Methods Mol Biol.*, **543**, 563–576.
- [12] Rubtsov, M. A., Polikanov, Y. S., Bondarenko, V. A., Wang, Y. H., and Studitsky, V. M. (2006) Chromatin structure can strongly facilitate enhancer action over a distance. *Proc. Natl. Acad. Sci., U.S.A.*, **103**(47), 17690–17695.

- [13] Langowski, J. (2006) Polymer chain models of DNA and chromatin. *Eur. Phys. J. E. Soft Matter*, **19**, 241–249.
- [14] Arya, G. and Schlick, T. (2006) Role of histone tails in chromatin folding revealed by a new mesoscopic oligonucleosome model. *Proc. Natl. Acad. Sci., U.S.A.*, **103**, 16236–16241.
- [15] Jacobson, H. and Stockmayer, W. H. (1950) Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.*, **18**, 1600–1606.
- [16] L. Czapla, D. S. and Olson, W. K. (2006) Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theor. Comp.*, **2**, 685–695.
- [17] Lee, S. Y., de la Torre, A., Yan, D., Kustu, S., Nixon, B. T., and Wemmer, D. E. (2003) Regulation of the transcriptional activator NtrC1: structural studies of the regulatory and AAA+ ATPase domains. *Genes Dev.*, **17**, 2552–2563.
- [18] Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., and Kornberg, R. D. (2001) Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Angstrom resolution. *Science*, **292**(5523), 1876 – 1882.
- [19] Flory, P. (1969) *Statistical Mechanics of Chain Molecules*, Interscience, New York.

## Appendix A

### Determination of nucleosome step parameters

In our models, the nucleosome is modeled as a rigid body, so that, for any two nucleosomes there exist six degrees of freedom to describe their spatial relationship, including three rotational variables and three translational variables. We define a set of nucleosome step parameters to describe those degrees of freedom: two angles of bending ( $\xi_1$ ,  $\xi_2$ ), two in shearing displacement ( $\xi_4$ ,  $\xi_5$ ), one angle of twisting ( $\xi_3$ ), and one stretching displacement ( $\xi_6$ ), analogous to the definition of base-pair step parameters of DNA [1] (See Fig. A.1). For any two planes, given the origins ( $\mathbf{O}_i$ ,  $\mathbf{O}_{i+1}$ ) and orientations ( $\mathbf{A}_i$ ,  $\mathbf{A}_{i+1}$ ), the six step parameters  $\{\xi_i; i = 1, 2, \dots, 6\}$  can be determined following published procedure to analyze DNA base-pair step parameters [2, 3].

(1) The orientation of a nucleosome/cylinder, is described by the local unit vectors  $\mathbf{X}$  and  $\mathbf{Y}$  in the plane of a circular cross section, with the  $X$  axis pointing toward the dyad, and the local normal unit vector  $\mathbf{Z}$  along the cylindrical axis. To determine the step parameters between nucleosomes  $i$  and  $i + 1$ , we first want to find the line of interaction of the two ( $X$ ,  $Y$ ) planes ( $i$  and  $i + 1$ ). This line, also called the hinge axis ( $\mathbf{h}$ ), can be obtained from the vector product of the two normal cylindrical axes

$$\mathbf{h} = \frac{\mathbf{Z}_i \times \mathbf{Z}_{i+1}}{|\mathbf{Z}_i \times \mathbf{Z}_{i+1}|}. \quad (\text{A.1})$$

In this step, we also find the angle between the two planes, i.e., the angle between the two normal/cylindrical axes,

$$\varphi = \cos^{-1}(\mathbf{Z}_i \cdot \mathbf{Z}_{i+1}). \quad (\text{A.2})$$

(2) We next rotate the two planes about the hinge axis with two different angles:  $+\frac{\varphi}{2}$

for plane  $i$ , and  $-\frac{\varphi}{2}$  for plane  $i+1$ . The rotation is carried out by applying rotation matrices to the local coordinate axes.

$$\mathbf{A}_i^* = \mathbf{R}_h(+\frac{\varphi}{2})\mathbf{A}_i; \mathbf{A}_{i+1}^* = \mathbf{R}_h(-\frac{\varphi}{2})\mathbf{A}_{i+1}. \quad (\text{A.3})$$

Here the function  $\mathbf{R}_v(\theta)$  is a matrix describing a rotation of magnitude  $\theta$  along an arbitrary unit vector  $\mathbf{v} = (v_1, v_2, v_3)$ , with elements defined by:

$$R_{ij} = (1 - \cos \theta)v_i v_j - \sin \theta \sum_{k=1}^3 \epsilon_{ijk} v_k + \delta_{ij} \cos \theta, \quad (\text{A.4})$$

where,  $\delta_{ij}$  is the Kronecker delta function and  $\epsilon_{ijk} = \pm 1$  when  $i, j, k$  are even or odd permutations of 1, 2, 3, respectively, and vanishes otherwise. After such rotations, the two planes become parallel, with the transformed axis  $\mathbf{Z}_i^*$  parallel to  $\mathbf{Z}_{i+1}^*$ .

(3) A “middle” plane can then be constructed by averaging the two sets of local axes obtained after rotation, i.e.,

$$\mathbf{V}_m = \frac{\mathbf{V}_i^* + \mathbf{V}_{i+1}^*}{|\mathbf{V}_i^* + \mathbf{V}_{i+1}^*|}, \quad (\text{A.5})$$

where,  $\mathbf{V} = \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ . The orientation of the “middle” plane can be expressed as a  $3 \times 3$  matrix,

$$\mathbf{A}_m = \begin{bmatrix} \mathbf{X}_m & \mathbf{Y}_m & \mathbf{Z}_m \end{bmatrix}. \quad (\text{A.6})$$

(4) The translational step parameters (Shift, Slide, Rise) correspond to projections of the displacement vector between the two nucleosome centers onto the “middle” plane:

$$(\xi_4, \xi_5, \xi_6) = (\mathbf{O}_{i+1} - \mathbf{O}_i)\mathbf{A}_m. \quad (\text{A.7})$$

(5) Twist  $\xi_3$  is the angle between  $\mathbf{Y}_i^*$  and  $\mathbf{Y}_{i+1}^*$ , with the same sign as  $\mathbf{Q} = (\mathbf{Y}_i^* \times \mathbf{Y}_{i+1}^*) \cdot \mathbf{Z}_m$ , i.e.,

$$\xi_3 = \frac{\mathbf{Q}}{|\mathbf{Q}|} \cos^{-1}(\mathbf{Y}_i^* \cdot \mathbf{Y}_{i+1}^*), \quad (\text{A.8})$$

(6) Tilt and Roll are defined with respect to the phase angle ( $\psi$ ) between the hinge axis ( $\mathbf{h}$ ) and the Y axis on the “middle” plane ( $\mathbf{Y}_m$ ) as

$$\xi_1 = \varphi \cos(\psi), \xi_2 = \varphi \sin(\psi), \quad (\text{A.9})$$

where  $\psi = \cos^{-1}(\mathbf{h} \cdot \mathbf{Y}_m)$ .

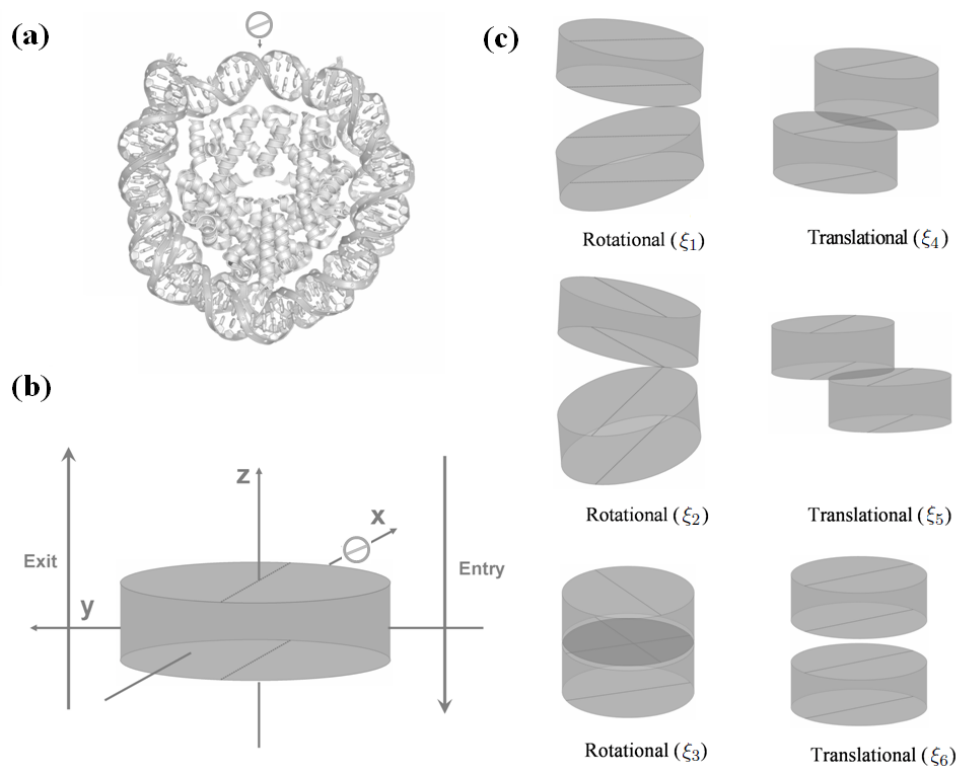


Figure A.1: Illustration of nucleosome step parameters. (a) A top-down view of NCP-147 [4]. (b) The NCP is modeled as cylinder with a reference frame defined as in Chapter 6. The structural dyad is noted by the “ $\odot$ ” symbol. (c) The six rigid-body parameters that relate the coordinate frames on adjacent nucleosome core particles: three rotational angles,  $\xi_1$ ,  $\xi_2$ ,  $\xi_3$ , rotating a nucleosome respectively along the  $\mathbf{X}$ -,  $\mathbf{Y}$ -,  $\mathbf{Z}$ - axes of the other; three translational displacement,  $\xi_4$ ,  $\xi_5$ ,  $\xi_6$ , moving a nucleosome respectively along the  $\mathbf{X}$ -,  $\mathbf{Y}$ -,  $\mathbf{Z}$ - axes of the other.



## References

- [1] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. H., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., and Olson, W. K., et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 781–791.
- [2] Lu, X. J. and Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
- [3] Lu, X. J. and Olson, W. K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protoc.*, **3**, 1213–1227.
- [4] Davey, C. A., Sargent, D. F., Luger, K., Mader, A. W., and Richmond, T. J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, **319**, 1097–1113.

## Vita

### Guohui Zheng

- 1999-03** University of Science and Technology of China, Hefei, China.
- 2003** B.S. in Space Physics, University of Science and Technology of China.
- 2003-05** Johns Hopkins University, Baltimore, MD.
- 2005** M.A. in Physics and Astronomy, Johns Hopkins University.
- 2005-09** Rutgers, The State University of New Jersey, New Brunswick, NJ.
- 2009** M.S. in Mathematics, Rutgers University.
- 2010** Ph.D in Computational Biology and Molecular Biophysics, Rutgers University.
- 2006** W. K. Olson, A. V. Colasanti, Y. Li, W. Ge, G. Zheng and V. B. Zhurkin. (2006). DNA simulation benchmarks as revealed by X-ray structures. In J. Sponer and F. Lankas, Editors, *Computational Studies of RNA and DNA*, Page 235-237, Springer, Dordrecht, The Netherlands.
- 2008** M. A. Karymov, M. Chinnaraj, A. Bogdanov, A. R. Srinivasan, G. Zheng, W. K. Olson and Y. L. Lyubchenko. (2008). Structure, dynamics, and branch migration of a DNA Holliday junction: a single-molecule fluorescence and modeling study, *Biophys. J.*, 95(9), 4372-4383.
- 2008** W. K. Olson, A. V. Colasanti, L. Czapla and G. Zheng. (2008). Insights into the sequence-dependent macromolecular properties of DNA from base-pair level modeling. In G. A. Voth, Editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Page 205-223, Taylor and Francis Group, Boca Raton, FL.
- 2009** G. Zheng, X.-J. Lu, and W. K. Olson. (2009). Web 3DNA — a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.*, 37(Web Server issue): W240-W246.
- 2009** G. Zheng, A. V. Colasanti, X.-J. Lu, and W. K. Olson. (2009). 3DNALandscapes: a database for exploring the conformational features of DNA. *Nucleic Acids Res.*, In press.
- 2009** G. Zheng, L. Czapla, A. R. Srinivasan, and W. K. Olson. (2009). How stiff is DNA? *Phy. Chem. Chem. Phys.*, In press.