MECHANICAL ANALYSIS OF COLLAGEN AND DNA

A Dissertation

by

XIAOJING TENG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Wonmuk Hwang
Committee Members,	Roland R. Kaunas
	Alvin T. Yeh
	Steven E. Wheeler
Head of Department,	Anthony Guiseppi-Elie

December 2016

Major Subject: Biomedical Engineering

ABSTRACT

It is known that mechanics plays a central role in many biological events. Tissue can remodel and turnover to adapt to new mechanical environment, such as hypertension and exercising. During the remodeling, hydrolysis of collagen is a key step. It is found that extension will change the cleavage rate of both collagen monomers and fibrils. The specificity of the collagen cleavage site is explained as the unique local mechanical environment of the cleavage site. DNA is another important filament molecule, and its behavior is also regulated by mechanics. The sequence-dependence of mechanical property has been observed, and is related to the specific interaction between proteins and DNA.

On the pursuit of understanding the role of mechanics in those biological events as well as connecting atomistic to mesoscale properties of biopolymers, we used molecular dynamics (MD) simulation to study collagen and DNA. In collagen study, from the local bending stiffness calculated around cleavage site, we found it is transitioned from stiff to flexible across the cleavage site, which agrees with the classic model and can be seen as the structural feature recognizable by MMPs. We showed that the α-chain registry can determine the local conformation of collagen, and hence the cleavability of collagen. The resistance of homotrimer form to hydrolysis is interpreted as the stabilization role of arginines downstream to the cleavage site. Homotrimer form is found mainly in fetal tissue and carcinomas, and related to osteogenesis imperfecta. This resistance mechanism can help people to better understand its role in these processes. We further resolved controversial findings in experiments regarding the relationship between extension and collagen cleavage rate published the same year on the same journal. By mimicking the pulling conditions in the experiments, we found it is their different ways of pulling that induces different conformations, and therefore, different relationship of cleavage rate vs extension.

This indicates the importance of mechanical environment on collagen.

In our DNA investigation, we further developed our triad method to make it being capable for local isotropic mechanics study. We demonstrated the mechanical property is mainly determined at the dinucleotide-level sequence. The sequence-dependent flexibility can be applied to mechanical property prediction of any DNA sequence, as well as DNA nanostructures construction. We found the overwhelmingly used helicoidal parameters are not suitable for dynamic study, due to their degeneracy in describing conformational changes. Based on our data, we built a coarse-grained model that can capture the mechanical properties measured in experiments. This model bridges the atomistic dynamics and mesoscale property of DNA. By using the obtained stiffness and equilibrium data, we calculated energy of crystal structures of dsDNA-protein complexes without non-standard bases and paring. The results provided quantitative insight into the DNA-protein interaction dynamics. We further analyzed DNA methylation, a fundamental epigenetic modification that generates profound impact on gene regulation. We showed methylation generally will cause the immediate neighbor steps to be stiffer, whereas the methylated step itself less affected in mechanics. This is mainly because the steric interaction between methyl groups of methylated cytosine with other groups. We also demonstrated the hydration distribution change upon methylation could play a role in the stiffness variation, as well as affect the binding affinity to different proteins, since hydration force is key in molecular interactions. The findings in this study display influence of methylation in high resolution, and are potentially helpful to elucidate the mechanism of methylation in gene regulation.

Currently we are investigating interaction between kinesin-1 motor head and tubulin. Its dimer or tetramer form can walk unidirectionally on microtubules (MTs), in an out-of-phase manner. The motion can be attributed to different binding affinity when pulled in different directions and various nucleotide binding modes. We will simulate those different conditions to understand the atomistic mechanism.

DEDICATION

To my mother, my father, and my grandparents.

ACKNOWLEDGMENTS

First of all, I would like deeply appreciate my advisor Dr. Wonmuk Hwang, for his patient guidance and persistent help throughout my PhD study. During the six years in his lab, he kept motivating me to challenge new projects and learn new techniques, and inspiring me with fresh ideas. From his hard-working, being cautious on every detail in research, and strict time arrangement, I learned a lot on how to be a competent researcher. Under his advisement, I also progressed hugely on presentation and communication with other people, which will definitely benefits my future career. I am also grateful for the opportunities he provided me to share my research in many conferences, especially the wonderful trip to Korea. I believe under his instruction, I have become a better research and person.

I thank the other committee members, Dr. Roland R. Kaunas, Dr. Alvin T. Yeh, and Dr. Steven E. Wheeler, for their valuable suggestions to improve my research. I also thank Dr. Jay D. Humphrey and Dr. Kumbakonam R. Rajagopal for their insightful and enjoyable courses.

I would like to thank my lab members, Krishna, Kaushik, Weewen, Andrew, Esma, Amol, Tyler, Jie, Eric, Bruce, and Ana. They helped me on my presentation and poster preparation. I also learned a lot from discussion with them. I am very thankful to Krishna and Kaushik, for their help on studying Linux and molecular dynamics simulation when I just came to the lab, as well as their previously developed analysis methods that I can refer and further develop. My first project was about collagen, which was based on Krishna's preliminary research. Thanks for his excellent job, I can carry out mine smoothly.

My life at Collage Station is colorful, largely because of the company of my friends, particularly my soccer teammates. It is a real joy to play every game with them.

Lastly, I would like to thank my family, for my wife's accompanying with me most of my PhD life, and especially my parents, for their continuous supporting not only financially, but also mentally at my depression periods. Without their support, I couldn't accomplish what I have done today.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Wonmuk Hwang, Professor Roland R. Kaunas, and Professor Alvin T. Yeh of the Department of Biomedical Engineering, and Professor Steven E. Wheeler of the Department of Chemistry.

The coding of Coarse-grained model of DNA in Chapter III was accomplished by Professor Wonmuk Hwang.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by family, U.S. National Institutes of Health grant R01GM087677, and teaching assistantship from Department of Biomedical Engineering and First-Year Engineering Program of Dwight Look College of Engineering of Texas A&M University.

NOMENCLATURE

MD	molecular dynamics
MMP	matrix metalloproteinase
ECM	extracellular matrix
GPO	Glycine-Proline-Hydroxyproline
dsDNA	double-stranded DNA
C-G modle	Coarse-grained model
mCYT	5-methyl-cytosine
GBSW	generalized Born with a simple switching
PME	particle-mesh Ewald

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION: PREVIOUS STUDIES ON TYPE-I COLLAGEN AND B-DNA AND MOTIVATION

1.1 Type-I collagen

Collagen is the most abundant proteins in vertebrates, and the major component of the extracellular matrix (ECM). Among the 28 types of collagens identified up-to-date[168, 211], type-I collagen is the most widespread one that can be found in skin, bone, tendon, blood vessel, *etc*. Collagen monomer is composed of three 1000 amino-acids long α-chains. They are twisted by staggering one residue to form a triple-helical structure, being the hallmark of collagen. The monomer of type-I collagen generally contains 2 α1-chains and 1 α2-chain, but homotrimer with 3 α1-chains are also found. Besides its prominent mechanical property to provide strength and stability to the tissue [50] and capability of hierarchical assembly into fibrils and fibers, collagen is able to be digested when tissue needs to remodel during wound healing or to adapt to changing mechanical environment, such as hypertension. Its cleavage is almost exclusively accomplished by matrix metalloproteinases (MMPs).

This specific yet important cleavage event poses several interesting questions. First of all, along the long chain of type-I collagen, there are more than 15 G-I/L bonds [222], but only one of them can be specifically cleaved by MMPs. To explain this specificity, Gregg Fields did a pioneer work and proposed that cleavage site is proceeded by a compact triple-helical region and followed by a loose region [46]. This structural feature makes it recognizable to MMPs. This model is widely accepted. However, the atomistic picture is missing. On the other hand, intact collagen monomer is too large (diameter of 15 Å) for MMPs (active site of 5 Åin size) to hydrolyze [29]. Therefore collagen has to be unwound and open before cleavage. We are going to address how this process occurs.

As we just introduced above, collagen triple helix is formed in a staggering manner, and type-I collagen is a heterotrimer, so there could be three isomers. People have found different chain registries can influence its interaction with von Willebrand factors [20] and osteonectin [68]. It is reasonable to deduce the registry will also affect the cleavability of type-I collagen.

Homotrimer of type-I collagen is found in fetal tissues, fibrotic tissues, and carcinomas [122, 60]. It is also found to be resistant to cleavage [122, 60]. It is thought this homotrimer is resistant to local unwinding [60], but again, the mechanism at the atomistic level is unknown. Researchers using biased potential in umbrella sampling simulation to demonstrate α2-chain to be more easily unwound compared to α1-chain. Therefore, homotrimer without α2-chain is not disrupted. However, since the reaction coordinate they chose is arbitrary, it is hard to explain the results. It is likely homotrimer's resistance to cleavage plays a significant role in those physiological (fetal development) and pathological (cancer) processes. Understanding the underlying mechanism is quite important.

Another mystery was brought up by controversial findings in experiments regarding the relationship between collagen cleavage rate and pulling [1, 24]. One group claimed extension of collagen monomer would increase the cleavage rate [1], whereas the other reported totally contrary trend [24]. By using MD simulation, one explained as it was the difference in the behavior of hetero- *vs*. homotrimers used in the two experiments [25], and it has subsequently been shown that the cleavage rate increases in both cases [2]. We noticed they used slightly different ways of constraints to pull during experiments, which may result in the conflicting results. This further suggests mechanics as a key role in collagen cleavage.

In summary, we used MD simulation to investigate several type-I collagen segments (heterotrimer isomers, homotrimer, and mutants) and control groups, with different pulling schemes, to analyze their various properties, including mechanics, conformations, dynam-

ics, *etc*, to address following questions regarding type-I collagen:

- What is the local mechanics of the transition from compact to loose along cleavage site?
- Since unwinding is necessary for cleavage, how does it happen?
- What is the role of chain registry in the conformation and dynamics of type-I collagen heterotrimer isomers, and how does it affect cleavage?
- The molecular mechanism of the homotrimer's resistance to cleavage.
- How does the different boundary conditions influence collagen's cleavability?

1.2 B-DNA

With the success of our triad method in collagen and other filamentous biomolecules [162, 95, 210], we extended our scope to DNA, the carrier of genetic information of eukaryote. The building blocks of DNA are individual nucleotide composed of one phosphate group, one deoxyribose, and one nucleobase. There are only four different nucleotides, and they differs at the nucleobase, which are adenine (A), thymine (T), cytosine (C), and guanine (G). Hereafter, we use the nucleobases to refer the sequence of DNA. Pairing can form between A and T, C and G, and then two strands can twist into a double-stranded DNA (dsDNA). Several forms of dsDNA have been found physiologically relevant, and they are A-DNA, B-DNA, and Z-DNA, among which B-DNA is the most common one. It features a right-handed helical structure, and a major groove and a minor groove, which can interact with proteins [151, 165].

Sequence-dependent flexibility of DNA is essential for genomic organization, DNA packaging, and positioning of DNA-binding enzymes [65, 137, 213, 182, 53, 17]. It

has been extensively analyzed through surveys of available structures [41, 28] and experiments [66, 125, 157]. However, static crystal structure provides little information in dynamics, and experiments lack the resolution to show the anisotropic motion and dinucleotide-level mechanics. Therefore, MD simulation becomes an important supplementary method [22, 100, 99, 156, 154]. They studied dynamics by using helicoidal parameters [37, 41, 119], which depicts DNA at the atomistic level very well and was designed to describe the geometry of dsDNA. However, their validity as an order parameters set to describe DNA as an elastic rod and hence to obtain a mesoscale model is questionable. In a study, it is found base pair stacking energy was better described by non-standard sets of conformational variables [132], further suggesting that helicoidal parameters may not be ideal for describing energetics. Hence we further developed our triad method to make it being capable of analyzing local anisotropic bending motion. This decomposition of bending is purely from local dynamics motion. Consequently, our analysis is optimal for representing the elasticity of DNA, and the mesoscale model from our analysis would be more accurate.

One intriguing feature of DNA is that the expression regulation can be accomplished by decorations [15, 85, 186]. A fundamental one is the cytosine methylation at the C5 position, which is the most prevalent one in vertebrates. Only the one occurs at the CG/CG step is heritable [200, 186]. It can lead to gene suppression [166, 14, 127], chromosome inactivation [153, 207], and genomic imprinting [111, 83]. Abnormal hyper-methylation of the CpG-rich region (the CpG island), is frequently observed in cancer cells [82, 84, 10]. Its heritability highlights the importance of methylation in development and disease progression.

Although the functionality of methylation is well understood, its mechanism is not known yet. The newly added methyl group may prevent the binding of transcription factors [220, 208, 72] or allows binding of other proteins such as methyl-CpG-binding domain

proteins that blocks transcription [18, 202]. Methylation can alter chromatin structure, thereby restrict accessibility of transcription factors [105, 139]. However, controversial results are reported. We think mechanics plays a central role in this process. Flexibility significantly influences protein binding to DNA [88], and packing of DNA such as in nucleosome formation [155, 139]. This dependence on local DNA sequence needs to be established. On the other hand, hydration force is known to be a fatal factor in biomolecular interaction [201, 26, 164]. Therefore, it is also worth of investigation by a method developed in our group [164].

Here is the key questions and applications we will study in this dissertation:

- Sequence-dependent anisotropic mechanical property of B-DNA.
- Whether our method is superior to traditionally used helicoidal parameters set in describing DNA as an elastic rod?
- Building a coarse-grained model of DNA from our mechanics analysis, to connect atomistic properties with mesoscale model.
- How does the methylation change the B-DNA properties, in mechanics and hydration distribution?

2. CHAIN REGISTRY AND LOAD-DEPENDENT CONFORMATIONAL DYNAMICS OF COLLAGEN*

2.1 Introduction

Collagens possess distinct properties as the main building blocks of the extracellular matrix. They assemble hierarchically into near-macroscopic order, making up both soft and hard tissues [19, 221]. Of 28 known types, fibrillar collagens including type-I, II, and III, are dominant [168]. To achieve structural diversity and larger-scale compliance while maintaining precise local order within the extracellular matrix, fibrillar collagens adopt residue-specific interactions [19] as well as other less specific interactions such as water-mediated force that also exists in other biopolymers [108, 109, 164]. Such balance between crystallinity and disorder [221] likely applies to other types of collagens as well. The 'order-and-disorder' features are based on the domain organization within a single collagen molecule, which consists of the stable imino-rich (Pro or Hyp) domains for which the representative structural motif is GPO (Gly-Pro-Hyp; O is the single-letter code for hydroxyproline), and 'labile' domains where X and Y in the GXY triplet are not imino acids [133]. The imino-rich domain is thermally stable due to the constraint on the backbone dihedral angle imposed by the imino rings, which prefers the polyproline type II conformation of the α chain in a collagen triple helix [9, 93, 204]. Hydroxyproline in the Y position of the GXY motif provides further stabilization due to a stereo-electronic effect that favors the α chain backbone dihedral angles in the triple helical conformation [191, 192] and also via possible water-mediated hydrogen bond (H-bond) formation [161, 9, 141, 204]. Labile domains are thought to be more loosely wound and flexible

*Reprinted with permission from "Chain registry and load-dependent conformational dynamics of collagen" by Xiaojing Teng and Wonmuk Hwang, 2014. *Biomacromolecules*, 15, 3019-3029, Copyright [2014] by American Chemical Society.

compared to the imino-rich domain [91, 133, 163]. The collagen cleavage site hydrolyzed by MMP is located in a labile domain, about $\frac{3}{4}$ along the length of the molecule [46, 205]. Since a well-folded collagen triple helix is highly resistant to protease cleavage [46, 61], local unfolding of the labile domain is critical for cleavage by MMP [46, 29, 123]. Our earlier study using molecular dynamics (MD) simulation showed that, in the case of an imino-poor domain of type-III collagen, unwinding initiates at a typical cleavage bond (Gly-Ile) at temperatures as low as 300 K [163]. Spontaneous unwinding of the labile domain is likely implicated in the instability of isolated type-I collagen molecules at body temperature [110], and it may also contribute to recognition and additional disruption of the triple helix by MMPs [29, 123].

Apart from the general picture for the collagen molecule as a whole, far less is known regarding the mechanism by which the individual α chains forming a collagen triple helix affect the conformational behavior. This is especially important for type-I collagen, a heterotrimer comprised of two $\alpha 1$ chains and one $\alpha 2$ chain. Herein, we call the $\alpha 1$ and $\alpha 2$ chains simply as $\alpha 1$ and $\alpha 2$, respectively. Compared to the native heterotrimer, a homotrimer comprised of three $\alpha 1$ is more stable [122], assembles less efficiently [130, 129], and is more resistant to MMP cleavage [122]. The $\alpha 1$ homotrimer is found in fetal tissues, fibrotic tissues, and carcinomas [122, 60], and is implicated in osteogenesis imperfecta [27]. Comparing the primary structure, $\alpha 1$ has a net charge of $+11e$ ($e = 1.6 \times 10^{-19}$ C) and 64 large non-polar residues (Ile, Leu, Met, Phe, Tyr, and Val). $\alpha 2$ has $+31e$ and 106 large non-polar residues (for comparison, sequences for the triple helix part in the Uniprot P02452 for $\alpha 1$ and P08123 for $\alpha 2$ were used). More non-polar residues in $\alpha 2$ would mean greater hydrophobic attraction, promoting assembly, whereas a higher net charge may keep the molecule hydrated, thus it may allow axial sliding of collagen molecules in a bundle that is crucial for proper ordering [109, 164]. On the other hand, $\alpha 2$ has smaller number of imino acids, which supports its destabilizing role. However, beyond the sequence-level

information, structural mechanisms for different α chains in modulating the stability and conformation of a collagen triple helix are unclear. Since the three α chains in a collagen triple helix is staggered by one residue [9], three isomers of type-I collagen are possible depending on whether $\alpha 2$ is in the leading (the most N-terminal side), middle, or trailing position. While a modest-resolution (5.16 Å) x-ray fiber diffraction structure of rat tail tendon suggests that $\alpha 2$ is in the middle [149], a systematic study of the dependence of the conformational properties on chain registry is lacking. A related issue is the load-dependent cleavage of collagen by MMP. It is generally accepted that collagen fibrils under tensile load are more resistant to cleavage [73, 135, 174, 117, 89, 13, 48, 58]. However, single-molecule experiments yielded conflicting results, with cleavage rate either decreased [24] or increased by as much as 100-fold [1]. While one suggestion was the difference in the behavior of hetero- *vs*. homotrimers used in the two experiments [25], it has subsequently been shown that the cleavage rate increases in both cases [2]. One of the difficulties in studying collagen is its long length ($\sim$300 nm) that is organized into different domains for numerous ligand binding and signaling [205]. Model collagen mimetic peptides (also called triple helical peptides) have thus been instrumental for analyzing behaviors of specific sub-domains or chain registry [92, 178, 175, 112, 79, 47]. They also have potential for biomedical applications [224, 146].

Here we use MD simulations of various collagen mimetic peptides containing the MMP cleavage domain of type-I collagen, to systematically analyze its properties. We find that chain registry plays a critical role for the stability and flexibility of the triple helix. A heterotrimer with $\alpha 2$ in the leading position behaves similar to the stable $\alpha 1$ homotrimer, despite the general destabilizing role of $\alpha 2$. The inter-chain H-bond formed by the arginine side chain, together with clustering of non-polar residues, is a major determinant for the registry dependence, in agreement with experiment [175]. The heterotrimer with $\alpha 2$ in the middle is mechanically the most labile at and downstream to the MMP cleavage site,

suggesting that this isomer may be the most prone to cleavage. The imino-rich domain upstream to the MMP cleavage site is unwound but is stiff, supported by long-lived H-bonds. The MMP cleavage domain is thus characterized by a rapid transition in stiffness and stability. In contrast, the backbone H-bond occupancy and lifetime for the stable GPO peptide is much smaller. The rapidly forming H-bonds allow the GPO peptide to remain flexible while maintaining the triple helical structure. We also find that the conformational behavior and mechanical response of the triple helix depend sensitively on how loads are applied to the ends of the molecule. The loading-condition dependence addresses recent debates about whether mechanical load increases [1, 2] or decreases [24] the MMP cleavage rate of a monomer. Present results elucidate dynamic versus static mechanisms for stabilizing the collagen triple helix and their relation to mechanics. Furthermore, simple 'rules of thumb' such as regarding $\alpha 2$ as generally destabilizing, or the stabilizing role of arginine, should be exercised with caution.

2.2 Methods

2.2.1 Peptide generation

We used 30-residue long α chains to build collagen-like peptides. Residues 7–24 have the corresponding sequence from the MMP cleavage domain of human type-I collagen (residues 766–783 [222], with the cleavage site between 775 and 776) (Table 2.1). Residues 1–6 and 25–30 are GPO triplets that stabilize the ends [91]. For comparison, we also considered α chains made only of the GPO triplet. The mutant chains $\alpha 1_{(\mathrm{R21O})}$ and $\alpha 2_{(\mathrm{O21R})}$ had Arg21 of $\alpha 1$ and Hyp21 of $\alpha 2$ switched, to investigate the role of arginine for the triple helix stability. The five α chains in Table 2.1 were used to build the triple helices in Table 4.1. Backbones of the triple helical structures were built using the THeBuScr program [160] and sides chains were added by using CHARMM [21].

Triad				2	4	6	8	10	12	14	16	18	20		
Residue	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29
$\alpha 1$	GPO	GPO	GAO	G	T	P	G	P	Q	**G**	**I**	A	G	Q	R	G	V	V	G	P	O	G	P	O
$\alpha 2$	GPO	GPO	GPO	G	T	O	G	P	Q	**G**	**L**	L	G	A	O	G	I	L	G	P	O	G	P	O
gpo	GPO	GPO	GPO	G	P	O	G	P	O	G	P	O	G	P	O	G	P	O	G	P	O	G	P	O
$\alpha 1_{\text{(R21O)}}$	GPO	GPO	GAO	G	T	P	G	P	Q	**G**	**I**	A	G	Q	*O*	G	V	V	G	P	O	G	P	O
$\alpha 2_{\text{(O21R)}}$	GPO	GPO	GPO	G	T	O	G	P	Q	**G**	**L**	L	G	A	*R*	G	I	L	G	P	O	G	P	O

Table 2.1: Sequences of α chains used [222]. Residues forming the MMP cleavage bond are in boldface. Mutated residues in $\alpha 1_{\text{(R21O)}}$ and $\alpha 2_{\text{(O21R)}}$ are in italic. The first row shows triad numbers that start from residue 6 of the leading chain in a triple helix. Pro12 in $\alpha 1$ is left non-hydroxylated, based on other studies [149, 206].

Name	Leading	Middle	Trailing
$huco1$	$\alpha 2$	$\alpha 1$	$\alpha 1$
$huco2$	$\alpha 1$	$\alpha 2$	$\alpha 1$
$huco3$	$\alpha 1$	$\alpha 1$	$\alpha 2$
$homo$	$\alpha 1$	$\alpha 1$	$\alpha 1$
$homo2$	$\alpha 2$	$\alpha 2$	$\alpha 2$
$gpo10$	gpo	gpo	gpo
$homo_m$	$\alpha 1_{\text{(R21O)}}$	$\alpha 1_{\text{(R21O)}}$	$\alpha 1_{\text{(R21O)}}$
$homo2_m$	$\alpha 2_{\text{(O21R)}}$	$\alpha 2_{\text{(O21R)}}$	$\alpha 2_{\text{(O21R)}}$
$huco1_m$	$\alpha 2_{\text{(O21R)}}$	$\alpha 1_{\text{(R21O)}}$	$\alpha 1_{\text{(R21O)}}$

Table 2.2: Composition and chain registry of triple helices used in this study.

2.2.2 Basic simulation procedure

For simulation, we used the CHARMM program [21] with the param27 all-atom force field [121] and additional parameters for Hyp [3]. Before solvation, a brief energy minimization (2000 steps) was carried out in the generalized Born with a simple switching (GBSW) implicit solvent model of CHARMM [77]. The peptide was solvated in an orthorhombic box of about $135 \times 55 \times 55$ Å^3. The box size was chosen so that there is at least 20-Å gap between the molecule and the boundary of the box, which is larger than the 12-Å non-bond interaction cutoff. Ions were added to neutralize the system, at approximately 150 mM NaCl and 10 mM $MgCl_2$ [60]. The solvated system was energy minimized again by 1600 steps. Simulation proceeded by heating from 0 K to 300 K for 30 ps followed by equilibration at 300 K for 170 ps. Production runs were either 8 ns or 24 ns, with most measurements done during the last 12 ns of the 24-ns runs. Coordinates were saved every 5 ps. The total simulation time was over 2.5 μs.

2.2.3 Triad-based description of the triple helix conformation

We used local coordinate bases $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ (triads) to describe torsional and bending motion of the molecule along its length [163]. Triads were assigned based on adjacent backbone carbonyl C atoms from each α chain. We chose C since its radial position from the axis of the initial straight triple helix varies less compared to C_α or N atoms, so that the resulting triads are more uniformly aligned. To eliminate end effects, we only considered the region spanning residue 6 to 25 on the leading chain. Due to chain staggering, C atoms of residue 6, 5, and 4 from the leading, middle, and trailing chains, respectively, constitute triad 1, and so on. The three C atoms for each triad make a triangle, whose centroid is the origin of the triad and the unit vector normal to the triangle and pointing to the C-terminus is set as $\boldsymbol{e_3}$. The unit vector pointing from the centroid to the midpoint of the C atoms of the leading and middle chains is $\boldsymbol{e_1}$, which fixes $\boldsymbol{e_2} = \boldsymbol{e_3} \times \boldsymbol{e_1}$. Local torsional angle was

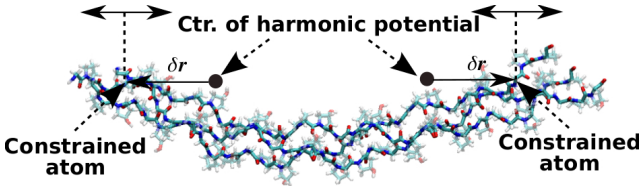

Figure 2.1: Illustration of the simulation with harmonic constraints at the ends of a peptide. Displacement of the constrained C_α atoms relative to the center of the harmonic potential is $\delta \boldsymbol{r}$. The tug-of-war sampling analyzes the fluctuation of $\delta \boldsymbol{r}$ (double arrow) to calculate the force exerted on the constrained atom at the center of the harmonic potential.

measured as the Euler angle between two successive triads relative to $\boldsymbol{e_3}$ [163].

2.2.4 Calculation of mechanical properties

Force-extension relationship. To control extension of the molecule, harmonic potentials were applied to the C_α atoms of G4 and G28 in the leading chain at a given distance (Fig. 2.1). By restraining only one α chain, rotation or unwinding of the triple helix is allowed. In some cases, we restrained ends of all three α chains to study the effect of the loading condition on the conformational behavior of the triple helix. To avoid large abrupt changes in extension, we gradually changed it. Starting with 72-Å (the distance between the restrained atoms in the initially built triple helix), the extension was either increased or decreased in 4-Å intervals with 100-ps equilibration for each, covering 60–84 Å. At each extension, the production run was 8 ns (Fig. 2.2a, open symbols). In the physiologically more relevant region (see Results), we carried out another set of 24-ns simulations in 0.8-Å intervals (Fig. 2.2a,b, solid symbols).

The force exerted by the molecule at a given extension was calculated by using the tug-of-war sampling method [75]. Briefly, if we denote the i-th Cartesian component of the deviation of the restrained atom from the center of the potential during the simulation by δr_i ($i = 1, 2, 3$; Fig. 2.1), the i-th component of the force F_i exerted by the restrained

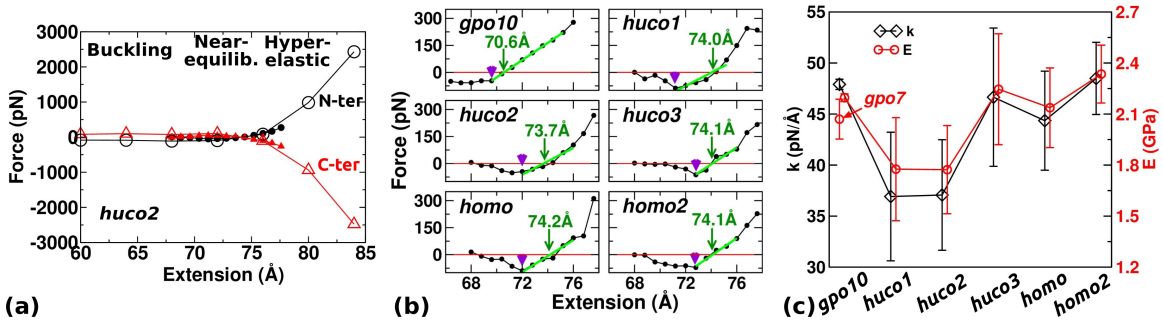

Figure 2.2: Extensional behavior. (a) Overview of the force-extension relation with $huco2$ as an example (see Table 4.1 for peptide names). Open/solid symbol: 8-ns/24-ns simulation in 4-Å/0.8-Å steps. The last half of each simulation period was used to calculate force. N-ter/C-ter: Axial forces exerted by the restrained C_α atoms at 4th/28th residues of the leading chain. Transverse components of the force are very small (less than 10 pN), isotropic, and are independent of extension. (b) Force-extension relations from 24-ns simulations, as solid symbols in (a). Sign of the C-terminal force is reversed and averaged with the N-terminal force. Arrowhead: extension below which buckling occurs. Arrow: L_{eq} where local linear fit to the near-equilibrium regime (thick green line) crosses the zero-force point. (c) Extensional stiffness k (diamond) and Young's modulus E (circle). Young's modulus of $gpo7$ is shown in the $gpo10$ column.

atom at the center of the potential is given by:

$$F_i \simeq -k_B T \left(\frac{\langle \delta r_i \rangle}{\mathrm{var}(\delta r_i)} - \sum_{j \neq i} \langle \delta r_j \rangle \frac{\mathrm{cov}(\delta r_i, \delta r_j)}{\mathrm{var}(\delta r_i)\mathrm{var}(\delta r_j)} \right) \tag{2.1}$$

where $\langle \cdot \rangle$ denotes an average over coordinate frames, k_B is Boltzmann constant, and $T(=300$ K) is temperature. The harmonic potential had a spring constant 10 kcal/(mol·Å^2). This choice does not affect the measured force since Eq. 2.1 is independent of the spring constant [75], which we confirmed by performing simulations using 5 kcal/(mol·Å^2) (Fig. 2.3). For simulations with all three chains restrained, the spring constant was reduced by 1/3.

Bending Stiffness. Local bending stiffness was measured by analyzing the fluctuation in the polar angle of $\boldsymbol{e_3}$ between two triads. Let the deviation of this angle from its average during the simulation by $\delta\theta$, and the distance between two triads by s. The bending

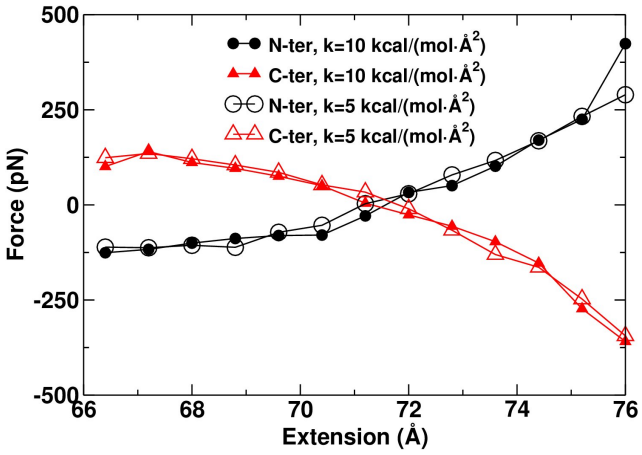

Figure 2.3: Force-extension relation of $gpo10$ measured using two different spring constants of restraining potentials.

stiffness κ_f between these triads is given by [94]

$$\kappa_f = \frac{2k_BTs}{< \delta\theta^2 >}. \tag{2.2}$$

For Eq. 2.2, we carried out simulations without any positional restraints. To prevent self-interaction through the periodic boundary by the freely diffusing peptide, we used a larger cubic box of side length 125 Å. For each peptide, we carried out a 24-ns simulation and used the last 12 ns for calculation.

2.3 Results

2.3.1 Peptide design

Among the peptides tested (Table 4.1), $gpo10$ serves as a stable control. $huco1$, $huco2$, and $huco3$ are the three isomers of the human type-I collagen cleavage domain. $homo$ is an $\alpha1$ homotrimer where the corresponding full-length molecule is known to be stable and resists cleavage. $homo2$ is an $\alpha2$ homotrimer that is expected to be less stable. $homo_m$, $homo2_m$, and $huco1_m$ are designed to test the role of Arg21 on $\alpha1$ (Table 2.1).

2.3.2 Extensional behavior

Our initial force-extension curve based on 8-ns simulation in 4-Å steps possesses approximately three regimes of behavior: buckling, near-equilibrium, and hyper-elastic (open symbols in Fig. 2.2a). In the buckling regime, the molecule takes a bent conformation. The near-equilibrium regime is around the region where the force is close to zero. In the hyper-elastic regime, force increases sharply, beyond physiologically relevant levels (see Discussion). Based on the initial characterization, we carried out 24-ns simulations to obtain refined force-extension curves surrounding the near-equilibrium regime in 0.8-Å steps. They cover 66.4 Å–76.0 Å for $gpo10$ and 68.0 Å–77.6 Å for other peptides. We used the last 12 ns of these simulations for calculating forces (Fig. 2.2b). Compared with 8-ns simulations, forces decrease in magnitude in small and large extensions due to conformational relaxation. Taking $huco2$ as an example, in 8-ns simulation the force in the buckling regime is non-zero (open symbols in Fig. 2.2a), while it decreases to zero in 24-ns simulation (Fig. 2.2b). In the compressed state, there is an extension below which the force does not increase in magnitude (arrowheads in Fig. 2.2b). Below this extension, conformational change occurs, such as breakage of existing hydrogen bonds and/or formation of new contacts. Only $gpo10$ maintains a non-zero force (Fig. 2.2b) as it remains stably wound even in the buckling regime (explained below).

In the near-equilibrium regime, the point where the force-extension curve crosses the zero-force point defines the equilibrium length L_{eq} (Fig. 2.2b, arrows). Except for $gpo10$ ($L_{eq} = 70.6$ Å), it is similar among other peptides ($\sim$74 Å) with $huco2$ being the shortest (73.7 Å). Compared to the initial canonical triple helix (72 Å), $gpo10$ wound more tightly thus became shorter, whereas others containing the labile domain became longer due to unwinding. In the hyper-elastic regime, in addition to conformational relaxation, more extensive unfolding can occur. For example, the reduced force of $huco1$ at the largest

extension (Fig. 2.2b, 77.6 Å) is due to splaying of one of α chains on its C-terminal end at 16.1 ns.

For $gpo10$, the force-extension curve is fairly linear in the near-equilibrium regime. Other peptides show less linear behavior. Nevertheless, it is informative to measure the extensional stiffness k and Young's modulus E to compare with previous estimates. Linear fit to the force-extension curve around L_{eq} gives k (Fig. 2.2c, diamond). Using $r = 7.0$ Å as the radius of a hydrated collagen molecule [163], Young's modulus is given by [70] $E = k\frac{L_{eq}}{\pi r^2}$ (Fig. 2.2c, circle). While k depends on the system size, E is a material property. The calculated E (1.77–2.34 GPa) lies on the lower end of previous experiments [63, 184] and simulations [217, 116], 2.4–9 GPa. The large variation in previous works is due to different experimental methods used and choices for the radius r. The largest value, 9.0 GPa was obtained using inelastic light scattering, which the authors suggested to be an overestimate [63]. Ref. [184] used x-ray diffraction and obtained $E = 2.9$ GPa. They used $r = 6.15$ Å. If we use this radius, our estimate becomes 2.3–3.0 GPa, which agrees well with their result. In simulations, steered molecular dynamics (pulling the molecule with a constant speed) is frequently used [116, 52]. In this case, E tends to be over-estimated due to the lack of conformational relaxation and E increases with the pulling speed [52]. Relaxation can be seen in our simulation by measuring k in 4-ns intervals, which generally decreases with time before 8 ns (Fig. 2.4). If we use the 4–8 ns period, the calculated E indeed increases to 2.2–2.5 GPa. To test independence of E on the length of the peptide, we carried out another set of 24-ns simulations using 7 GPO repeats, gpo_7, whose E is comparable to that for $gpo10$ (Fig. 2.2c).

Among the peptides tested, $huco1$ and $huco2$ possess the smallest k, which is also seen in calculations over 4-ns intervals (Fig. 2.4). Since MMP locally unwinds or deforms collagen for cleavage [29, 11], $huco1$ or $huco2$ may possess the native registry of α chains among the three type-I collagen isomers, which we examine further below.

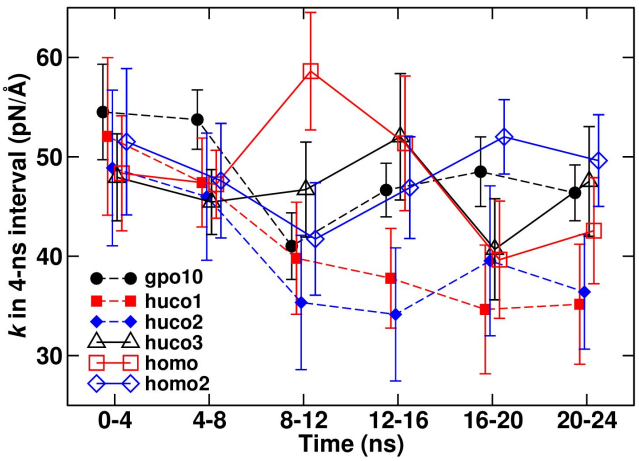

Figure 2.4: Extensional stiffness k calculated in 4-ns intervals. After 8 ns, $huco1$ and $huco2$ take the two lowest values of k. Due to the nonlinear nature of Eq. 2.1, values of k over 12–24 ns (Fig. 2.2c) are not equal to the averages of the last three 4-ns intervals.

2.3.3 RMSD

At each extension, we calculated the root-mean-square deviation (RMSD) of the positions of backbone heavy atoms in the triad region from those at the beginning of simulation (Fig. 2.5). In most cases, RMSD increases during the first few nanoseconds, and stays fluctuating after about 6 ns, which partly supports making measurements during 12–24 ns (Fig. 2.5a). Additional analysis of time scale is in Discussion. As expected, RMSD generally decreases with extension (Fig. 2.5b), although the trend is not strictly monotonic. Comparing different peptides, RMSD of $gpo10$ is the smallest, reflecting its stability.

2.3.4 Local bending stiffness

In simulations for measuring local bending stiffness, no restraint was applied. During 12–24 ns, distances between the 4th and 28th C_α atoms of the leading chains were: 70.9±0.9 Å ($gpo10$; average±standard deviation), 74.4±0.6 (huco1), 74.4±2.3 (huco2), 72.6±1.0 (huco3), 74.0±0.8 (homo), and 72.7±1.0 (homo2), which are comparable to L_{eq} in Fig. 2.2b. To use Eq. 2.2, the interval s between two triads needs to be chosen. If it is

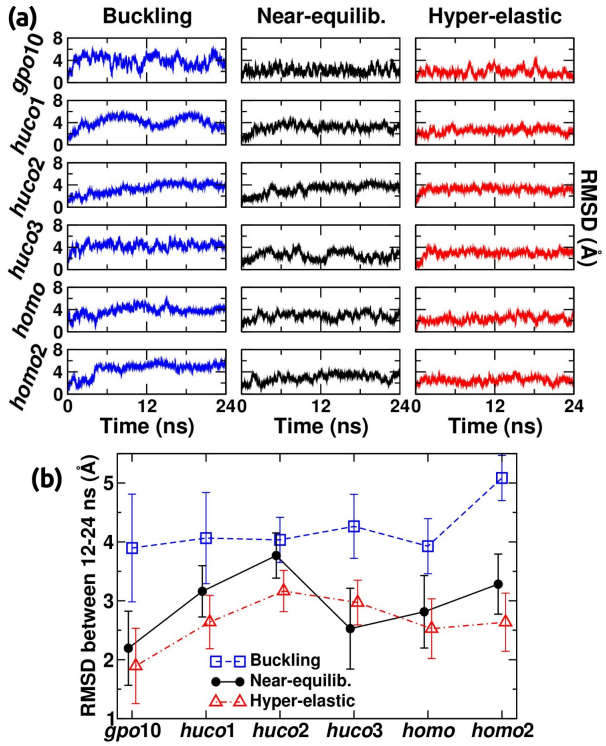

Figure 2.5: Average root-mean-square deviation (RMSD) of backbone heavy atoms from those at the beginning of the production run. The two GPO triplets at each end of the peptide were excluded from calculation. (a) Time trace. (b) RMSD averaged over 12–24 ns. Error bar: standard deviation. The same representative extensions of the three regimes (buckling, near-equilibrium, and hyper-elastic) in Fig. 2.7 are used.

too short (*e.g.*, between two immediately neighboring triads), κ_f may reflect properties of atomic-level covalent bonds rather than representing a local average for the peptide as a filament. On the other hand, if s is too long, fluctuations of all atoms within this interval will contribute to the measurement, so that the meaning of κ_f as a local property will be unclear. Due to the staggering of chains, MMP cleavage bonds (boldface in Table 2.1) occur over three triads. We thus used triad i and $i+3$ ($i = 1 \cdots 17$) for calculating s. For each pair of triads, we took s as an average over 12–24 ns and used it for calculating κ_f. Averaged over all triads in each peptide, s follows the same trend as the average end-to-end distance, which is the shortest for $gpo10$ (8.94 ± 0.02 Å) and the longest for $huco1$

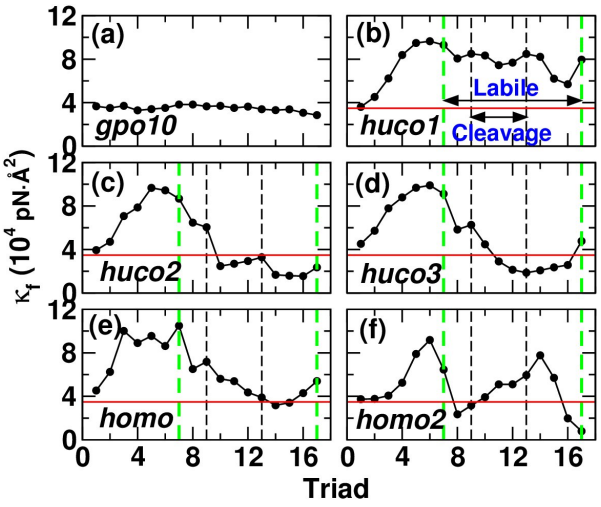

Figure 2.6: Local bending stiffness κ_f. (a) $gpo10$, (b) $huco1$, (c) $huco2$, (d) $huco3$, (e) $homo$, and (f) $homo2$ (see Table 4.1 for peptide names). Horizontal line (red) is the average κ_f for $gpo10$ (3.49×10^4 pN·Å^2), as a guide. While there are 20 triads (Table 2.1), since triads i and $i+3$ are used to calculate κ_f, the last data point ends at triad 17. Likewise, the MMP cleavage bond appears across triad 9–13. In triad 11, all three α chains contain the cleavage bond. The cleavage and the imino-poor labile (triad 7–17) domains are marked by vertical lines (noted in panel (b)).

(9.49 ± 0.16 Å).

For $gpo10$, κ_f is nearly constant (34900±2600 pN·Å^2; Fig. 4.3a). In other peptides, κ_f in the imino-rich domain (triad 1–7) is overall higher. This is because this region unwinds to make the three α chains rather parallel and suppresses bending motion (*cf.* Fig. 2.7). Among peptides other than $gpo10$, $huco2$ and $huco3$ have the two lowest κ_f in the cleavage domain (Fig. 4.3c,d). As discussed above, taking compliance of the cleavage domain as an attribute utilized by MMP, $huco2$ and $huco3$ may be better choices than $huco1$ with regard to bending. Combined with the results for the extensional stiffness, $huco2$ is mechanically the most compliant in both extension and bending, thus it may be the best candidate for MMP binding and cleavage. As explained below, this is due to the arrangement of residues in $huco2$ that destabilizes the labile domain and leads to unfolding (row *Free* in Fig. 2.7c).

When κ_f is calculated in 4-ns intervals, $gpo10$ shows no time dependence (Fig. 2.8).

In other peptides, κ_f varies over time to different degrees, reflecting their conformational motion. Yet, $huco2$ and $huco3$ are still more flexible in the cleavage domain than $huco1$ and $homo$. To compare our measurement with experiment, we calculated the persistence length $l_P = \frac{\kappa_f}{k_B T}$. It ranges between 84.2 nm ($gpo10$) to 181 nm ($huco1$), which lie well within the experimental estimates, 14.5–180 nm [194]. For our estimation, κ_f in each peptide was averaged over triads. For a full-length collagen, the apparent l_p may be dictated by highly flexible, locally unfolded regions such as the cleavage domain of $huco2$, which may be smaller. A recent study using atomic force microscopy reports 12–40 nm [118].

2.3.5 Torsional behavior

Twist of a triple helix is an important descriptor of collagen conformation [91, 163, 145], which may also be functionally important as it affects binding of MMP and cleavage of collagen [46, 29, 163, 123]. In simulations where the ends of only the leading chain are restrained, torsional angle decreased with extension, indicative of unwinding (Fig. 2.7a-f). Consistent with its stability, $gpo10$ unwound the least (Fig. 2.7a). In other peptides, the region around the MMP cleavage site underwent the greatest unwinding (darker color in Fig. 2.7b-f). In the buckling regime, kinking of $huco1$, $huco2$, and $huco3$ was observed at the cleavage site. These results further corroborate its labile nature. In simulations without any restraint, cleavage domains of $huco2$ and $huco3$ disrupt compared to that of $huco1$ (row *Free* in Fig. 2.7b–d). The extent of disruption is the greatest in $huco2$, which supports it as the most cleavable isomer. The imino-rich domain upstream to the cleavage site unwinds, likely due to Ala8 in $\alpha1$ and Thr11 in both $\alpha1$ and $\alpha2$ (Table 2.1). However, further unfolding of this region does not occur and the three α chains stay aligned, resulting in elevated bending stiffness (left of the cleavage site in Fig. 2.7b-f).

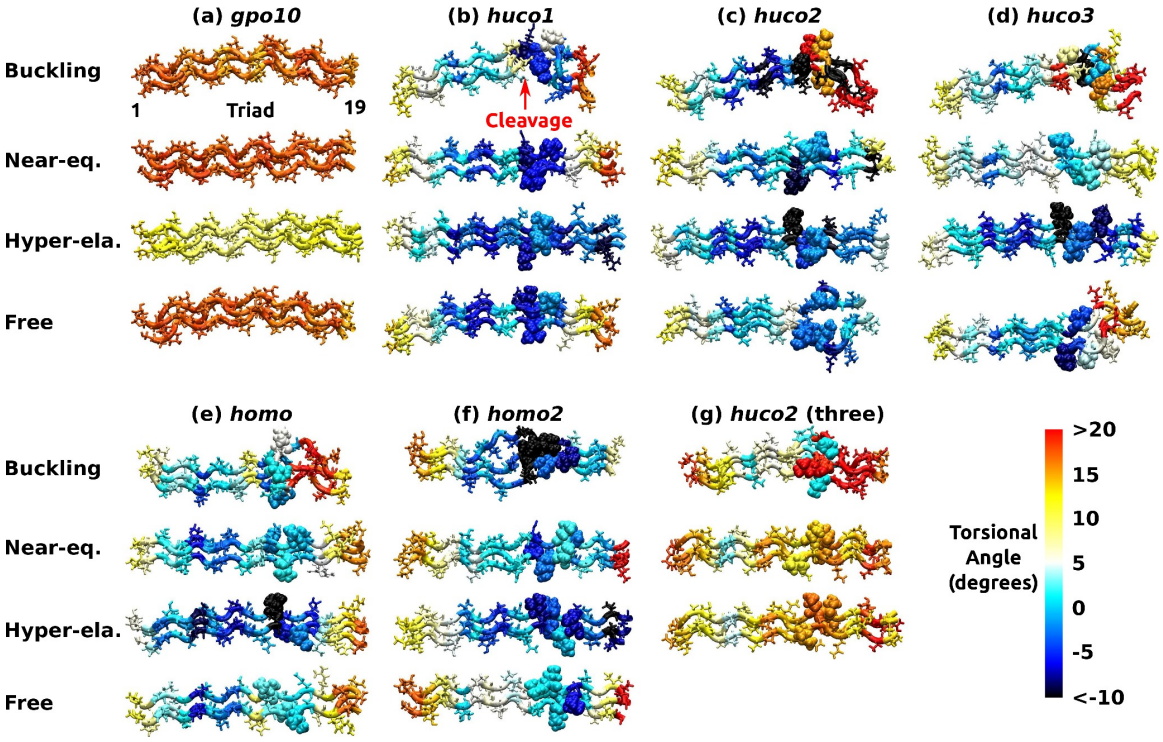

Figure 2.7: Torsional angles between successive triads averaged over 12–24-ns, displayed on conformations at the end of each simulation. Two GPO triplets at each end of the peptide are not shown. (a) $gpo10$, (b) $huco1$, (c) $huco2$, (d) $huco3$, (e) $homo$, (f) $homo2$, and (g) $huco2$ with ends of all three α chains restrained (see Table 4.1 for peptide names). *Buckling*, *Near-eq.*, and *Hyper-ela.*, are representative structures from respective regimes, where the extensions are 66.4, 70.4, and 76.0 Å for $gpo10$, 68.0, 73.6, and 77.6 Å for $huco2$, and 68.0, 74.4, and 77.6 Å for all other peptides. These are based on differences in L_{eq} (Fig. 2.2b). *Free* is for simulation without any restraint. The same extensional regimes are used in Figs. 2.9 and 2.10. For torsional angle measured between triad i and $i+1$, residues of triad i are colored (marked triad 1–19 in (a)). Free: simulation without any restraint. Ile17 in $\alpha1$ and Leu17 in $\alpha2$ at the cleavage site (Table 2.1) are shown in van der Waals representation to show their location (marked in (b)). Molecular structures are rendered using VMD [74].

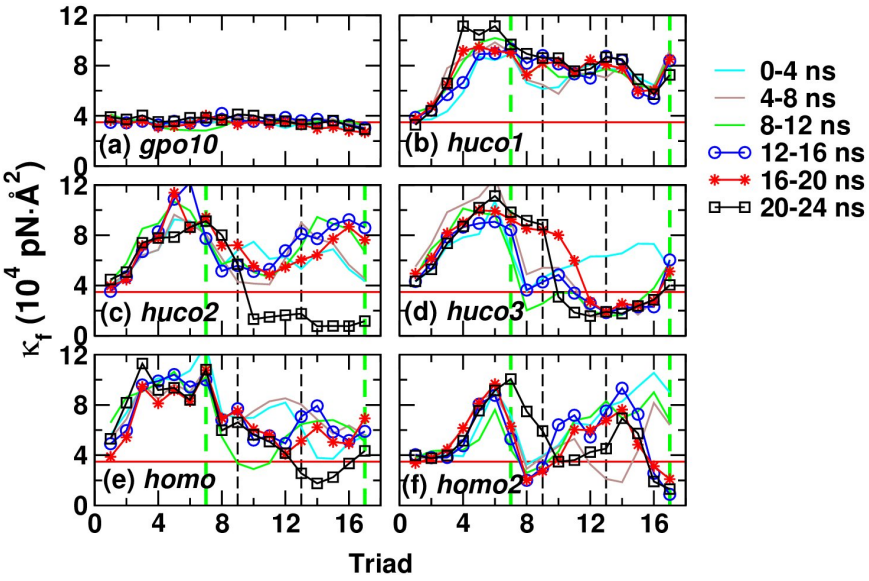

Figure 2.8: Local bending stiffness κ_f calculated in 4-ns intervals. Red horizontal line and vertical dashed lines are explained in Fig. 4.3. Among $huco1$, $huco2$, and $huco3$, the latter two have κ_f in the cleavage region (triad 9–13) consistently lower than that of $huco1$ after 12 ns. The large decrease in κ_f of $huco2$ in triads 9–17 during 20–24 ns is due to unfolding of this region that occurred at around 22.5 ns, as shown in Fig. 2.7c.

2.3.6 Dependence on loading condition

We restrained the ends of only one α chain when studying the extensional behavior, which allowed conformational (especially torsional) motion under load. To test the effect of disallowing torsional motion of the end, for $huco2$, we applied restraints to three C_α atoms of residue 4, 3, and 2, respectively from the leading, middle, and trailing chain, and likewise restrained residues 28, 27, and 26. In this case, the extensional stiffness was $k = 186.4 \pm 6.6$ pN/Å (Fig. 2.11), which is about 5 times greater than the case with only one α chain restrained. The corresponding Young's modulus, 8.75 ± 0.31 GPa is comparable to the maximum among previous estimates [63]. Furthermore, the triple helix unwinds far less, with much reduced dependence on extension (Fig. 2.7g). These results highlight the sensitivity of the conformational behavior on the loading condition.

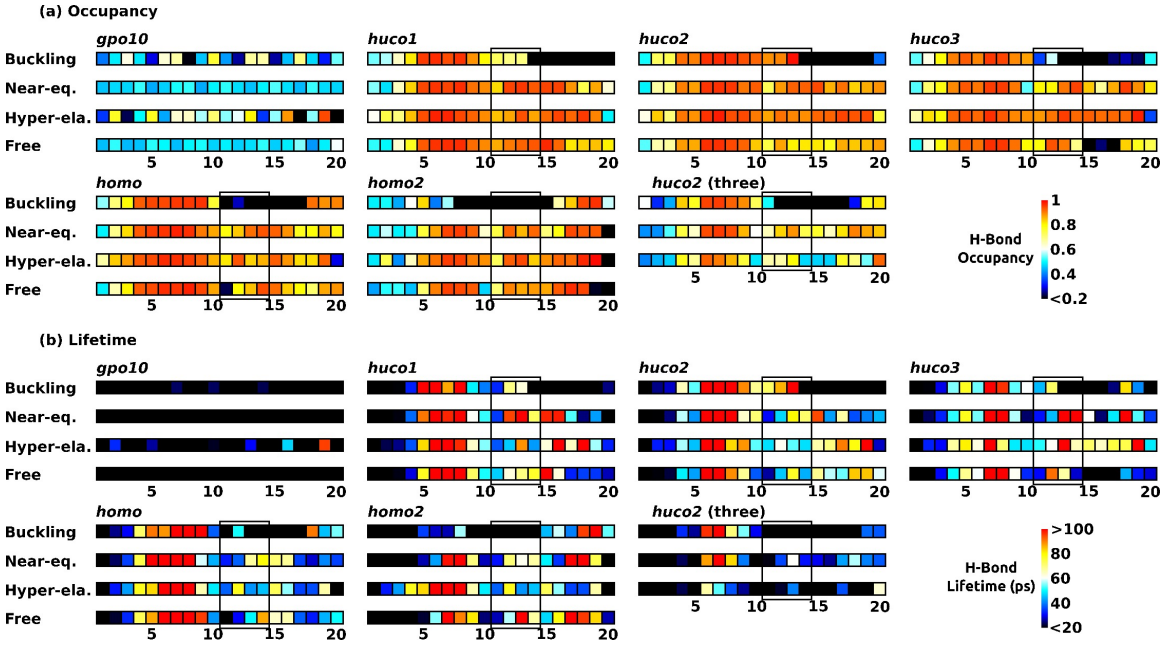

Figure 2.9: Dynamics of native backbone H-bonds (see Table 4.1 for peptide names, and Fig. 2.7 for *Buckling*, *Near-eq.*, *Hyper-ela.*, and *Free*). (a) Occupancy and (b) average lifetime. Standard deviations of lifetimes are in Fig. 2.12a. Measured values with glycines as H-bond donors are averaged for each triad and represented in color. Triads 11–14 contain cleavage sites (solid box). The imino-poor domain spans triad 10–20. Compared to other peptides, H-bonds of $gpo10$ have notably smaller occupancy, lifetime, and standard deviation (Fig. 2.12a), suggesting a dynamic stabilization mechanism.

Its implication for MMP cleavage is considered in Discussion.

2.3.7 Hydrogen bonding events

H-bonds are critical for the stability of the collagen triple helix [169, 221, 19]. We classified them into 'native' and 'non-native.' Native H-bonds are formed between backbone amide hydrogen of glycine in a GXY triplet to the backbone carbonyl oxygen of the residue at the X position of a neighboring α chain [19]. They are thus formed in a helical manner, between leading-middle, middle-trailing, and trailing-leading chains. Since atoms forming native H-bonds are present in any GXY sequence, native H-bonds can form in both imino-rich and imino-poor domains [91]. Non-native H-bonds refer to all others,

including those between backbone to backbone, backbone to side-chain, and side-chain to side-chain. For identification of a H-bond, a cutoff distance of 2.4 Å between hydrogen and oxygen atoms was used [115]. H-bonding events were quantified by occupancy (number of coordinate frames where a H-bond is formed, divided by the total number of frames), average lifetime (average duration of consecutive frames where a H-bond is formed), and standard deviation of the lifetime. The H-bond occupancy and lifetime can together provide a dynamic picture of the H-bonding events. For example, two bonds may have the same occupancy but differ in lifetimes, as one bond may rapidly form and break, while the other may be longer-lived but forms more sparsely. The converse may also hold, with similar lifetimes but different occupancies depending on the frequency of H-bond formation.

Strikingly, the stable $gpo10$ has the lowest native H-bond occupancy compared to those of other peptides (Fig. 2.9a). Its average native H-bond lifetime and fluctuation (standard deviation) are also the shortest (Figs. 2.9b and 2.12a). This suggests that the native H-bonds of $gpo10$ stabilize the structure dynamically, by rapid formation and breakage in a uniform manner. The native H-bond occupancy of $gpo10$ becomes uneven along its length in the buckling and hyper-elastic regimes as strain builds up in the structure. In the hyper-elastic regime, the native H-bond occupancy overall increases, which is also observed in other peptides (Fig. 2.9a). An exception is $huco2$ with all three α chains restrained ('$huco2$ (three)' in Fig. 2.9a), where the native H-bond occupancy is lower in the hyper-elastic compared to the near-equilibrium regime. In this case, the peptide becomes more wound with extension (Fig. 2.7g), becoming conformationally closer to $gpo10$ whose native H-bond occupancy is low. These results indicate that unwinding of the triple helix actually promotes native H-bond formation. Consistent with this, triads 5–10 that are upstream to the MMP cleavage site, have elevated occupancy and longer lifetime (Fig. 2.9). As explained earlier, this region unwinds without α chains falling apart (Fig. 2.7). The higher

occupancy of native H-bonds in this region likely contributes to its larger bending stiffness (Fig. 4.3).

Non-native H-bonds show more punctate behavior (Fig. 2.10). The well-folded $gpo10$ has very few non-native H-bonds. This is also the case in other peptides upstream to the MMP cleavage domain (triads 5–10) that are unwound without falling apart. Non-native H-bonds occur downstream to the cleavage site (triad 15–20) as this region is more disrupted (Fig. 2.7). In particular, triads 17–18 of $homo$ have high-occupancy non-native H-bonds in all extensional regimes and also in the restraint-free case (Fig. 2.10). They mainly involve H-bond between Arg21 of $\alpha 1$ (Table 2.1) and the backbone oxygen atom in a neighboring chain (Fig. 2.13a). We call it the *Arg-bridge*. Although several other very short-lived non-native H-bonds in these triads caused the average lifetime below 50 ps, the Arg-bridge can persist beyond 100 ps, so it can play a substantial role in local stabilization.

2.3.8 Molecular origin of the dependence on chain registry

In addition to the Arg-bridge, we found that two bulky non-polar residues Leu17 and Leu18 in $\alpha 2$, located right next to the cleavage bond (Table 2.1), play a critical role in determining registry-dependent conformational behavior. In $huco1$, since $\alpha 2$ is in the leading position, Arg21 of $\alpha 1$, being farther downstream, can form a bridge while Leu17 and Leu18 interact with surrounding residues (Fig. 2.13b). In $huco2$, placement of $\alpha 2$ in the middle separates Arg21 in two $\alpha 1$, resulting in the greatest destabilization (Fig. 2.13c). In $huco3$, since the two leucines of $\alpha 2$ are close to Arg21, their hydrophobic stabilization requires local deformation of the molecule and interferes with Arg-bridge formation, which again have a destabilizing effect, but to a less extent compared to $huco2$ (Fig. 2.13d).

To test the stabilizing role of the Arg-bridge, we constructed models of three mutant peptides, $huco1_m$, $homo_m$ and $homo2_m$ where Hyp21 and Arg21 in respective chains are switched (Table 4.1). For each peptide, we carried out 24-ns MD simulation without any

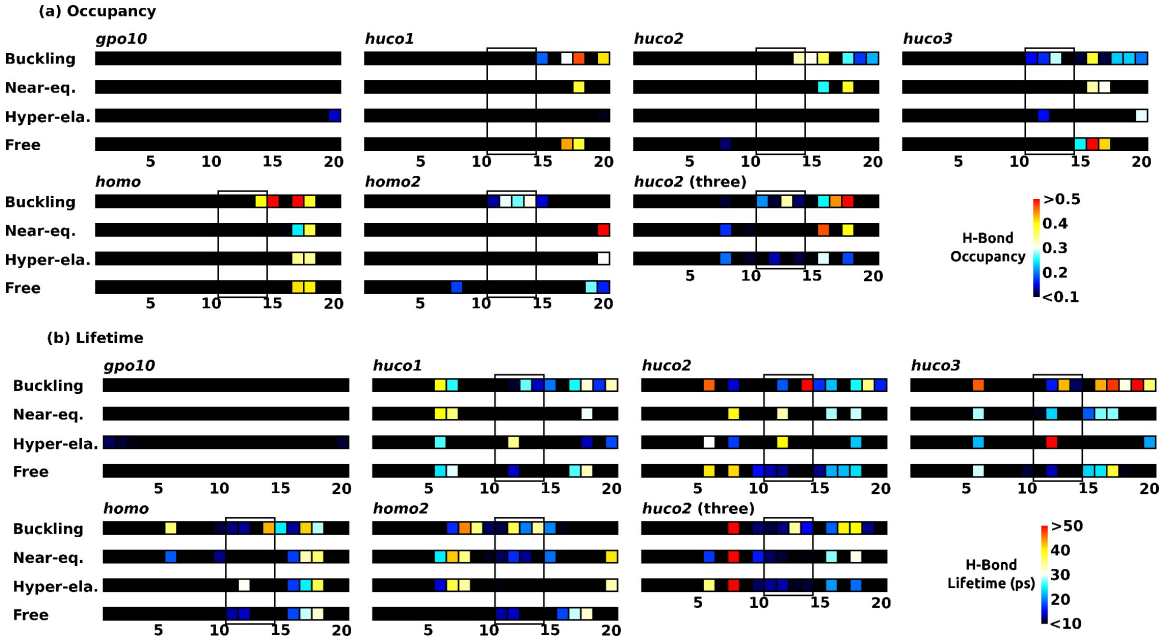

Figure 2.10: Dynamics of non-native H-bonds (see Table 4.1 for peptide names, and Fig. 2.7 for *Buckling*, *Near-eq.*, *Hyper-ela.*, and *Free*). (a) Occupancy and (b) average lifetime. Standard deviations of lifetimes are in Fig. 2.12b. For each triad, H-bonds are counted only when residues in the triad serve as H-bond donor, to avoid double counting across different triads. Triads 11–14 contain cleavage sites (solid box). High-occupancy bonds in triads 17–18 in $homo$ (also in $huco1$) are due to Arg-bridges (Fig. 2.13).

restraint applied. Whereas triads 16—19 in $huco1$ and $homo$ remained wound (Fig. 2.7b,e; 'Free'), this region in $huco1_m$ and $homo_m$ unwound, with very low occupancy of non-native H-bond (Fig. 2.14a,b). $homo2_m$ behaves oppositely, where triads 16–19 are wound more compared to $homo2$, and have high-occupancy non-native H-bonds (Fig. 2.14c), mainly Arg-bridges. These results corroborate the importance of the Arg-bridge, which may contribute to the stability and cleavage resistance of the type-I collagen homotrimer [60, 122].

2.4 Discussion

Present results elucidate mechanical and conformational differences between homo- *vs.* heterotrimers of collagen, or between isomers with different registry of α chains. Al-

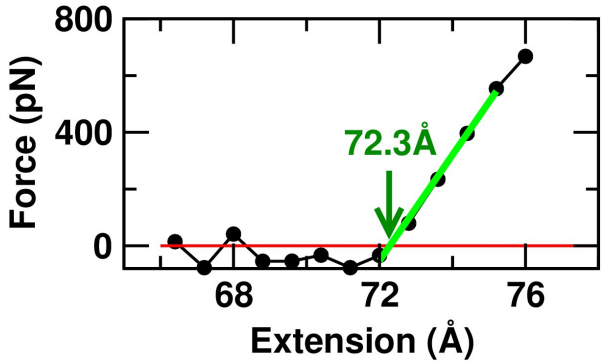

Figure 2.11: Force-extension relation of $huco2$ with ends of all three chains constrained. Lines and symbols are defined in Fig. 2.2b. The stiffness, 187 pN/Å (slope of the thick green line), is much higher than the case of $huco2$ with only one α chain restrained (37.1 pN/Å).

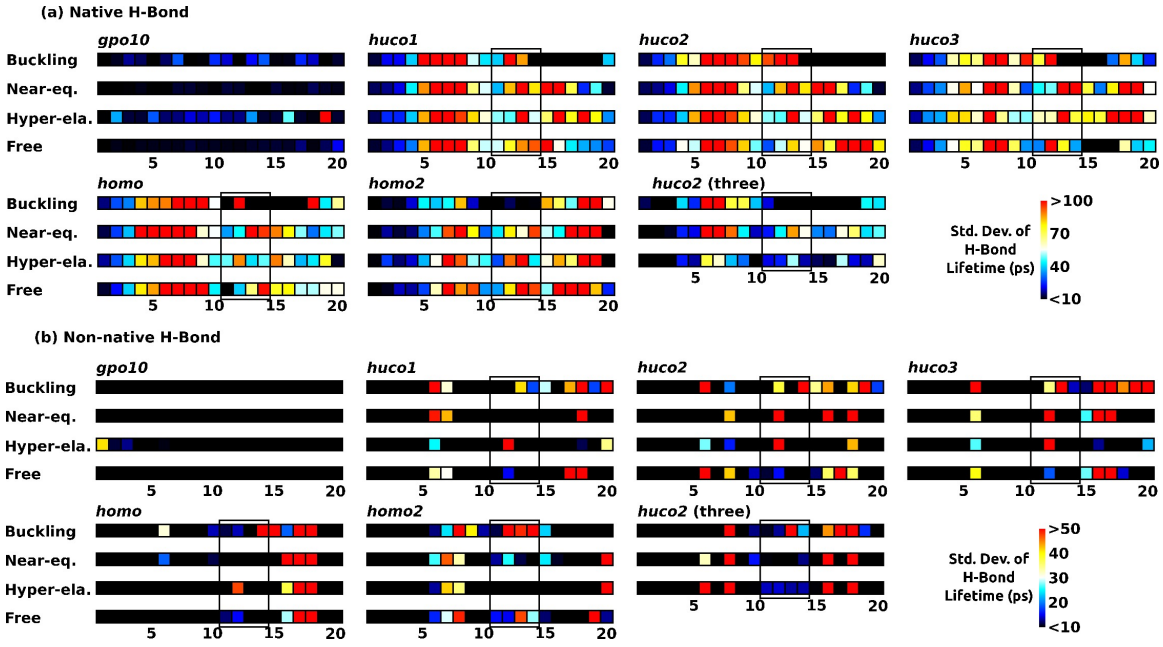

Figure 2.12: Standard deviation in the hydrogen bond lifetime. (a) Native and (b) non-native. *cf.*, Figs. 2.9 and 2.10.

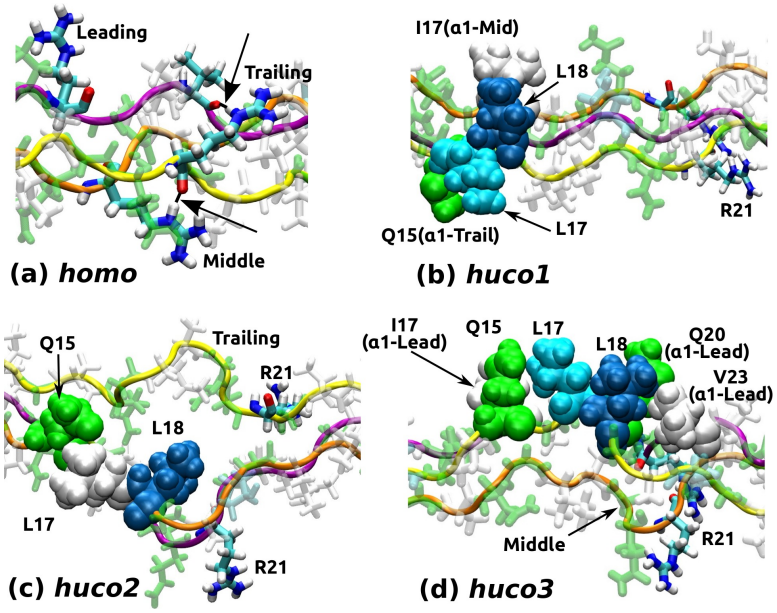

Figure 2.13: Role of the Arg-bridge and chain registry on the conformation of the imino-poor domain (see Table 4.1 for peptide names). Structures are taken after 24-ns MD without any restraint. (a) $homo$. Arg-bridges are marked by arrows. Arg21 in the leading chain does not form a bridge. (b) $huco1$. Leu17 and Leu18 of the leading $\alpha 2$ are held by Ile17 in middle and Gln15 in trailing chains. (c) $huco2$. Leu18 inserts between α chains and the trailing chain separates. Arg-bridges are absent. (d) $huco3$. Leu17 and Leu18 of the trailing $\alpha 2$ are held by residues in the leading chain and by Gln15 of $\alpha 2$. The longitudinal compaction causes the middle chain to bend severely. (a) is rendered larger than (b-d).

though our calculation shows that $huco1$ and $huco2$ (see Table 4.1 for peptide names) have the two lowest extensional stiffness, variation in extensional stiffness among peptides tested (37–49 pN/Å) is well within 2-fold (Fig. 2.2c). By comparison, the local bending stiffness κ_f varies by as much as 5-fold (Fig. 4.3). It is thus a more sensitive measure of local conformational properties. For $gpo10$ that is uniform in flexibility, we can calculate Young's modulus using bending stiffness, $E = \kappa_f / I$, where $I = \frac{\pi}{4} r^4$ ($r = 7$ Å) is the second moment of inertia of cross section for a circular cylinder of radius r [70]. Using the average κ_f for $gpo10$, 3.49×10^4 pN·Å^2, we get $E = 1.85$ GPa, which is comparable to the value based on extensional stiffness (Fig. 2.2c). This reflects consistency of our

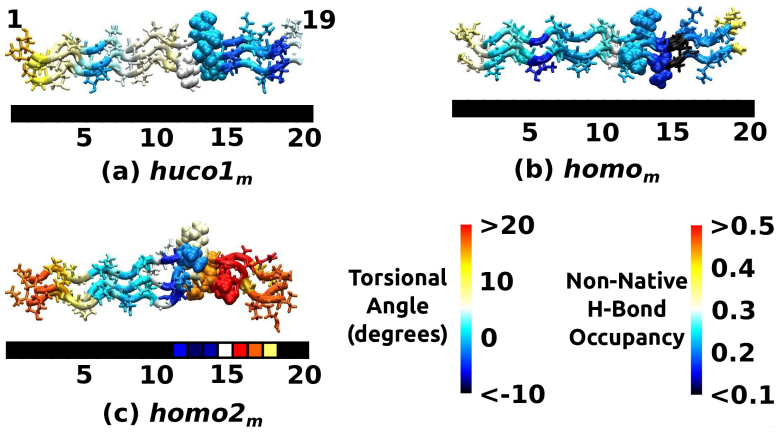

Figure 2.14: Conformational behavior of mutant peptides in Table 4.1. (a) $huco1_m$, (b) $homo_m$, (c) $homo2_m$. Structures are taken after 24 ns MD without any restraint. Coloring schemes are the same as in Fig. 2.7b,e,f (torsional map) and Fig. 2.10a (non-native H-bond occupancy). Since the molecular structure is 3-dimensional, its triads do not align exactly with the triad numbers of the color strip below. Without the Arg-bridge, triads 16–19 of $huco1_m$ and $homo_m$ undergo unwinding. Conversely, $homo2_m$ stays wound due to the presence of the Arg-bridge that manifests as a high-occupancy non-native H-bond.

measurements in simulations with and without restraints.

Among the three extensional regimes, the near-equilibrium regime is likely the most physiological. In tissues, collagen bundles form macroscopic crimps [51] so that molecular-level buckling is unlikely to happen under compression. On the extensional side, we can estimate a typical tensile load by considering tendon. The cross sectional area of a human tendon is on the order of cm^2, and it bears $\sim$kN forces. Assuming that the entire cross section of a tendon consists of collagen molecules 7-Å in radius, there are about 6.5×10^{13} collagen molecules, so that each molecule will bear about 15 pN. Thus, up to 100 pN in Fig. 2.2b will be physiological, which lies within the near-equilibrium regime.

In addition to force, we examine the relevant time scale. Lifetimes of contacts are at most a few hundred picoseconds, majority of which being less than 100 ps (Figs. 2.9 and 2.10). Thus, the 12-ns measurement time during the later half of 24-ns simulation was sufficient for monitoring the dynamics of contacts, We also observed transient forma-

tion of β-sheets, consistent with experiment [57]. They are mostly short, formed by two backbone H-bonds between two α chains in parallel, and are rarer than individual contacts. Additional analysis would be necessary to elucidate the role of transient β-sheets in conformational dynamics of the molecule.

Even though individual bonds form and break rapidly, the overall conformational fluctuation of the peptide may be slower. The RMSD undulates on the order of a few nanoseconds (Fig. 2.5). We estimate the slowest relaxation time of the peptide as an elastic rod suspended in a viscous medium [70]. For a rod of length L, diameter d, and bending stiffness κ_f, its slowest relaxation time τ in a solution of viscosity η is given by $\tau = \frac{c_\perp}{\kappa_f}(\frac{L}{\omega_1})^4$, where $c_\perp = 4\pi\eta/[\ln(L/d) + 0.84]$ is the transverse drag coefficient per unit length of the rod, and ω_1 is a constant of order 1 for the slowest vibrational mode, which depends on the boundary condition of motion. For $gpo10$, we have $L = 87.5$ Å, $d = 14$ Å, and $\kappa_f = 3.49 \times 10^4$ pN·Å^2 (Fig. 4.3). This gives $\tau \simeq 140$–330 ps, which is comparable to the longest H-bond lifetimes. This shows that the 24-ns simulation time was much longer than the equilibrium fluctuation time of the peptide. However, large deviations from the triple helical conformation can occur over a longer time scale, such as formation and breakage of β-sheets mentioned above, and water molecules can break in and out between α chains in the locally unfolded labile domain. A more detailed analysis of such events would require at least an order of magnitude longer simulation, which would be impractical. Nevertheless, the 24-ns simulation time employed in the present study was sufficient for distinguishing between the relative stability and region-specific conformational behavior of the triple helical peptides, which is further supported by the agreement of our calculations with available experimental data. On the other hand, although biasing potentials *e.g.*, in umbrella sampling, may further drive conformational changes [138], unless reaction coordinates are properly chosen, it is difficult to interpret the observed changes [6].

A fundamental aspect that we revealed is the dynamic stabilization mechanism for the

GPO repeat: $gpo10$ has low-occupancy, short-lived native backbone H-bonds (Fig. 2.9), which is an effective strategy to maintain the triple helical structure while remaining flexible (Fig. 4.3a). This is reminiscent of the stabilization of single α-helical domains by dynamic and 'malleable' contacts between appropriately located charged side chains that hold the 'brittle' α-helix backbone [195, 196, 199]. Analogously, in the case of the collagen GPO domain, dynamic backbone H-bonds hold the triple helix tertiary structure.

For the MMP cleavage domain, there are stronger regional variations in flexibility and stability. We found that $huco2$ is the isomer that is likely the most cleavage-prone, as it has the highest mechanical compliance and the greatest unfolding of the region at and downstream the cleavage site. Instability of $huco2$ in the labile domain is due to the axial separation of Arg21 in the leading and trailing $\alpha1$, together with the two tandem leucines of $\alpha2$ that locally destabilize the region and hampers the Arg-bridge formation. Although configurations in Fig. 2.13 may not be the only way how these residues organize locally, they illustrate the unfavorable arrangement of arginines and leucines in $huco2$ compared to other isomers. The stabilizing role of the Arg-bridge has been shown experimentally in a model heterotrimer [175], and similar roles for lysine and glutamic acid were also suggested for type-IV collagen [178]. However, as Fig. 2.13 shows, placement of these residues within the molecule affects the extent of stabilization.

The high stiffness of the imino-rich domain N-terminal to the cleavage site (Fig. 4.3) is a result of unwinding without separation of α chains (Fig. 2.7) that appears to promote native backbone H-bond formation (Fig. 2.9). While there are many sites along the collagen molecule whose amino acid sequences are partially similar to the actual MMP cleavage site, the latter is distinguished by local imino-rich GXY repeats followed by an imino-poor domain [46, 222]. This suggests that the abrupt transition in local bending stiffness may be unique to the MMP cleavage site, thus it may provide a mechanical recognition signal as MMP diffuses over collagen and searches for the cleavage site [183].

The present analysis also makes a testable prediction: $huco1$, possessing the Arg-bridge (Figs. 2.10, 2.13b) behaved similarly to $homo$. While the presence of $\alpha 2$ may make $huco1$ not as cleavage-resistant as $homo$, compared to $huco2$ or $huco3$, it is likely to be so and also be more stable, which will be an interesting subject for experiments. The possible stabilizing role of arginine in type-III collagen has been previously proposed [179], although the atomistic basis was unclear. Our analysis shows that Arg-bridges are dynamic, whose strength depends on their location relative to other residues (Fig. 2.13).

Last, we discuss the conflicting reports of the cleavage rate of single collagen molecules, that either increased [1, 2] or decreased [24] with load in similar magnetic tweezer experiments. In the former case, a homotrimeric peptide containing the MMP cleavage domain [1] (similar to $homo$) or a full-length type-I collagen heterotrimer [2] (corresponding to $huco2$) was linked between a glass coverslip and a magnetic bead. Since the bead can rotate, unwinding of the molecule is possible regardless of the number of α chains in a molecule linked to the bead or to the coverslip. As in our simulation (Fig. 2.7), stretching will result in more unwinding, which may assist with cleavage by MMP. In comparing between homotrimer and heterotrimer, cleavage of the former was more sensitive to load, which was interpreted to be due to its higher propensity to unwind under load while the heterotrimer is unwound even without load [2]. Our simulation supports this, as $homo$ unwound substantially in the hyper-elastic regime compared to near-equilibrium or load-free case (Fig. 2.7e), whereas $huco2$ is already unwound (near-equilibrium) or unfolded (load-free) (Fig. 2.7c). It should be noted that unwinding can either stabilize or destabilize the triple helix, depending on whether the domain is imino-rich or imino-poor, as seen in our analysis of bending stiffness and H-bonds.

In another study, cleavage rate of collagen decreased by nearly 10-fold with load [24]. In this case, antibody-functionalized beads were exposed to a large volume of type-I collagens to achieve conjugation. This may result in multiple collagen molecules attaching to

a single bead. Even though a single collagen tether may be formed between the bead and the surface, neighboring collagen molecules can bind to the tethered collagen, affecting its conformational motion. Likewise, in tissues, other neighboring molecules in a bundle may limit conformational motion of the cleavage domain under load, thereby protect it from cleavage. While presence of many other factors makes it difficult to directly apply analysis of a single triple helix to a tissue, our study demonstrates that susceptibility to MMP cleavage depends sensitively on the loading condition and the local arrangement of molecules, and not simply on the magnitude of load. Experimentally, when studying load-dependent cleavage of collagen by MMP, it would thus be necessary to probe or control the mechanical environment around collagen molecules in limiting or promoting local unfolding.

2.5 Conclusions

Fibrillar collagen is the major load-bearing component of the tissue, so that continuum mechanical description of a collagen molecule as a biopolymer is needed. On the other hand, its biological function involves residue-specific behaviors as in any globular proteins. Our study elucidates the atomistic origin for the mechanical and conformational properties of the MMP cleavage domain of type-I collagen. Fundamental aspects that we found, such as the local conformational behavior of the triple helix under load, static *versus* dynamic origin for the flexibility and stability, and the effect of chain registry, will also be useful for understanding the behaviors of other domains or other types of fibrillar collagens.

3. ELASTIC ENERGY PARTITIONING IN DNA DEFORMATION AND BINDING TO PROTEINS*

3.1 Introduction

Sequence-dependent flexibility of DNA is essential for genomic organization, DNA packaging, and positioning of DNA-binding enzymes [65, 137, 213, 182, 53, 17]. Deformability of possible combinations of base pairs have been extensively analyzed, through surveys of available structures [41, 147, 148, 213, 28] and using various experimental methods that measure bending [125, 144, 187, 225, 53] and twist [219, 66, 69, 114, 157] motion. Yet, the issue remains unclear due to several reasons [59, 42, 188, 53, 214, 103, 39]. X-ray structures are static and DNA conformations may be affected by crystal contacts, where discrepancy between crystal analysis and cyclization-based flexibility measurements have been observed [53]. However, experiments based on the dynamics of DNA oligos do not address anisotropy in bending and results are interpretive in nature. Also, single-molecule experiments report higher flexibility of DNAs than in measurements based on bulk cyclization experiments [214, 103, 23]. Molecular dynamics (MD) simulation thus plays an increasingly important role [22, 100, 99, 128, 150, 43, 197, 12, 156]. With the development of force fields for nucleic acids [150, 64] and increasing computing power, MD simulations can address not only equilibrium conformational behaviors [102, 12, 33, 197, 39] but also conformational transitions of DNA [5, 143, 154].

For analyzing DNA conformation, helicoidal parameters have been predominantly used [37, 41, 119]. Their fluctuation in x-ray structures or simulations have also been used to study elastic properties of DNA [147, 100, 99, 102]. However, while helicoidal

*Reprinted with permission from "Elastic Energy Partition in DNA Deformation and Binding to Proteins" by Xiaojing Teng and Wonmuk Hwang, 2016. *ACS nano*, 10, 170-180, Copyright [2016] by American Chemical Society.

parameters effectively describe DNA structure at near-atomic level, it is unclear whether they are suitable for the motion of DNA as an elastic rod, which is a continuum mechanical concept [97]. Previous simulations found coupling between helicoidal parameters, and non-Gaussian or bimodal behaviors of some of the helicoidal parameters [147, 152, 101, 170, 7, 33, 154], so that the corresponding stiffness matrix or its diagonalized form cannot faithfully reflect elastic properties of DNA. By analogy, a protein undergoing hinge motion is better described by parameters associated with its hinge angle rather than by a set of parameters related to internal coordinates. In the case of an elastic rod, basically four order parameters are needed: Bending about two principal axes, twist (not a helicoidal parameter) about an axis orthogonal to principal axes, and extension [97, 30, 55]. The relation between the $\sim$16 helicoidal parameters [119] and these order parameters have not been evaluated carefully. In another study, *ab initio* optimization of base pair stacking energy was better described by non-standard sets of conformational variables [132], further suggesting that helicoidal parameters may not be ideal for describing energetics. This leads to a question regarding the applicability of an elastic rod description to the atomistic behaviors of DNA. Another related question is the minimum number of base pair steps that need to be considered collectively for calculating local order parameters along DNA.

Here we develop an approach where the order parameters are calculated based on MD simulation trajectories, and show that calculation at the level of dinucleotide yields results consistent with available data on DNA flexibility. About the two principal axes of bending, a dinucleotide step is the most flexible in the *major* bending direction and the stiffest in the *minor* bending direction. Principal axes (major and minor) are orthogonal by construction and there is no coupling between the two in the associated elastic energy [97, 30, 55]. They also yield the local equilibrium curvature. While the two bending motions show linear elastic behavior, twisting motion is slightly skewed, so that unwinding of the double helix is easier than overwinding. The distributions of bending and twist angles are consistent

with the twist-stretch coupling of DNA observed in single-molecule experiments [54, 113, 190]. While we mainly focus on B-DNA, an oligo with only cytosine on one strand and guanine on the other, turns into an A-DNA [49, 5, 132], which is about 1.5 times stiffer than that of the CC/GG step in the B-DNA form. For B-DNA steps, the stiffness associated with bending and twisting are within experimentally measured ranges.

Using the calculated stiffness, we build the flexibility maps of DNA oligos that are cleavable and non-cleavable by type-II topoisomerase [106], which agree with the maps calculated directly from MD simulations of these oligos. Sequence-dependent persistence lengths of oligos in our calculation also capture the experimental trend [53]. The observation that the calculated values are overall smaller is consistent with single-molecule results that found DNA to be more flexible than observed in bulk cyclization-based experiments [103, 214, 23]. The calculated stiffness and equilibrium angles allow to build a coarse-grained (C-G) model of DNA, which efficiently captures its length- and sequence-dependent conformational behavior. Furthermore, analysis of the Protein Data Bank (PDB) structures of protein-DNA complexes shows that in most cases DNAs are only mildly deformed, where the twist energy takes up the highest portion of the total elastic energy per dinucleotide step. On the other hand, in DNAs with high elastic energy, bending in the major direction is the dominant deformational mode. The present results thus elucidate partitioning of different deformational modes of DNA in its conformational motion and interaction with proteins, and should be useful when considering mesoscale organization of DNA assemblies as well.

3.2 Methods

3.2.1 DNA oligo generation

We used 7 oligos to cover the 10 possible dinucleotide steps (AA/TT, AC/GT, AG/CT, AT/AT, CA/TG, CC/GG, CG/CG, GA/TC, GC/GC, and TA/TA; Table 4.1). Steps con-

Sequence	Name	Number of dinucleotide steps used for analysis									
		AA/TT	AC/GT	AG/CT	AT/AT	CC/GG	CA/TG	CG/CG	GA/TC	GC/GC	TA/TA
d(CCCC**CCCCCCCC**CCCC)· d(GGGG**GGGGGGGG**GGGG)	[CCCC*]	–	–	–	–	4	–	–	–	–	–
d(CGCG**CCG**C**CG**C**C**GCGC)· d(GCGC**GGC**G**GC**G**G**CGCG)	[CCCC]	–	–	–	–	3	–	–	–	–	–
d(CGCG**ATATATAT**CGCG)· d(CGCG**ATATATAT**CGCG)	[ATAT]	–	–	–	4	–	–	–	–	–	3
d(CGCG**CGCGCGCG**CGCG)· d(CGCG**CGCGCGCG**CGCG)	[CGCG]	–	–	–	–	–	–	4	–	3	–
d(CGCG**AAAAAAAA**CGCG)· d(CGCG**TTTTTTTT**CGCG)	[AAAA]	4	–	–	–	–	–	–	–	–	–
d(AGAG**AGAGAGAG**AGAG)· d(CTCT**CTCTCTCT**CTCT)	[AGAG]	–	–	4	–	–	–	–	3	–	–
d(ACAC**ACACACAC**ACAC)· d(GTGT**GTGTGTGT**GTGT)	[ACAC]	–	4	–	–	–	3	–	–	–	–

Table 3.1: Names and sequences of DNA oligos tested. Base pairs used for analysis are in boldface in the sequence column. The first and last 4 base pairs were excluded from analysis to avoid end effects. [AAAA] and [CCCC*] respectively contain 7 AA/TT and CC/GG steps, out of which we used 4 alternating steps for consistency with other steps. [CCCC*] turns into an A-DNA (Fig. 3.1c). Most analysis for CC/GG was done with [CCCC], which maintains the B-DNA form.

sisting of only A and T were less stable [181] so [ATAT] and [AAAA] had d(CGCG)$_2$ at both ends to prevent fraying [107, 143, 226]. For CC/GG, we used two different oligos. [CCCC*] that has only C on one strand (d(C·G)$_{16}$) turned into an A-DNA soon after the simulation started (Fig. 3.1c), which is consistent with previous studies [49, 5, 132]. However, in simulations of oligos containing an isolated CC/GG, the B-DNA form was maintained. In [CCCC], CC/GG are thus interrupted by CG/CG. We also generated two 30-bp oligos with mixed sequences that are respectively cleavable (CLV) and non-cleavable (NONCLV) by type-II topoisomerase [106] (Table 3.2). All oligos were built initially in the B-DNA form by using X3DNA [119]. Missing hydrogen atoms in the structure generated by X3DNA were added by using CHARMM [21].

Name	Sequence					
CLV	GCTCA	CTCAA	AGGCG	GTAAT	ACGGT	TATCC
NONCLV	CATCG	ATAAG	CTTTA	ATTAA	AGCTT	ATCGA

Table 3.2: Two oligo sequences used for the simulations in Fig. 3.16 [106]. Sequence for one strand in the DNA duplex is shown in each case. Four base pairs on each end were excluded from calculating stiffness.

3.2.2 MD simulation

For simulation, we used CHARMM version 40a1 [21] with the param36 all-atom force field [64]. All simulations were carried out in explicit water using the TIP3P model [87]. After a brief initial energy minimization (1600 steps), the DNA was solvated in a cubic water box of side length about 84 Å (16-bp oligos) or 135 Å (30-bp CLV or NONCLV). DNAs were at least 15 Å away from the boundary in all directions, which is larger than the 12 Å cutoff distance for non-bonded interactions. To neutralize the system, sodium ions were added, resulting in about 90 mM concentration. Using only monovalent ions is a standard protocol in all-atom MD simulation [33, 142, 156]. Divalent ions are avoided in part due to their low mobility and long residence time once bound to DNA [134, 156]. Persistence length generally decreases with ionic strength, which is more substantial for divalent ions [8, 23]. However, in the $\sim$90-mM range of monovalent ions that we used, l_p depends only weakly on concentration [8, 23]. Inclusion of divalent ions such as Mg^{2+} is expected to reduce the stiffness of DNA, but it will require significantly longer simulation time [134]. Also, experiments testing the effect of divalent ions used very low concentration of monovalent ions. In physiological media where monovalent ions are typically much higher in concentration, multivalent ions are expected to have comparatively less influence on DNA flexibility.

The solvated system was energy minimized again for 600 steps using the steepest de-

scent method and then 1000 steps using the adapted basis Newton-Raphson method. During energy minimization, heavy atoms of DNA were harmonically restrained with a spring constant of 2 kcal/(mol·Å^2). The system was heated from 0 K to 300 K in 30 ps and equilibrated for 70 ps. Heavy atoms of DNA were harmonically restrained with a spring constant of 1 kcal/(mol·Å^2) during heating, and 0.2 kcal/(mol·Å^2) during equilibration. Heating and equilibration were done using the constant pressure (1 atm) and temperature (CPT) method [45]. The 100-ns production run was at 300 K under constant volume (NVT), without any restraints. We used the SHAKE [176] algorithm to fix the length of the covalent bonds for hydrogen atoms, and used a 2 fs integration time step. Particle-mesh Ewald (PME) summation was used to account for long-range electrostatic interactions [44]. The Domain Decomposition (DOMDEC) module of CHARMM was used to achieve efficient parallelization of simulation, using 280 (16-bp oligos) or 640 (30-bp oligos) processors [76]. Coordinates were saved every 10 ps (10^4 coordinate frames for 100 ns). Analysis of the coordinate trajectory was done for the last 50 ns for stable oligos. In [ATAT], base pairs broke more extensively (Fig. 3.2). The 5000 frames selected for analysis included only those with intact base pairs [39]. Similarly, the last 50 ns was analyzed in simulations of CLV and NONCLV, except for using 45–95 ns in step 19 of CLV (AT/AT) and 35–85 ns in steps 5 (GA/TC) and 6 (AT/AT) of NONCLV (*cf.*, Table 3.2).

Calculation of standard helicoidal parameters, number of base pair duplexes, and assessing A- and B-DNA structures for a coordinate frame (Figs. 3.1, 3.2, 3.11, and 3.12, and Table 4.4) were done using X3DNA. Molecular structures in figures were rendered using VMD [74].

3.2.3 Triad construction

The triad $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ for a given base pair (local coordinate basis) was constructed by adopting the method we previously developed [162, 94, 210]. We first assigned $\boldsymbol{e_3}$

as the unit vector normal to the best-fit plane of the base pair. For an oligo, we chose $\boldsymbol{e_3}$ to point approximately from the 5' to 3' direction of a strand selected as a reference. Next we determined a unit vector $\tilde{\boldsymbol{e}}_{\boldsymbol{2}}$ that aligns with the line joining C8 of purine and C6 of pyrimidine. The direction of $\tilde{\boldsymbol{e}}_{\boldsymbol{2}}$ was chosen to point to the reference strand of the oligo. Note that $\tilde{\boldsymbol{e}}_{\boldsymbol{2}}$ is not exactly orthogonal to $\boldsymbol{e_3}$. Using these two unit vectors, we set $\boldsymbol{e_1} = \tilde{\boldsymbol{e}}_{\boldsymbol{2}} \times \boldsymbol{e_3}$. This fixes $\boldsymbol{e_2} = \boldsymbol{e_3} \times \boldsymbol{e_1}$, which is along the projection of C8–C6 on the best-fit plane of a base pair. The midpoint between C8 of purine and C6 of pyrimidine was used as the centroid for each triad. In a triad, $\boldsymbol{e_1}$ is approximately in the direction of the major groove of DNA.

3.2.4 Identifying local principal axes

Conformational motion of a dinucleotide step was analyzed using the corresponding triads, which we call $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ and $\{\boldsymbol{e'_1}, \boldsymbol{e'_2}, \boldsymbol{e'_3}\}$ for respective base pairs. The latter triad is on the 3' side along the reference DNA strand mentioned above. Express $\boldsymbol{e'_3}$ in the coordinate frame i of $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ as $\boldsymbol{u}^i$ ($i = 1, \cdots 5000$). It forms a set of points on a unit sphere spanned by $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ (Fig. 3.3a). Its centroid $\boldsymbol{p_t}$ can be found numerically by minimizing the sum S of geodesic distances between $\boldsymbol{p_t}$ and $\boldsymbol{u}^i$:

$$S = \sum_i \arccos(\boldsymbol{u^i} \cdot \boldsymbol{p_t}). \tag{3.1}$$

Next we find a great circle on the unit sphere that passes through $\boldsymbol{p_t}$ and whose mean square arc distance from $\{\boldsymbol{u^i}\}$ is a minimum. The unit vector normal to this circle, $\boldsymbol{p_M}$, is the major principal axis. We thus call the circle as the *major bending circle* (red solid in Fig. 3.3a). To find it, we consider a series of great circles passing through $\boldsymbol{p_t}$, for which we use a normal vector $\boldsymbol{n'}$ that is perpendicular to $\boldsymbol{p_t}$. The contour $\boldsymbol{c}$ of the circle can be

parametrized by the angle θ:

$$\boldsymbol{c} = \cos\theta\,\boldsymbol{p_t} + \sin\theta\,(\boldsymbol{n'} \times \boldsymbol{p_t}), \quad \theta \in [0, 2\pi) \tag{3.2}$$

The great circle distance d_i between $\boldsymbol{u}^i$ to this circle can be found by minimizing $\arccos(\boldsymbol{u}^i \cdot \boldsymbol{c})$ while varying θ, which is equivalent to maximizing $\boldsymbol{u}^i \cdot \boldsymbol{c}$:

$$\begin{aligned} d_i &= \arccos[\max(\boldsymbol{u}^i \cdot \boldsymbol{c})] \\ &= \arccos\left(\max[\boldsymbol{u}^i \cdot \{\cos\theta\,\boldsymbol{p_t} + \sin\theta\,(\boldsymbol{n'} \times \boldsymbol{p_t})\}]\right) \\ &= \arccos\sqrt{(\boldsymbol{u}^i \cdot \boldsymbol{p_t})^2 + (\boldsymbol{u}^i \cdot [\boldsymbol{n'} \times \boldsymbol{p_t}])^2} \end{aligned} \tag{3.3}$$

Hence the normal $\boldsymbol{n}$ of the major bending circle, *i.e.*, $\boldsymbol{p_M}$, can be determined by minimizing the sum

$$\sum_i d_i^2 = \sum_i \left[\arccos\sqrt{(\boldsymbol{u}^i \cdot \boldsymbol{p_t})^2 + (\boldsymbol{u}^i \cdot [\boldsymbol{n} \times \boldsymbol{p_t}])^2}\right]^2. \tag{3.4}$$

The direction of $\boldsymbol{p_M}$ was set by imposing $\boldsymbol{p_M} \cdot \boldsymbol{e_2} > 0$. Once $\boldsymbol{p_t}$ and $\boldsymbol{p_M}$ are found, we can determine the minor principal axis $\boldsymbol{p_m} = \boldsymbol{p_M} \times \boldsymbol{p_t}$ and the corresponding *minor bending circle* (black dashed in Fig. 3.3a). See Appendix A for MATLAB code.

3.2.5 Elastic stiffness calculation

We projected $\boldsymbol{u}^i$ onto the major and minor bending circles and measured the respective angles θ_M^i and θ_m^i relative to $\boldsymbol{p_t}$. The signs of θ_M^i and θ_m^i (in the $(-\pi, \pi)$ range) were set equal to those of the dot products, $\boldsymbol{u}^i \cdot \boldsymbol{p_m}$ and $\boldsymbol{u}^i \cdot \boldsymbol{p_M}$, respectively (Fig. 3.4). This choice makes positive values of θ_M^i and θ_m^i to correspond to $\boldsymbol{p_t}$ bending in the directions of $+\boldsymbol{e_1}$ and $+\boldsymbol{e_2}$. By the definition of $\boldsymbol{p_t}$, the bending angles have zero average. The bending stiffness in the major and minor directions were found by applying the equipartition theorem [167]. If we denote the average distance between the centroids of the two neighboring

triads by s (Table 4.2), the bending stiffness are

$$\kappa_M = \frac{k_B T s}{\text{var}(\theta_M^i)} \tag{3.5}$$

$$\kappa_m = \frac{k_B T s}{\text{var}(\theta_m^i)}. \tag{3.6}$$

Here, k_B is Boltzmann constant and $T = 300$ K is the temperature. Variances of angles were measured across dinucleotide steps of the same type within an oligo (Table. 4.2). Twist stiffness κ_t can be calculated in a similar manner. At frame i, we express $\{\boldsymbol{e_1'}, \boldsymbol{e_2'}, \boldsymbol{e_3'}\}$ relative to its neighbor triad $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$, as $\{\boldsymbol{u_1^i}, \boldsymbol{u_2^i}, \boldsymbol{u}^i\}$. We calculate twist as the Euler angle θ_t^i about $\boldsymbol{p_t}$ for the rotation between $\{\boldsymbol{u_1^i}, \boldsymbol{u_2^i}, \boldsymbol{u}^i\}$ and $\{\boldsymbol{p_m}, \boldsymbol{p}_M, \boldsymbol{p_t}\}$ [162]. Then

$$\kappa_t = \frac{k_B T s}{\text{var}(\theta_t^i)}. \tag{3.7}$$

One potential issue in defining twist as an Euler angle is that, since $\boldsymbol{p_t}$ and $\boldsymbol{u}^i$ are not aligned, the rotation is not on the plane spanned by $\boldsymbol{u_1^i}$ and $\boldsymbol{u_2^i}$. We thus tested another definition of twist, where $\boldsymbol{u}^i$ is rotated on the plane spanned by $\boldsymbol{u}^i$ and $\boldsymbol{p_t}$, to be aligned with the latter, then twist angle is measured for aligning $\{\boldsymbol{u_1^i}, \boldsymbol{u_2^i}\}$ to $\{\boldsymbol{p_m}, \boldsymbol{p}_M\}$. However, the resulting twist angle distribution was nearly identical to that based on the Euler angle. Extensional stiffness of a dinucleotide step was similarly calculated as $\kappa_E = k_B T/\text{var}(s)$. Note that this stiffness is for a single dinucleotide step and is different from the stiffness associated with stretching a DNA as in single-molecule experiments.

3.2.6 Calculation of persistence length

There are different ways of calculating persistence length, such as fitting the wormlike chain model to the fluctuation of the end-to-end distance [125] and extrapolating distance-dependent fluctuations of bending angles [142]. The former has difficulty in assigning local bending stiffness, while the latter shows oscillatory behavior due to the helical nature

Step	s	$\boldsymbol{p_M}$	$\boldsymbol{p_m}$	θ_t	$\sigma(s)$	$\sigma(\theta_M)$	$\sigma(\theta_m)$
AA/TT	3.4021	(-0.3578 0.9338 -0.0009)	(0.9308 0.3566 -0.0801)	14.22±4.62	0.32	6.69	2.98
AC/GT	3.6500	(-0.1948 0.9808 0.0044)	(0.9760 0.1943 -0.0985)	19.77±6.00	0.37	7.69	3.02
AG/CT	3.4941	(-0.3697 0.9291 0.0052)	(0.9230 0.3679 -0.1126)	10.58±4.87	0.30	5.77	2.80
AT/AT	3.7687	(-0.1753 0.9845 -0.0106)	(0.9745 0.1720 -0.1438)	18.28±7.48	0.45	7.71	3.10
CA/TG	3.4012	(-0.3755 0.9268 -0.0034)	(0.9240 0.3740 -0.0798)	15.05±5.49	0.39	7.22	3.43
CC/GG	3.4823	(-0.2660 0.9640 0.0032)	(0.9613 0.2655 -0.0738)	20.53±5.25	0.43	5.99	3.43
CG/CG	3.4389	(-0.2992 0.9542 0.0004)	(0.9502 0.2979 -0.0911)	19.60±5.31	0.37	6.13	3.44
GA/TC	3.3905	(-0.4140 0.9101 0.0187)	(0.9100 0.4143 -0.0156)	14.14±4.48	0.34	5.51	3.05
GC/GC	3.4859	(-0.2589 0.9659 -0.0014)	(0.9615 0.2576 -0.0953)	16.91±6.83	0.33	5.67	2.85
TA/TA	3.6104	(-0.2901 0.9570 0.0099)	(0.9546 0.2901 -0.0671)	18.77±7.67	0.58	10.19	4.48

Table 3.3: Equilibrium conformations of dinucleotide steps. s: average distance between centroids (Å); $\boldsymbol{p_M}$ and $\boldsymbol{p_m}$: Coordinates of the major and minor principal axes relative to the triad of the reference base pair in a dinucleotide step. θ_t: Twist angle (degrees). $\sigma(s)$: standard deviation of the distance between centroids (Å). $\sigma(\theta_M)$ and $\sigma(\theta_m)$: standard deviations in the major and minor bending angles (degrees). Since θ_M and θ_m are measured relative to the centroid $\boldsymbol{p_t}$ ($= \boldsymbol{p_m} \times \boldsymbol{p_M}$; Fig. 3.3a), their averages are zero.

of DNA [142]. In our case, since we have obtained bending stiffness associated with local principal axes, persistence length can be estimated directly by $l_p^{\text{step}} = 2[k_B T(\kappa_M^{-1}+\kappa_m^{-1})]^{-1}$, where κ_M and κ_m are major and minor bending stiffness [96]. Using this, the effective persistence length l_p^{eff} of an oligo containing different dinucleotide steps can be calculated as [53]

$$\frac{1}{l_p^{\text{eff}}} = \sum_{\text{step}} \nu_{\text{step}} \frac{1}{l_p^{\text{step}}} \tag{3.8}$$

where ν_{step} is the fraction of a given dinucleotide step type in the sequence.

3.2.7 Analysis of protein-DNA complexes

By using the Nucleic Acid Database, we identified 2318 structures of protein-DNA complexes and downloaded them from the PDB web site. Among them, structures with only a single DNA strand or non-standard nucleotides were excluded. After this screening, 1381 structures remained. To avoid end effects, one base pair from each end of the double

helix was excluded. Steps with non-standard base pairing were also skipped. In the end, 33360 steps were analyzed.

3.2.8 Coarse-grained simulation of DNA oligos

We used the equilibrium angles of triads (Table 4.2) and stiffness (Table 4.3) to construct C-G models of DNA oligos. Motion of each dinucleotide step is described by three angles, θ_M (major bending), θ_m (minor bending), and θ_t (twist). For simulation, thermal forces are applied directly to them. This is analogous to an approach of all-atom MD simulation where internal coordinates such as dihedral angles are used as dynamical variables rather than Cartesian coordinates of atoms [78]. Apart from its efficiency, using order parameters directly for equations of motion eliminates the need to control them indirectly by introducing additional C-G atoms and associated force fields. Time evolution of an angle $\theta \in \{\theta_M, \theta_m, \theta_t\}$ was done by the Brownian dynamics method [75]:

$$\zeta \frac{d\theta}{dt} = \xi(t) - \kappa_\theta \theta. \tag{3.9}$$

Here, θ is measured as deviation from the equilibrium angle, and κ_θ is the corresponding stiffness (Table 4.3). On the left hand side of Eq. 3.9, ζ is the friction coefficient for θ. The Langevin force $\xi(t)$ is a Gaussian white noise that satisfies the fluctuation-dissipation theorem [216]. For integration of Eq. 3.9, we use the stochastic velocity Verlet algorithm that enables using $\sim$10-fold larger integration step compared to conventional Brownian dynamics integration algorithms [177]. We ignored extension since in the absence of any large axial load, contribution of extension to the conformational motion of an oligo is expected to be minimal. Since the equilibrium angles and their stiffness do not specify dynamical properties, we adjusted ζ in Eq. 3.9 so that the fluctuation of the corresponding angle for a given dinucleotide step over time differs from its standard deviation for the atomistic MD (Table 4.2) by less than 1%. The integration step size was $\Delta t = 0.001$. Since

only the angles are evolved in time, their values at a given time step were used to build Cartesian coordinates of the centroids and two arms of a triad for each dinucleotide step. Center of mass translation and rotation of the oligo are irrelevant to its conformational motion. For visualization, we placed the oligo's center of mass at the coordinate origin and aligned it along the horizontal direction.

To measure l_p of a dinucleotide step (or its combination; Fig. 3.17), we used oligos with lengths ranging from 10 to 1000 base pairs. Each simulation was run for 10^7 steps. At every 2000 steps, the end-to-end distance R_{ee} of the oligo was measured. For the mean square $\langle R_{ee}^2 \rangle$, we used the last half of the simulation time, although averaging over shorter time intervals showed that the system relaxes to equilibrium much earlier during simulation. We used the equation for the wormlike chain model [125], $\langle R_{ee}^2 \rangle = 2l_p[L + l_p(\exp(-L/l_p) - 1)]$ (L: length of the oligo), to obtain the persistence length l_p. The same approach was used to measure l_p of 200-bp oligos in Fig. 3.15. The simulation code was written in C++. For a 1000-bp oligo (the longest we tested), a 10^7-step simulation took approximately 1 hour on a single Intel X5560 2.8 GHz processor.

3.3 Results and Discussion

3.3.1 Analysis of dinucleotide step motion

Oligos stayed mostly stable during 100 ns without any significant breakage of base pairs (Fig. 3.1). We used the last 50 ns for analysis except [ATAT] where base pairs broke more extensively (Figs. 3.1f and 3.2). For [ATAT] we thus selected 5000 frames where base pairs stayed intact [39]. Although we calculated elastic stiffness of the CC/GG step in the A-DNA state using [CCCC*] (Fig. 3.1c), most of our analysis for the CC/GG step was based on [CCCC] which maintained the B-DNA state.

Conformation of DNA was measured by assigning local triads $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ to each base pair [163, 94, 210] (Method). An example analysis of AG/CT is shown in Fig. 3.3

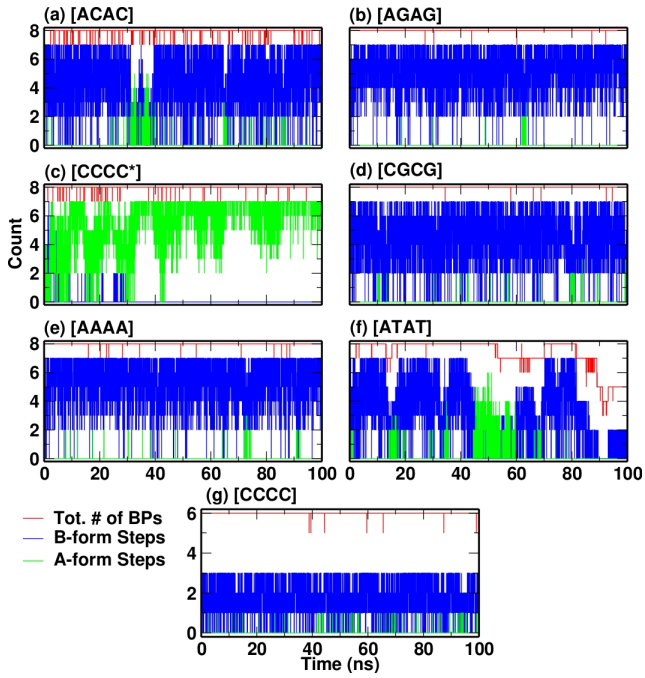

Figure 3.1: Stability and structure of DNA oligos during simulation. Base pairs marked in boldface in Table 4.1 were monitored. Red: total number of Watson-Crick base pairs, blue: number of B-form steps, green: number of A-form steps. In (a) to (f), the maximum number of base pairs is 8 so that up to 7 steps can form. In (g), a total of 6 base pairs and up to 3 CC/GG steps can form (Table 4.1). Except for (f), all oligos maintain the total number of base pairs close to the maximum value.

(see Figs. 3.5 and 3.6for other steps). The trajectory of $\boldsymbol{e_3}$ for G·C relative to the triad for A·T forms an ellipsoidal set of dots on a unit sphere (orange in Fig. 3.3a). From these the major and minor bending directions and the corresponding principal axes are found ($\boldsymbol{p}_M$ and $\boldsymbol{p}_m$ in Fig. 3.3a; Method). Two-dimensional (2D) histograms of bending angles along these directions can be approximated as products of two independent Gaussian distributions (Fig. 3.7). Thus, $\boldsymbol{p}_M$ and $\boldsymbol{p}_m$ indeed agree with the definition of principal axes where the elastic energy does not have a coupling term between the two angles [97]. Twist is measured about the centroid $\boldsymbol{p}_t$, whose direction also represents the local equilibrium curvature of the dinucleotide step (summarized in Table 4.2). The step is the most

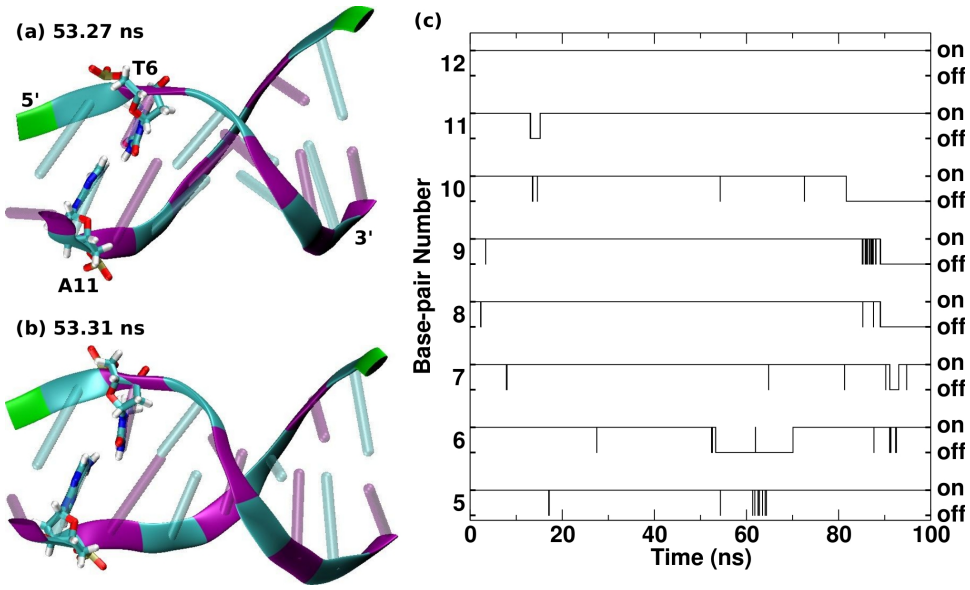

Figure 3.2: Formation and breakage of base pairs in [ATAT]. (a,b) Example breakage event. (a) The structure is distorted although all base pairs are intact. (b) The base pair slips normal to the plane of the pair and breaks. (c) Trajectory of base pairing. Panels (a) and (b) correspond to the base pair 6.

(least) likely to bend in the major (minor) directions (solid red and dashed black circles in Fig. 3.3a), which is the rotation about $\boldsymbol{p}_M$ ($\boldsymbol{p}_m$) (Fig. 3.4). The bending anisotropy is consistent with previous findings [227, 147, 100]. The major bending direction is approximately in the $\boldsymbol{e}_1$ direction facing the major groove (Figs. 3.3b and 3.6). Histograms of the projection angles of the dots on major or minor circles are fit well by unimodal Gaussian distributions (Fig. 3.8), suggesting that bending in the principal directions are linearly elastic. The distribution of twist is more skewed towards smaller angles [33], indicating that the double helix is easier to unwind than overwind (Fig. 3.8).

2D histograms of the bending *vs.* twist angles show no noticeable coupling except for CA/TG and TA/TA, where the steps untwist slightly as they bend in the direction of the major groove, as can be seen by the high-population region extending along the diagonal direction in Fig. 3.9c and Fig. 3.9j. On the other hand, twist mostly reduces as the step

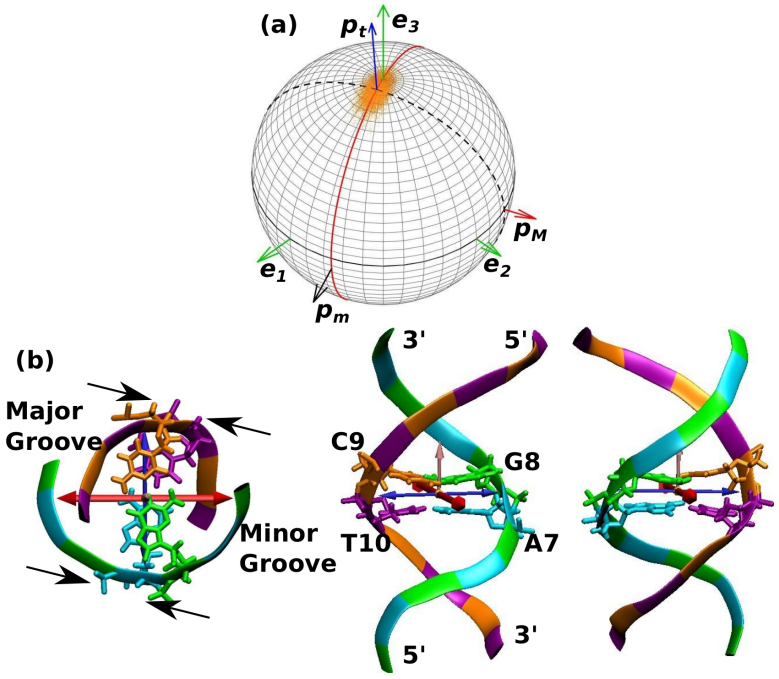

Figure 3.3: Conformational motion of the AG/CT step. Similar analysis for other steps are in Figs. 3.5 and 3.6. (a) Trajectory of $\boldsymbol{e_3}$ for G·C (orange dots) relative to the triad for A·T (marked $\boldsymbol{e_1}$–$\boldsymbol{e_3}$). $\boldsymbol{p_t}$: Centroid of the trajectory. Red solid (black dashed) circle denotes the major (minor) bending direction. $\boldsymbol{p_M}$/$\boldsymbol{p_m}$: Major/minor principal axis. (b) Illustration of major and minor bending directions (thick red and thin blue arrows). $\boldsymbol{p_M}$ and $\boldsymbol{p_m}$ are respectively normal to these directions. Pink arrow: $\boldsymbol{p_t}$. The last frame of the 100-ns simulation was used for visualization (only a part of the 16-bp oligo around the step is shown). Views are axial (left), and into the major (middle) and minor (right) grooves. Directions of motion for positive bending and twist angles are shown in Fig. 3.4.

distance increases (Fig. 3.10). It has been found that DNAs of several kilo base pairs in length overwind with stretch [54, 113, 190], which apparently contradicts the coupling between twist and step distance in our simulation. However, stretch measured by the end-to-end distance in experiments likely involves various deformational modes. As the major bending is the most compliant deformational mode (see below), with an applied tension, DNA stretches likely *via* reduction in the major bending angle rather than through increase in the step distance. As zero bending angles are measured about $\boldsymbol{p_t}$ possessing

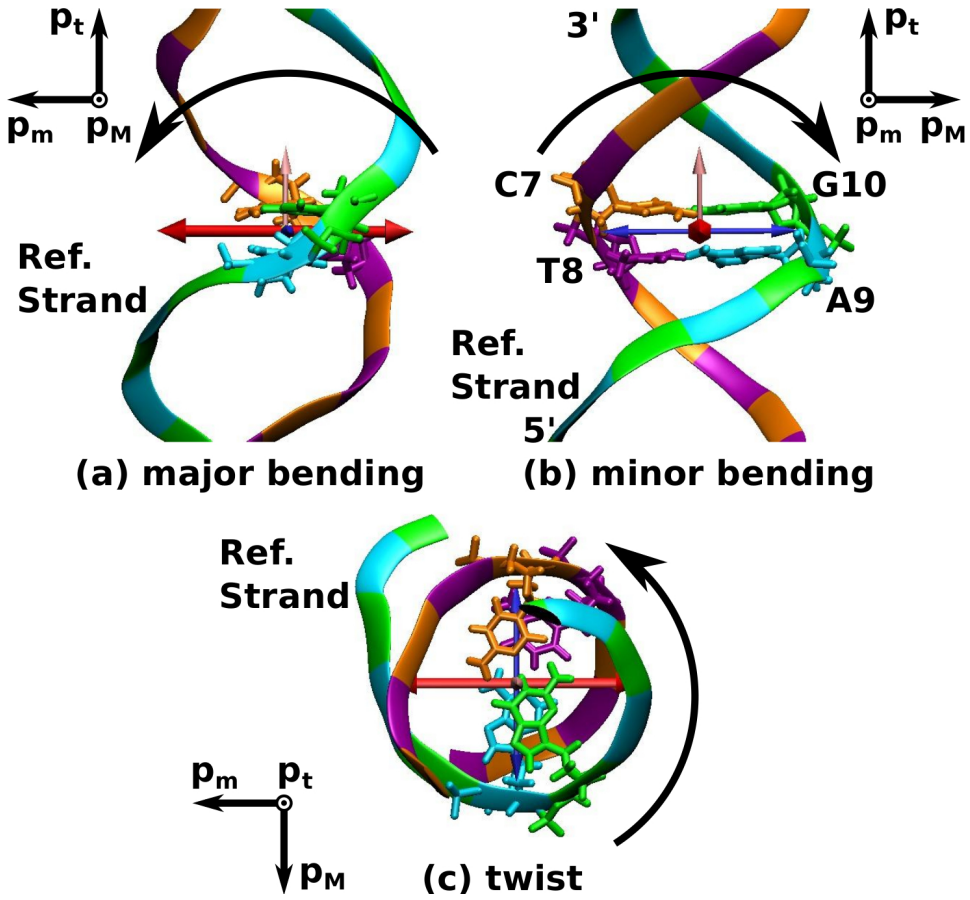

Figure 3.4: Directions of bending and twist that yield positive angles in Fig. 3.8 using the AG/CT as an example. (a) Major bending, (b) minor bending, and (c) twist. Thick red and thin blue arrows indicate the major and minor bending directions, which are perpendicular to respective principal axes.

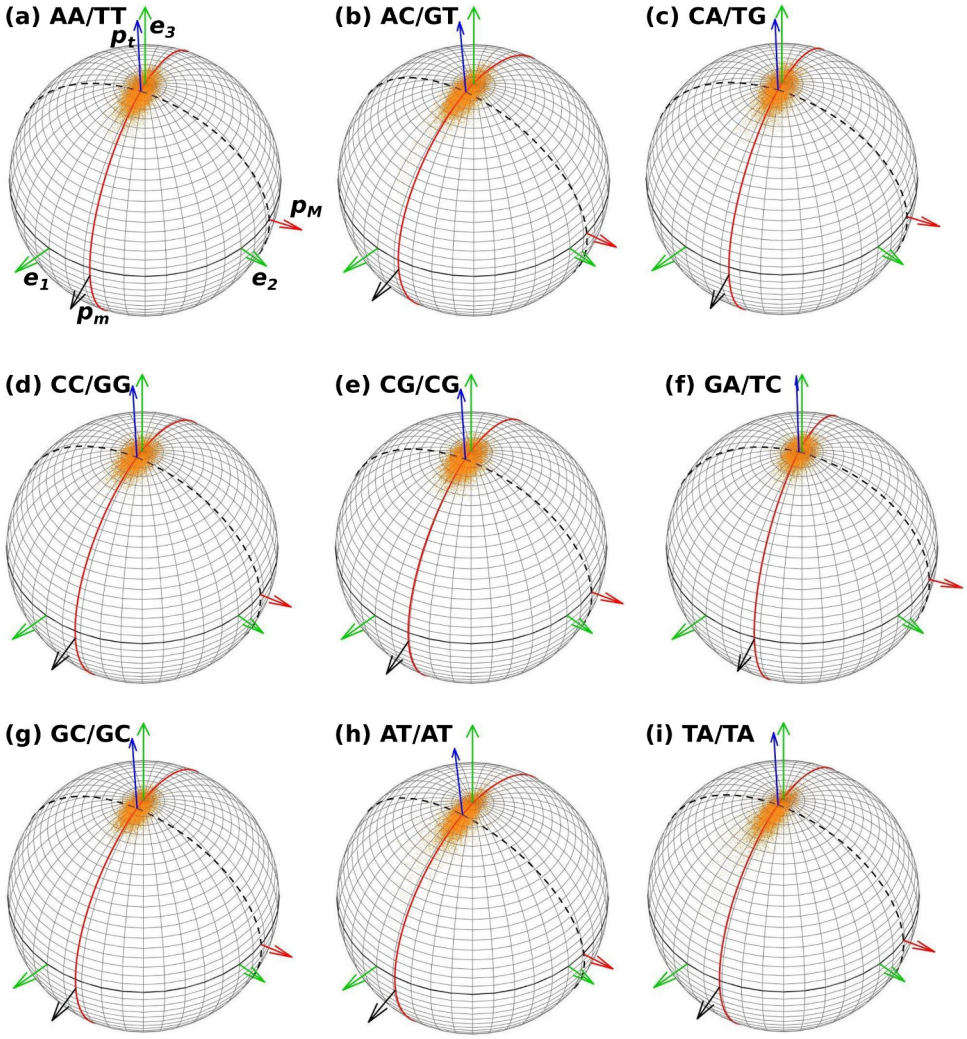

Figure 3.5: Trajectories of triads in individual dinucleotide steps (*cf.*, Fig. 3.3a and Table 4.4).

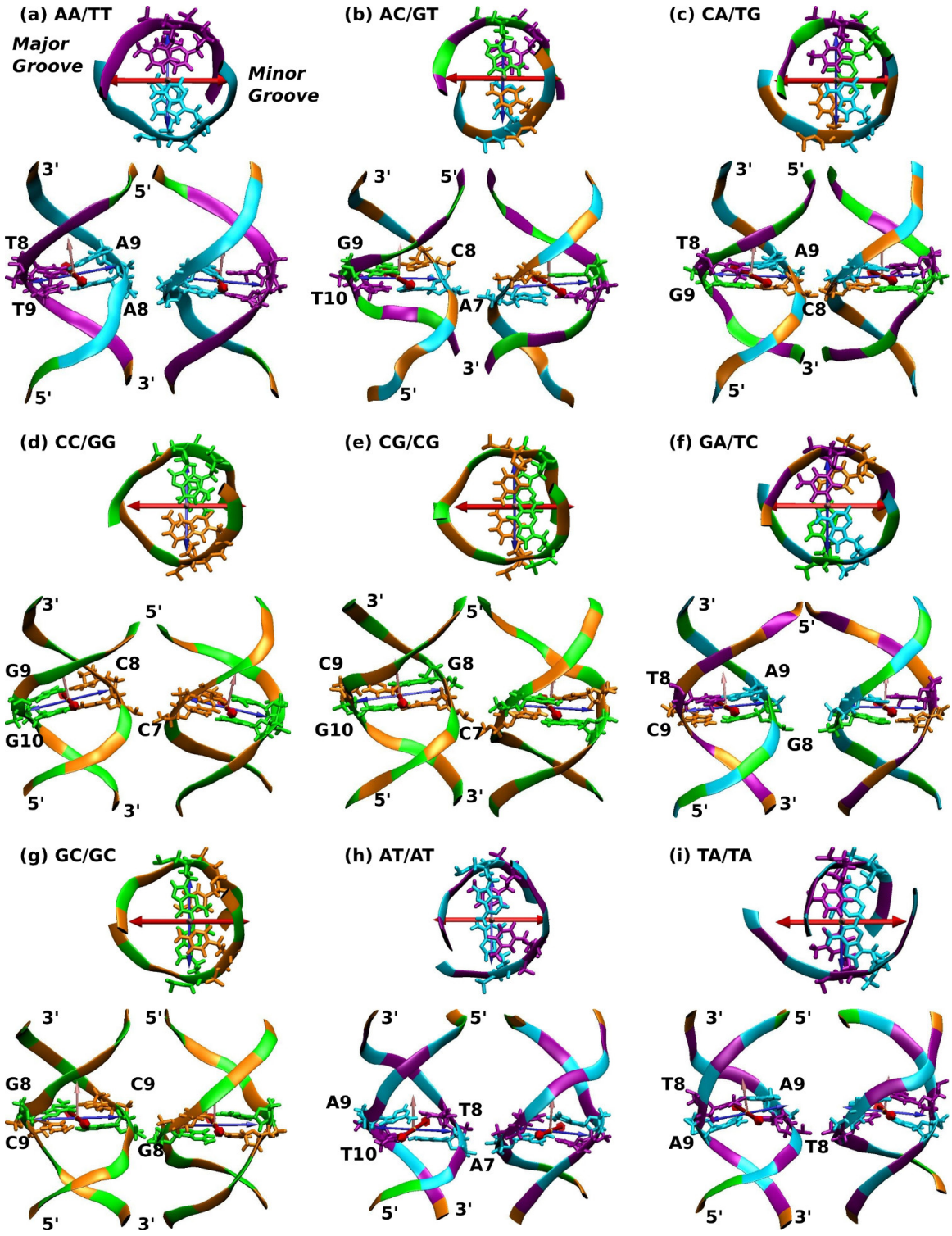

Figure 3.6: Principal axes and equilibrium curvature of of each step (*cf.*, Fig. 3.3b). Top: view in the direction of $\boldsymbol{p}_t$, lower left/right, view from the major/minor groove. For visualization, the last frames were used except for AT/AT and TA/TA (80 ns), and CC/GG (98 ns).

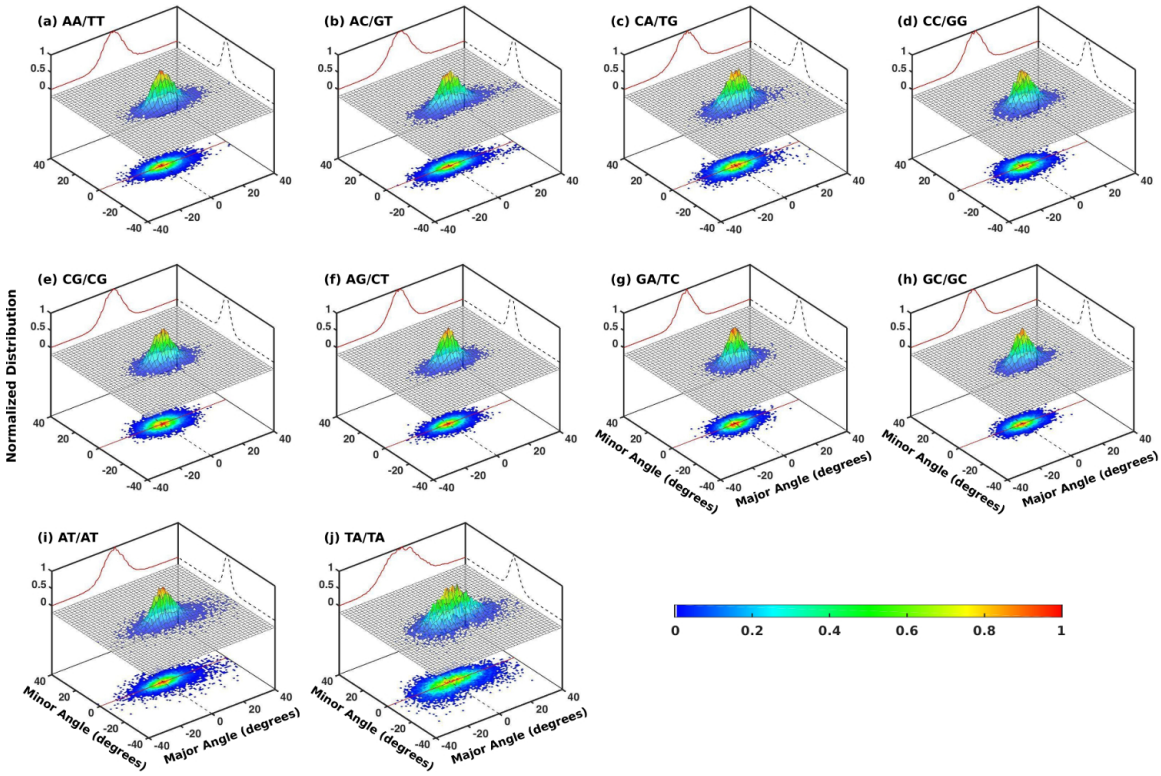

Figure 3.7: 2D histogram and surface plot of the major and minor bending angles for each dinucleotide step. Peak values are normalized to 1. Distributions are approximately symmetric with symmetry axes corresponding to the major and minor axes, indicating that the elastic energy does not have a coupling term between the two angles.

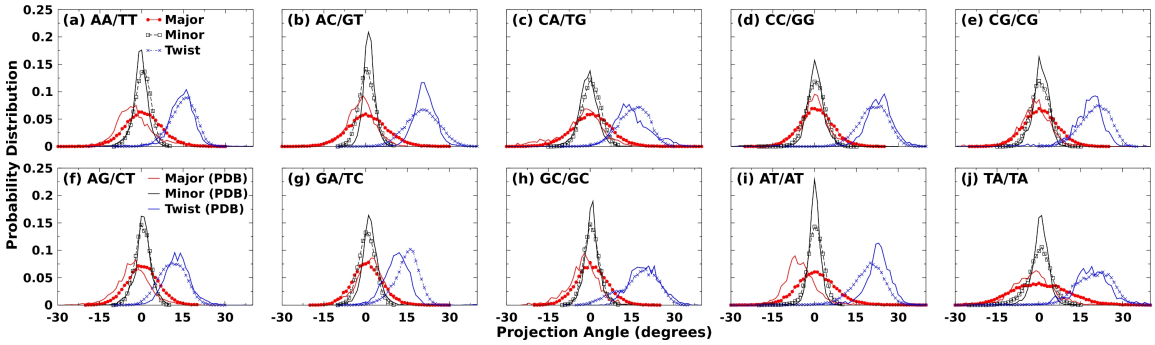

Figure 3.8: Distribution of the major bending (red), minor bending (black), and twist (blue) angles. Solid lines without symbols: distributions measured in PDB structures of protein-DNA complexes. Overall agreement between respective distributions indicate that during simulation, the dinucleotide steps sample conformational spaces comparable to those observed in crystal structures.

the equilibrium curvature (Fig. 3.3a), unbending by applied tension will cause the major bending angle to become negative. Within equilibrium fluctuation without load, although bending and twist are mostly decoupled, for more negative major bending angles reached by an applied tension, the twist angle is likely to increase as the conformation moves towards the upper left corners of the graphs for the major bending angle in Fig. 3.9. This is consistent with the experimentally observed overwinding under tension [54, 113, 190]. Yet, additional simulations with tensile loads are needed to systematically investigate the structural mechanism for the stretch-twist coupling of DNA.

Among the helicoidal parameters, roll, tilt, and helicoidal twist (we use 'helicoidal' to distinguish it from the twist as an order parameter in our analysis) describe relative angles between base pairs in a dinucleotide step [119]. To evaluate them for the equilibrium states in our simulation, we selected frames where each of the three angles in Fig. 3.8 falls within a half of its standard deviation from the average. In these frames, dinucleotide steps have conformations close to the equilibrium ones. We measured their distributions of roll, tilt, and helicoidal twist, and compared them with distributions for the entire 50-ns period

Step	Tilt			Roll			Twsit		
	selected	all	ratio (%)	selected	all	ratio (%)	selected	all	ratio (%)
AA/TT	3.64	4.49	81	3.60	6.35	57	2.70	4.98	54
AC/GT	4.09	5.09	80	3.01	6.69	45	3.15	6.17	51
AG/CT	3.63	4.21	86	2.98	5.78	51	2.78	4.98	56
AT/AT	4.22	4.91	86	3.26	7.41	44	2.92	8.96	33
CA/TG	3.97	5.19	76	3.86	8.12	48	3.08	5.74	54
CC/GG	3.82	5.24	73	3.23	6.12	53	3.25	5.46	59
CG/CG	4.42	5.82	76	3.26	7.37	44	3.10	5.39	57
GA/TC	3.52	4.69	75	2.88	5.83	49	2.75	4.72	58
GC/GC	3.87	5.14	75	2.84	5.71	50	3.15	6.80	46
TA/TA	4.28	5.78	74	4.41	9.54	46	3.23	7.89	41

Table 3.4: Standard deviations (σ) in tilt, roll, and helicoidal twist for frames that have the major and minor bending, and twist angles within 0.5σ from the respective average values (*selected*), compared with those over the entire 50-ns measurement period (*all*). Units are in degrees. The ratios of σ between the selected and the full data sets are also listed.

(Fig. 3.11). If the width (standard deviation) of the distribution of a helicoidal angle for the selected frames is comparable to that for the entire time interval, this would mean that the fluctuation in the helicoidal angle is insensitive to the equilibrium conformation of the dinucleotide step. The ratio of the widths for roll was 44% (AT/AT) to 57% (AA/TT), and for the helicoidal twist, it was 33% (AT/AT) to 59% (CC/GG). By contrast, for tilt, it was 73% (CC/GG) to 86% (AG/CT) (Table 4.4; Fig. 3.11). The large variation in tilt is likely because axes defining helicoidal parameters change orientations relative to the principal axis set $\{\boldsymbol{p}_m, \boldsymbol{p}_M, \boldsymbol{p}_t\}$ during simulation. Fig. 3.12 shows an example snapshot for the CA/TG step taking the equilibrium conformation while tilt deviates significantly from its average value. This indicates that the helicoidal parameter tilt is a degenerate quantity that does not take a well-defined value for the equilibrium conformation of a dinucleotide step, thus it cannot be used to describe the elastic motion of DNA. As a related issue, it was suggested that the dinucleotide step deformability can be achieved only by taking

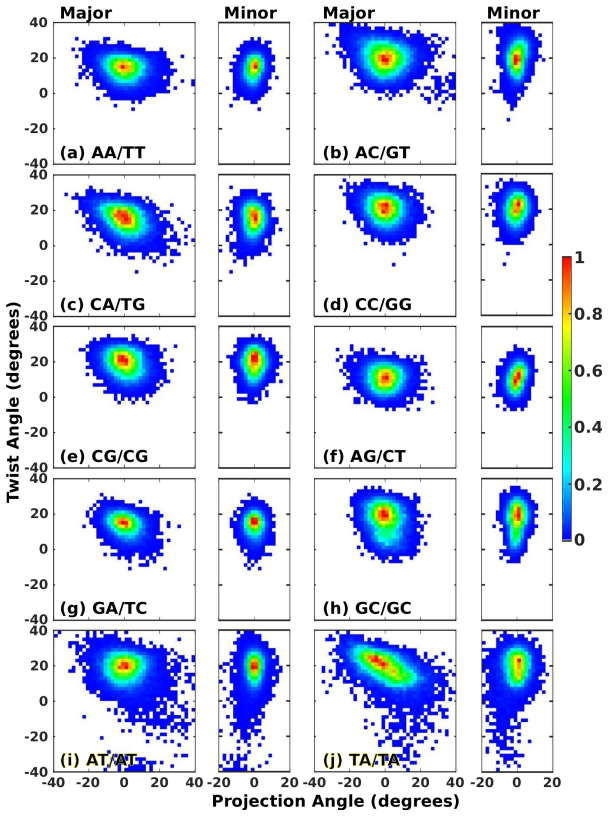

Figure 3.9: Histograms of the major/minor bending angles *versus* the twist angle (normalized by peak values). Although high-population regions are mostly round or vertical (little coupling between bending and twist), for large negative major bending angles, the histogram is populated more in the region with larger twist angles, suggesting a possible overwinding upon stretch by unbending of DNA.

into account the couplings between helicoidal parameters [101], and diagonalization of the stiffness matrix [100, 142] may not be effective due to the degeneracy of tilt. This further highlights the importance of calculating order parameters based on the principal axes for describing DNA's conformational motion as an elastic rod.

3.3.2 Sequence-dependent stiffness and persistence length

We apply the equipartition theorem [96] to the variance of bending angles to calculate the corresponding stiffness (Eqs. 3.5 and 3.6) [94, 210]. Even though twist is more

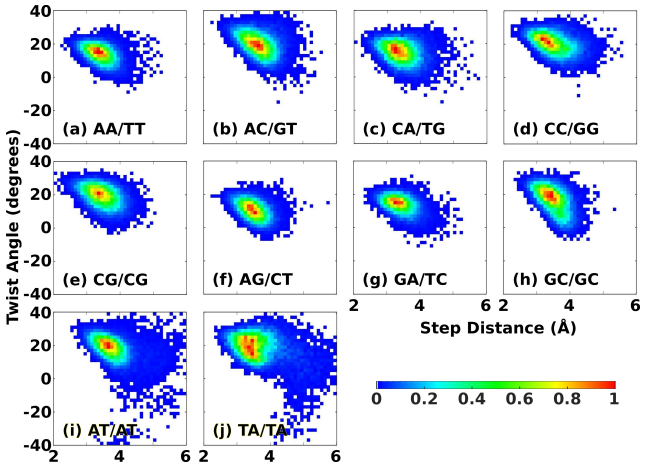

Figure 3.10: Histogram of the twist angle *versus* the step distance (normalized by peak values). In high-population regions, the step distance and twist angle are more negatively correlated than between twist and bending angles (Fig. 3.9).

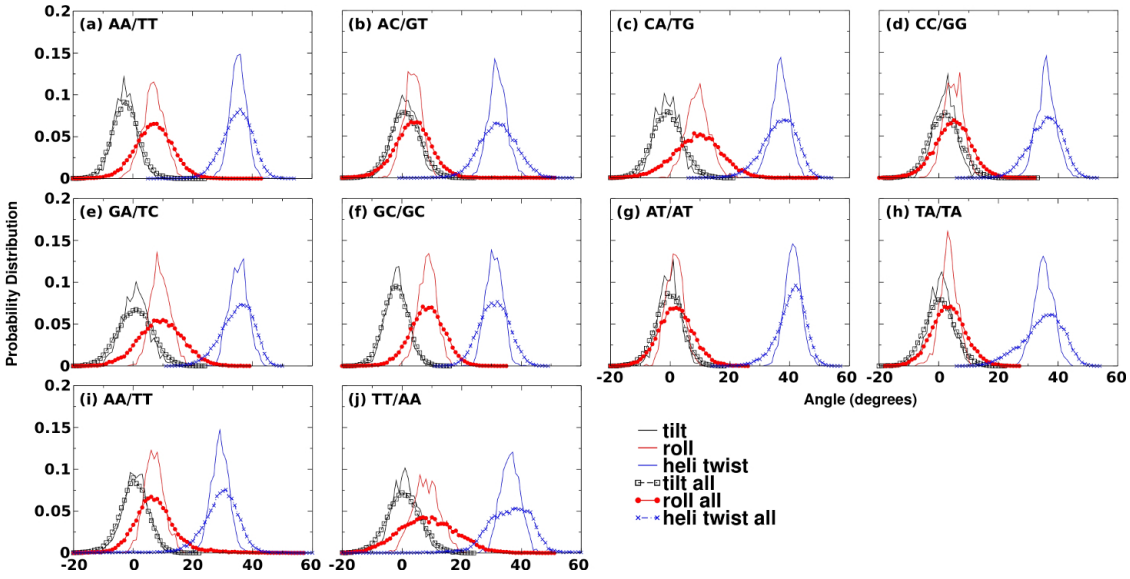

Figure 3.11: Assessing the degeneracy of the helicoidal parameters, roll, tilt, and helicoidal twist. Symbols: distributions obtained using all frames that we used for Fig. 3.8. Lines without symbols: distributions for frames where the two principal bending angles and twist (not helicoidal twist) lie within half the standard deviations from respective averages. For tilt, the two distributions have nearly the same width (Table 4.4).

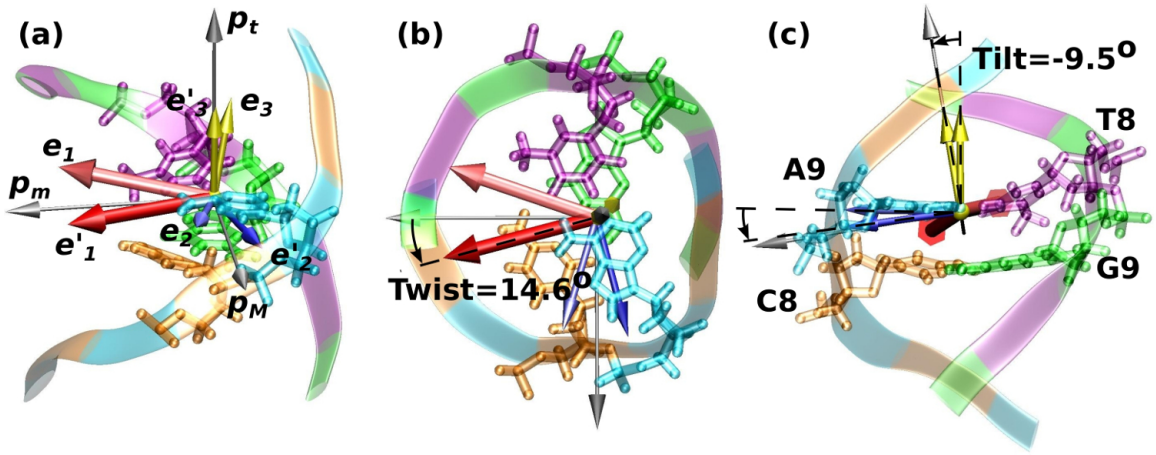

Figure 3.12: Example snapshot of a CA/TG step in near-equilibrium configuration despite having a large tilt. Triad $\{e_1, e_2, e_3\}$ for the C·G step is in lighter colors than the next triad $\{e'_1, e'_2, e'_3\}$ for the A·T step. Silver arrows: principal axis set. (a) e'_3 is aligned with p_t since bending is close to the equilibrium state. (b) Axial view. Twist (14.6°) is close to the average (15.1°; Table 4.2). (c) Another view showing tilt ($-9.5°$). The average±standard deviation for tilt is $-1.5° \pm 5.2°$ (Fig. 3.11c, Table 4.4).

skewed (Fig. 3.8), we calculate its stiffness as a measure of the torsional flexibility of DNA (Eq. 3.7). Stiffness calculated using the first and the last half of the 50-ns measurement interval were nearly identical except for [ATAT], where stiffness for the minor bending and twist decreased over time (Fig. 3.13), but the major bending stiffness that contributes the most to the overall flexibility of DNA, remains nearly constant over time. This suggests that the energy landscape in the most compliant major bending direction is the least rugged, so that the major bending angle equilibrates rapidly. In comparison, a previous estimate for the minimum simulation length required to attain equilibrium behavior of DNA was 20 ns [128]. More recently, microsecond-long simulations have been performed on systematically chosen oligos by the 'Ascona B-DNA consortium' (ABC) [154]. They revealed that tetramer-level (next-nearest neighbor) effects are pronounced in several helicoidal parameters. But roll and tilt that represent relative rotation between base pair steps, only had weak dependence of their average values on flanking steps. Although

fluctuations in roll and tilt were influenced more in some dinucleotide steps, an adaptive umbrella sampling study of various DNA oligos showed negligible cooperativity in energetics between adjacent roll angles [198]. And as explained above, tilt does not correlate strongly with the equilibrium fluctuation of order parameters (Table 4.4; Figs. 3.11, 3.12). Thus, it is unlikely that tetramer effects have any strong influence on the bending stiffness of dinucleotide steps measured using principal axes. Furthermore, we show below that distributions of order parameters obtained from simulation agree with those of PDB structures, which also suggests reasonable sampling (lines without symbols in Fig. 3.8).

In the ABC simulation, helicoidal twist had greater tetramer effects in several dinucleotide steps, resulting in non-Gaussian behaviors [154]. This is consistent with the skewed distribution of twist as an order parameter. In particular, helicoidal twist of the CG step was distributed with multiple (mostly two) peaks [33, 154]. Although our simulation time was not sufficient to observe additional peaks in helicoidal twist (data not shown), since distributions of twist (as an order parameter) in PDB structures do not possess multiple peaks (Fig. 3.8), any other peaks that might emerge in longer simulations are likely to be less important for the elasticity of DNA. The ABC simulations also observed behaviors such as spontaneous kink formation and base pair opening (*cf.*, Fig. 3.2). Whereas contributions by such transient states to the overall elasticity of DNA need additional investigation, our results below suggest that the dinucleotide-level order parameter analysis describes elasticity of DNA fairly well.

On average, the major bending stiffness varies from 0.47×10^4 (TA/TA) to 1.52×10^4 pN·Å^2 (GA/TC), and the minor bending stiffness from 2.43×10^4 (TA/TA) to 6.07×10^4 pN·Å^2 (AG/CT) (Fig. 3.14 and Table 4.3). The general trend that steps containing G and C are stiffer than those containing A and T is consistent with previous experiments [36, 66, 144] and simulations [126, 100, 99, 197]. Also, steps with purines on one strand and pyrimidines on the other have higher stiffness than those with them mixed, likely due

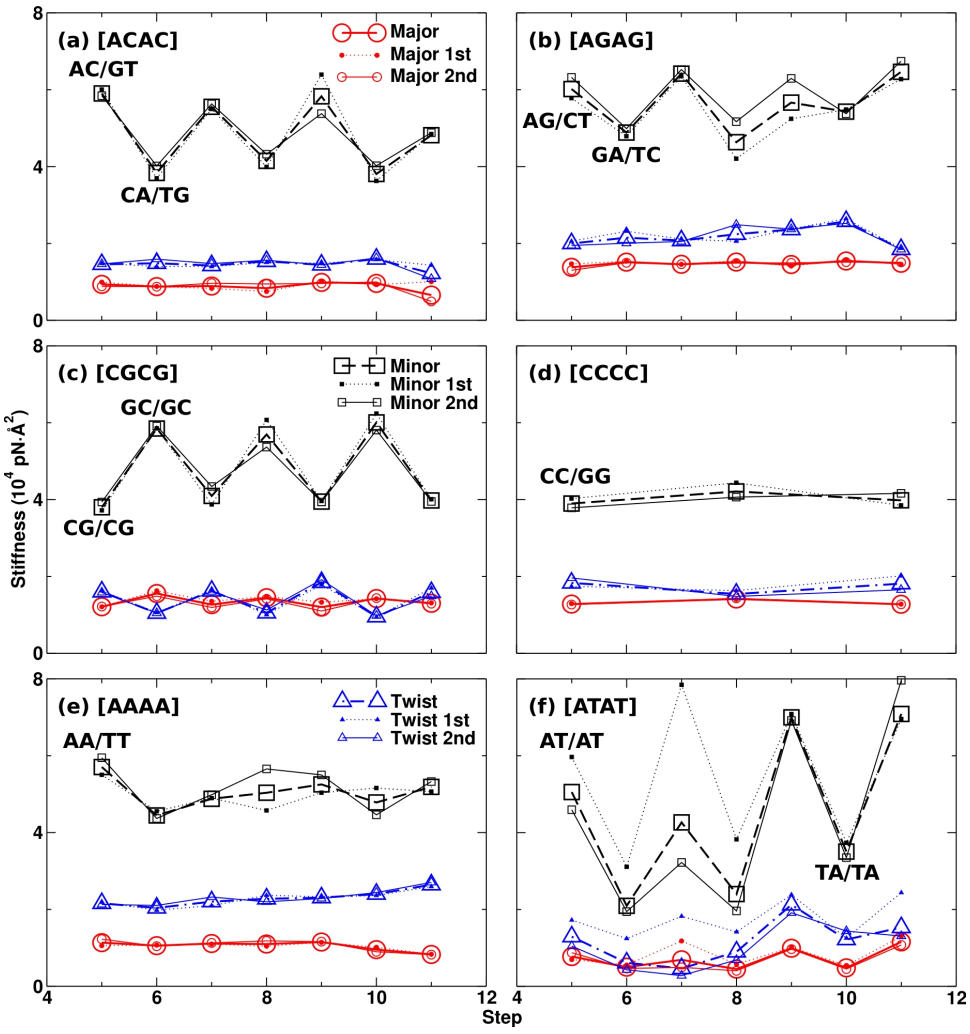

Figure 3.13: Stiffness of the three deformational modes between dinucleotide steps in each oligo. Stiffness measured for a given type of dinucleotide step is in Fig. 3.14 and Table 4.2. Legends in (a), (c), and (e) apply to all panels. Smaller symbols denote calculations based on the first and the last half of the 50-ns measurement intervals.

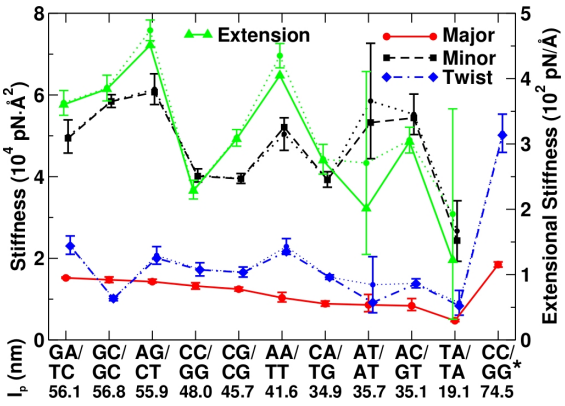

Figure 3.14: Stiffness and persistence length l_p of dinucleotide steps. Dots with error bars are average±standard deviation for calculations on individual steps in Fig. 3.13. CC/GG in the A-DNA conformation (from [CCCC*], marked by a star) is stiffer; its minor bending stiffness (9.37×10^4 pN·Å^2) and extensional stiffness (7.29×10^2 pN/Å) are not shown. See Table 4.3 for numerical values of stiffness and Fig. 3.13 for stiffness of individual steps.

to better stacking between neighboring bases. This is consistent with a previous structural analysis of protein-DNA complexes that found sharp bends occur mostly at pyrimidine-purine steps [147]. TA/TA has the lowest major bending stiffness. Since it is consistently low in all TA/TA steps in [ATAT] (Fig. 3.13f), it is unlikely due to destabilization of [ATAT]. Besides, the stiffness of the neighboring AT/AT step within the same [ATAT] oligo is in an intermediate range. Thus, the lower flexibility of TA/TA is unlikely because of the force field used [64].

Persistence length l_p of dinucleotide steps (Method; bottom in Fig. 3.14) can be further combined for a given DNA sequence (Eq. 3.8). For five ∼200-bp oligos previously studied by cyclization experiments [53], our calculation captures the trend of low *versus* high l_p though our values are shorter by 3.8–11.5 nm (Fig. 3.15). In fact, single-molecule DNA looping experiments revealed that oligos are more flexible than estimated by bulk cyclization experiments [103, 214]. Also, l_p in Fig. 3.15 average to 41 nm, which is very close to previous estimate, 43 nm, based on both simulation [142] and also a

Step	κ_M ($\times 10^4$)	κ_m ($\times 10^4$)	κ_t ($\times 10^4$)	κ_E
GA/TC	1.52	4.94	2.30	360
GC/GC	1.47	5.84	1.02	384
AG/CT	1.43	6.07	2.00	451
CC/GG	1.32	4.01	1.72	228
CG/CG	1.24	3.95	1.66	308
AA/TT	1.03	5.21	2.16	405
CA/TG	0.89	3.92	1.54	275
AT/AT	0.87	5.32	0.92	201
AC/GT	0.84	5.44	1.38	303
TA/TA	0.47	2.43	0.83	122
CC/GG*	1.85	9.37	5.01	729

Table 3.5: Stiffness of each step. κ_M: major, κ_m: minor, κ_t: twist, and κ_E: extension. Steps are listed in decreasing order of κ_M, as in Fig. 3.14. κ_M, κ_m, and κ_t are in units of pN·Å^2, and κ_E in pN/Å. CC/GG*: stiffness of the CC/GG step in A-DNA conformation.

recent experiment [23]. Similarly, the calculated twist stiffness, 0.83×10^4 (TA/TA) to 2.30×10^4 pN·Å^2 (GA/TC) (Table 4.3), are slightly lower than experimental estimates, $\{1.6\text{–}2.9\}\times10^4$ pN·Å^2 [59, 53, 114]. Note that we used mostly B-DNA conformations for calculation while DNA may transiently visit other conformations such as A-DNA (Fig. 3.1) whose stiffness in the case of CC/GG is much higher (Fig. 3.14). A complete absence of other particles or debris that bind to DNA and alter its conformational motion, as noted for DNA ligase in bulk cyclization experiments [103], may also contribute to the higher flexibility of DNA in simulation.

As an additional test, we calculated local stiffness of dinucleotide steps from simulations of two 30-bp oligos that are respectively cleavable (CLV) and non-cleavable (NON-CLV) by type-II topoisomerase (Method and Table 3.2) [106], and compared them with those from the oligos in Table 4.1 (Fig. 3.16). Since stiffness is inversely proportional

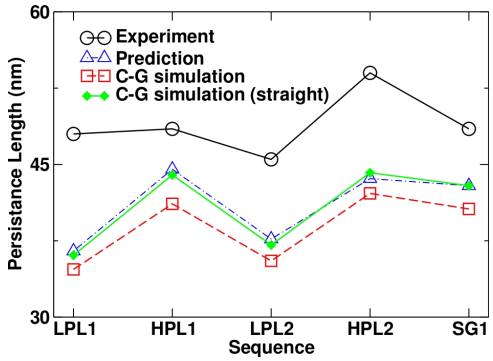

Figure 3.15: Comparison of persistence lengths between experiment and simulation. Oligo names and sequences are from Ref. [53]. 'LPL' and 'HPL' stand for low and high persistence length, respectively named based on sequence composition. Circle: bulk cyclization experiment [53]. Triangle: calculation using Eq. 3.8. Square: C-G simulation of oligos with equilibrium curvature. Diamond: C-G simulation of straight oligos.

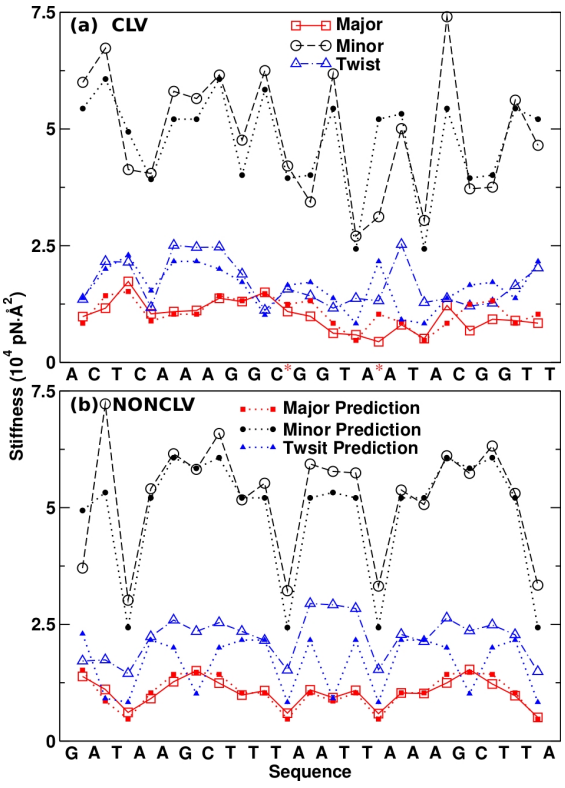

Figure 3.16: Predicting elastic stiffness of arbitrary DNA sequences. Open symbols: stiffness measured in simulations of (a) CLV and (b) NONCLV (Table 3.2). Solid symbols: calculation based on stiffness of individual steps (Fig. 3.14). In (a), cleavage sites by type-II topoisomerase are marked by stars in the sequence [106].

to the variance of measured angles (Eqs. 3.5–3.7), uncertainty is higher for smaller variance, which results in generally greater mismatch between measured and predicted values for the minor and twist stiffness. Nevertheless, the overall good agreement between the stiffness calculated based on different oligo sequences supports that local elastic properties of DNA are determined primarily by nearest-neighbor dinucleotide steps without any strong influence by tetramer effects, which is consistent with previous experiments and simulations [67, 53, 198]. As mentioned above, although long ABC simulations revealed tetramer effects in certain helicoidal parameters, the agreement between dinucleotide-level model described by principal axes and experiment [53] supports that our analysis provides a satisfactory picture of the elasticity of DNA. On the other hand, the CC/GG step was in either A- or B-DNA conformation depending on the sequence of the oligo (Fig. 3.1c *vs.* g), suggesting that interactions at the dinucleotide level do not completely determine DNA conformation. In case different conformational states appear within a long simulation, elasticity of those states can be calculated separately using our approach and then combined, weighted by their relative abundance.

Of note, in CLV, the major bending stiffness becomes lower starting from the first topoisomerase cleavage site (C*G in Fig. 3.16a), past the second cleavage site (A*A). Twist stiffness between the two cleavage sites are also lower than the surrounding regions, and the mismatch with predictions is higher in this region. As explained below, major bending and twist are dominant modes of deformation in protein-DNA interactions. The enhanced flexibility of the cleavage region is reminiscent of the local lability of the collagen triple helix for binding and cleavage by matrix metalloproteinase [210, 29] and also the actin binding sites in tropomyosin [94]. Further studies are needed to more carefully determine whether variations in the local flexibility of DNA serves as a recognition signal for topoisomerase [106].

With the bending stiffness known, we can estimate the conformational relaxation times

of oligos in water [210, 70]. Denoting the length of an oligo as L (3.4 Å per dinucleotide step), diameter $d = 20$ Å, and viscosity of water $\eta = 8.51 \times 10^{-4}$ Pa·s, its slowest relaxation time is [70] $\tau = \frac{c_\perp}{\kappa_f}(\frac{L}{\omega_1})^4$. Here, $c_\perp = 4\pi\eta/[\ln(L/d) + 0.84]$ is the transverse drag coefficient per unit length of the rod, and the constant ω_1 depends on the boundary condition, which is 4.73 for unconstrained ends of the oligo. Using the smallest bending stiffness $\kappa_f = 4.7 \times 10^3$ pN·Å^2 for TA/TA (Table 4.3), $\tau = 168$ ps for 16-bp oligos (*cf.*, Table 4.1) and 1.77 ns for 30-bp oligos (CLV and NONCLV). Thus, elastic fluctuation of oligos should be well-sampled within the 100-ns simulation time, albeit conformational transitions may involve longer time scale.

3.3.3 Coarse-grained model of DNA

We use the equilibrium values and stiffness of the order parameters identified from the atomistic MD simulations (Tables 4.2 and 4.3) to construct C-G models of B-DNA oligos (a hyphen is used for C-G to avoid confusion with DNA sequence). In our C-G model, each dinucleotide step has three degrees of freedom, the major bending, minor bending, and twist (see Methods for simulation details). Many previous C-G models of DNA use a larger number of effective atoms (thus more degrees of freedom) with the aim of capturing internal motion or semi-atomistic behaviors such as melting or fraying (see [34, 203] and references therein). Yet, as explained in Introduction, to describe the elastic motion of DNA, only the four orthogonal order parameters are needed, of which the three angles used in our C-G model are the most relevant. This makes our simulation highly efficient. Time evolution of the system is achieved by the Brownian dynamics method [75].

As a test, we used 10- to 1000-bp oligos with repeating sequence and measured their l_p by fitting the mean-square end-to-end distance to the expression for the wormlike chain model (Methods and Fig. 3.17). It increases with the oligo's length, reaching an asymptotic value that is close to the estimate based on the stiffness of dinucleotide steps (Fig. 3.17a).

The length dependence of l_p is due to the equilibrium curvature of oligos [185, 218]. This was verified by using modified oligos that are straight in equilibrium but have the same stiffness as the original, which had no length dependence (Fig. 3.17b). Since l_p based on dinucleotide stiffness only involves the two bending stiffness and not twist (Eq. 3.8), its agreement with that based on end-to-end distance fluctuation for long oligos indicates that twist plays little role. To confirm, we prepared another set of oligos that are torsionally rigid, which indeed had nearly the same l_p as the original ones (Fig. 3.17c). Furthermore, since long oligos had end-to-end distance-based l_p approaching the stiffness-based one, the static persistence length due to intrinsic curvature of DNA [185, 218] plays negligible role.

Using the C-G model, we measured l_p of oligos in Fig. 3.15. Since they are $\sim$200-bp in length, the measured l_p are shorter than the stiffness-based ones, while straight oligos having zero equilibrium curvature had the same l_p (square *vs.* diamond in Fig. 3.15). These results demonstrate that our C-G model can be used to elucidate various aspects of DNA's conformational behavior.

3.3.4 Partitioning of elastic energy in protein binding

An important question regards how the flexibility of an isolated DNA affects binding to proteins [193], for which we analyzed available x-ray structures of protein-DNA complexes. Out of 2318 PDB structures initially identified, we selected only those with standard bases and skipped steps containing broken base pairs, so that our analysis is applicable. This reduced the number of PDB structures analyzed to 1381, with 33360 steps (Method). In these structures, since the conformation of a DNA would be determined mainly by interaction with proteins, crystal contacts likely play little role compared to cases for DNA-only structures. We calculated distributions of order parameters in individual dinucleotide steps (solid lines in Fig. 3.8). Despite the fact that protein-DNA

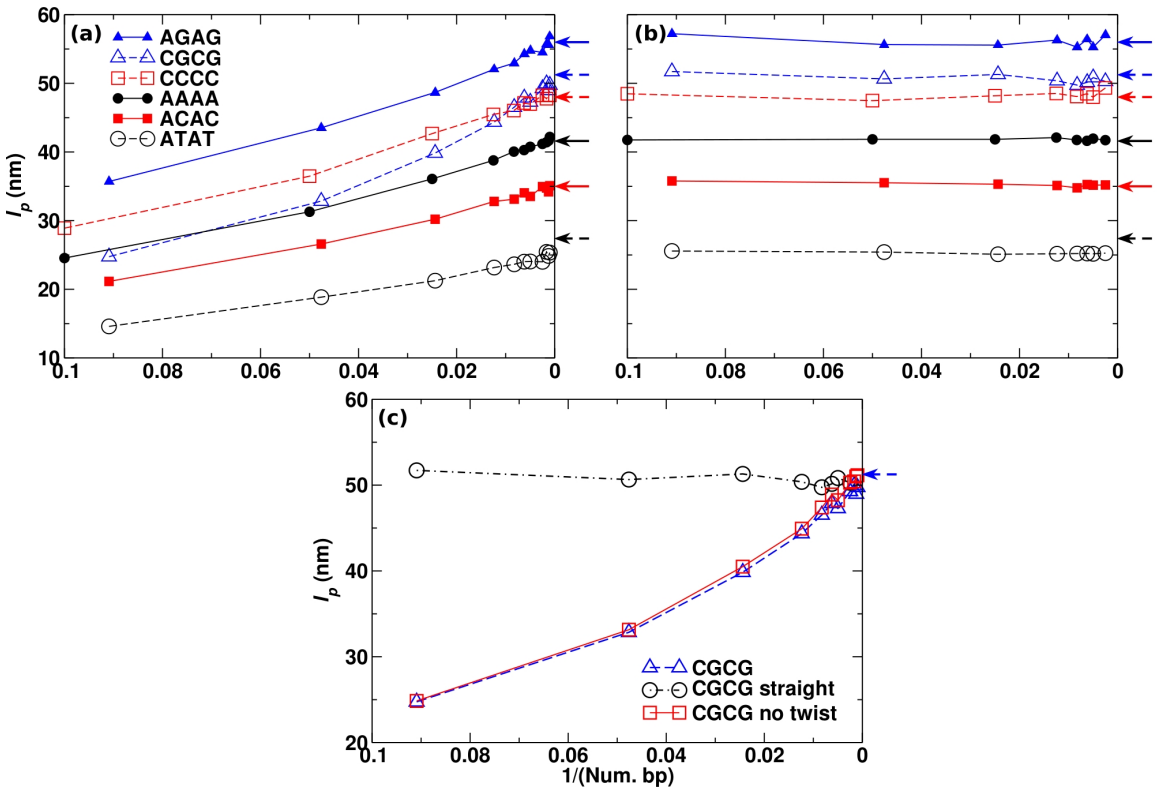

Figure 3.17: Length dependence of the persistence length l_p in C-G simulation. A reciprocal scale is used for the horizontal axis, so the oligo length increases to the right (Num. bp: Number of base pairs in an oligo). Horizontal arrows on the right of each plot are l_p calculated based on dinucleotide stiffness (Eq. 3.8). (a) Oligos with equilibrium curvature (Table 4.2). (b) Oligos with zero equilibrium curvature (straight). Their l_p show no length dependence. (c) CGCG oligo possessing equilibrium curvature but twist motion disabled (square), which behaves nearly the same as the original oligo undergoing twist motion (triangle). Oligo names here denote the repeating sequence and there are no capping sequences in the C-G models as in atomistic MD simulation (Table 4.1). For oligos with alternating sequences *e.g.*, CGCG, to have equal number of CG/CG and GC/GC steps, we used odd numbers of base pairs. The ATAT oligo has the largest discrepancy between the asymptotic value of l_p and the stiffness-based l_p (open circles in panel (a,b)). This is because AT/AT and TA/TA had distributions of major bending angles to be the least Gaussian (Fig. 3.8i,j): Their kurtosis were 2.87 and 1.83, respectively, while they were in the 0.14–0.42 range for major bending angles in all other dinucleotide steps. Thus, harmonic approximation to the major bending motion of the AT/AT and TA/TA steps (Eq. 3.9) becomes less accurate.

structures and simulations of isolated oligos are compared, the distributions match fairly well, suggesting that in most cases DNAs bind to proteins without significant deformation. Greater discrepancy in the distributions for the major bending and twist angles indicate that the two deformational modes are used more extensively when binding to proteins than the minor bending. Away from the peaks of distributions, a 2D histogram of bending angles shows that high curvature or sharp turns of DNA are accomplished by large major bending angles (star in Fig. 3.18a). As the positive major bending direction aligns approximately with that of the major groove, our results are consistent with previous suggestions that the wide major groove tends to interact with proteins [151, 165].

For each PDB structure, we calculated elastic energies of DNA for the major bending, minor bending, twist, and extension. Since the length of DNA varies among PDB structures, we averaged the energies per dinucleotide step for each structure, and plotted them in the increasing order of the total elastic energy per step (Fig. 3.18b and Appendix B; all energies below refer to average per step). Since partitioning of the elastic energy among the four modes fluctuates significantly, plots were smoothed using the Savitzky-Golay filter [159]. Fig. 3.18b thus reveals a general trend rather than features of individual structures.

Out of 1381 structures, 831 have the total elastic energy less than $2k_BT$ (82.8 pN·Å), confirming that majority of DNAs that we analyzed do not deform substantially when binding to proteins. Interestingly, in low-energy structures, twist takes up the major part of the total elastic energy (inset in Fig. 3.18b). The major bending energy occupies an increasingly higher portion, and becomes dominant for the total elastic energy above 4.16 k_BT (vertical arrows in Fig. 3.18b). For the last 33 structures (total elastic energy above 10.93 k_BT), extension takes the largest portion. They consist mainly of small antibiotic-bound structures (echinomycin-bound structures take the three highest elastic energy, with the side chains of echinomycin inserting between base pairs of DNA [32];

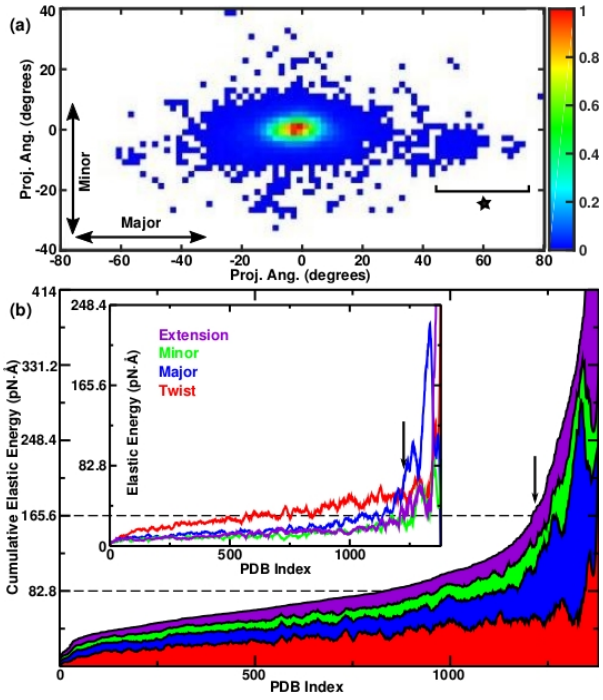

Figure 3.18: Conformation and energetics of DNAs in protein-DNA complexes. (a) 2D histogram of the major and the minor bending angles (normalized by the peak value). Star: structures with large major bending angles. (b) Decomposition of the elastic energy per dinucleotide step (Appendix B). Horizontal dashed lines mark 2 k_BT and 4 k_BT. Inset: Individual elastic energies (smoothed using the Savitzky-Golay filter [159]). Vertical arrow: PDB index above which the major bending energy becomes higher than the twist energy. Above about $4k_BT$ (base pair stacking energy) [56, 223], the linear elasticity assumption likely breaks down, so that energies in this regime should be regarded to represent the level of deformation rather than the actual storage of an elastic energy.

Appendix B) and complexes between DNA and TATA-box binding protein. Since the base pair stacking free energy is up to about 4 k_BT [56, 223], the assumption of linear elasticity likely breaks down in structures with the total elastic energy comparable to or greater than this. Instead of actually storing elastic energy, such highly deformed structures should thus be in different conformational states. Even in those cases, linear elasticity likely plays an important role during initial stages of binding when protein residues are not fully engaged into DNA.

3.4 Conclusions

A fundamental difficulty in understanding the dynamics of DNA is that it involves two scales: atomistic, and the mesoscale in which DNA behaves as a semi-continuum filamentous molecule. Although internal coordinates at the atomistic level completely specify DNA conformation, their large number of degrees of freedom make it difficult to describe mesoscale conformational motion. Helicoidal parameters significantly reduce the number while effectively describing atomistic-level structures of DNA [41, 119]. However, similarly as internal coordinates, they are not suitable order parameters for the motion DNA as an elastic rod. Without a proper choice of order parameters or reaction coordinates, energetics associated with conformational changes can be misleading [172]. The degeneracy of the helicoidal parameter tilt (Table 4.4 and Fig. 3.11) indicates that it partly reflects internal modes that have no direct relevance to the conformational motion of DNA as an elastic rod. The present study finds the order parameters by analyzing relative motions between base pairs in dinucleotide steps. During this procedure, many internal motion associated with $\sim$16 helicoidal parameters are expected to be integrated out. The calculated equilibrium conformation and stiffness (Tables 4.2 and 4.3) was used to further develop a C-G model of DNA which elucidates factors affecting the conformational motion of DNA. If necessary, our C-G model can incorporate nonlinear effects simply by changing the type of potential *e.g.*, by using a bimodal potential to capture kinking transition. Such extensibility of our model is possible due to its use of orthogonal order parameters.

We demonstrated that the stiffness calculated for the dinucleotide steps can be used to construct approximate flexibility maps of oligos with arbitrary sequence and estimate their persistence lengths. Although longer simulations or enhanced sampling methods may be needed to gain a more complete picture of DNA elasticity in presence of conformational transitions, the favorable agreement between our results and experiments in various aspects

suggests that our analysis based mainly on the B-DNA state in 100-ns production runs for each oligo, captures major features of the sequence-dependent DNA elasticity.

The present results also elucidate the role of sequence-dependent elasticity in protein-DNA interactions. Type-II topoisomerase binds to both CLV and NONCLV, but only bends CLV, indicating that mechanical compliance plays a significant role [106]. Our analysis suggests that the region containing the cleavage sites in CLV are more labile (Fig. 3.16). In the case of collagen, the site cleaved by matrix metalloproteinase (MMP) is characterized by an abrupt transition in stiffness and stability [46, 210]. Presence of a stiff region upstream to the MMP cleavage site may facilitate localized unfolding of the cleavage site. It is possible that a gradient in flexibility of DNA may also help topoisomerase to induce a regional bend [38]. Such local deformability is likely a general mechanism whereby proteins recognize target sites on DNA in addition to more static signatures such as kinks [193]. Our analysis further highlights the importance of twist and major bending in protein-DNA interactions (Fig. 3.18), which likely contribute to determining the binding free energy.

Of another note, in case DNA interacts with multiple ligands [131], establishing a more complete picture will require additional information including interfacial energies and kinetics of ligand binding. It is in principle possible to use our method to calculate the flexibility of a DNA-ligand complex by analyzing relative motions of suitably assigned triads. Weighted by the lifetime and number of ligands bound along a DNA, it would then be possible to calculate the effective persistence length as a function of the ligand concentration [131]. This is a subject of a future study for a specific DNA-ligand system.

Sequence-dependent flexibility and different deformational modes would be important for constructing DNA nanostructures as well [173, 180]. Since the present approach does not require any particular atomistic feature to work, it can also be applied to other types of filamentous systems including DNA in different structural states, RNA, and cytoskeletal filaments.

4. SEQUENCE-DEPENDENT EFFECT OF CYTOSINE METHYLATION ON DNA MECHANICS

4.1 Introduction

DNA methylation plays a crucial role in gene regulation [15, 85, 186]. For vertebrates, methylation typically occurs on the C5 atom of cytosine in the CpG dinucleotide step (Fig. 4.1a) [62, 186]. Cytosine methylation leads to gene suppression [166, 14, 127], chromosome inactivation [153, 207], and genomic imprinting [111, 83]. Abnormal hypermethylation of the CpG-rich region (the CpG island), is frequently observed in cancer cells [82, 84, 10]. The CpG methylation is heritable [200, 14, 40], highlighting its importance in development and disease progression.

The methyl group on mCYT (5-methyl-CYT) directly affects the interaction with DNA-binding proteins. For example, it prevents the binding of transcription factors [220, 208, 72] or allows binding of other proteins such as methyl-CpG-binding domain proteins that blocks transcription [18, 136, 86, 4, 202]. In addition to affecting the interaction with individual DNA-binding proteins, methylation can alter chromatin structure, thereby restrict accessibility of transcription factors [105, 155, 80, 31, 81, 104, 139]. However, controversial results are reported on the role of methylation for nucleosome formation, that is either promoted [105, 31, 104] or suppressed [155, 80, 81, 139]. Another study even suggests that CpG methylation has no effect on nucleosome stability [98]. Thus, methylation can affect the conformational behavior of DNA in different ways, likely depending on the local sequence around the methylation site. A central aspect in this regard is mechanics. Flexibility significantly influences protein binding to DNA [88, 16, 171, 215], and packing of DNA such as in nucleosome formation [155, 81, 140, 139]. Although it is generally agreed that DNA becomes stiffer upon methylation [35, 189, 155, 81, 139], its dependence

on local DNA sequence is not established.

Since experiments on DNA mechanics so far does not have the dinucleotide-level resolution, molecular dynamics (MD) simulation has been instrumental for obtaining insight into the sequence-dependent DNA mechanics [100, 102, 90, 154, 212, 120]. However, previous computational studies show inconsistent results regarding the effect of methylation, where it either stiffens [155] or does not affect [209, 90] the CpG step. Another issue with previous analyses is that the helicoidal parameters [41] were used for stiffness calculation. Although helicoidal parameters effectively describe the atomistic structure of DNA, they do not form a suitable set of order parameters describing deformation of DNA as an elastic rod. In the case of helicoidal tilt, its fluctuation is not centered about the equilibrium conformation [212]. We thus developed a principal-axes based description of DNA mechanics that uses four orthogonal order parameters: two principal axes of bending, twist about an axis orthogonal to principal axes (which is not the same as the helicoidal twist), and extension about this axis. This approach allows to link atomistic behavior of dinucleotide steps to the meso-scale conformational motion of DNA, as well as its deformation upon binding to proteins [212].

In this study, we perform MD simulation and apply the principal axis-based analysis to study the mechanical properties of methylated B-DNA. We find strong sequence dependence in the effect of methylation. The stiffness of the ^{Me}CpG step (^{Me}C: mCYT) itself remains similar to the un-methylated ones, whereas its neighbors become generally stiffer. However, when THY locates on the 5'-side of mCYT, the twist stiffness decreases for both ^{Me}CpG and the adjacent step. Also, hyper-methylation of DNA (more than two consecutive ^{Me}CpG) makes it stiffer compared to the case with isolated ^{Me}CpG steps. This is consistent with the presence of the 'methylation threshold' required to inactivate certain genes [71]. In comparison, the equilibrium curvature of DNA is unaffected by methylation except for oligos whose ^{Me}CpG step is flanked by T^{Me}C/GA. The sequence-dependent changes in

the stiffness of neighboring steps are due to the interaction between the methyl group of mCYT with the methylene group on the 5'-side deoxyribose and with the methyl group of THY. We also found that methylation alters the local hydration structure and may influence the conformational motion of DNA. The atomistic origin for the sequence-dependent effect of methylation as revealed in the present study will aid with understanding the physical role of methylation for DNA packaging and interaction with other proteins.

4.2 Methods

4.2.1 DNA oligo generation

Oligos used for simulation are listed in Table 4.1. Except for $[cg]_8$ and $[cg]_6$, both ends of an oligo were capped by $d(CGCG)_2$ to prevent fraying [107, 143, 226]. All oligos were built in the B-DNA form by using X3DNA [119]. Missing hydrogen atoms in the initially generated structure were added by using CHARMM [21]. Four oligos, [AAAcg], [TTTcg], $[cg]_8$, and [CGcg], were used for our main analyses. The first two oligos have a ^{Me}CpG dinucleotide step in the middle, flanked by Ac/gT and Tc/gA, respectively. We used $[cg]_6$ to compare with $[cg]_8$, for testing its length dependence, and [GCgc] to examine the effect of having a single Gp^{Me}C step.

4.2.2 MD simulation

We used CHARMM version 40a1 [21] with the param36 all-atom force field [64] and the TIP3P model [87]. Each DNA oligo was solvated in a cubic water box of side length about 84 Å to make the oligo at least 15 Å away from the boundary in all directions, which is larger than the 12-Å cutoff for nonbonded interactions. Sodium ions were added to neutralize the system, resulting in about 90 mM concentration [155, 142, 156]. The electrostatically neutral system was subjected to the initial energy minimization (600 steps of the steepest descent method followed by 1000 steps of the adapted basis Newton-Raphson method). During energy minimization, heavy atoms of DNA were harmonically restrained

Name	Sequence	Steps Studied
[AAAcg]	d(CGCGAAAcgTTTCGCG)· d(CGCGAAAcgTTTCGCG)	Ac/gT, cg/cg
[TTTcg]	d(CGCGTTTcgAAACGCG)· d(CGCGTTTcgAAACGCG)	Tc/gA, cg/cg
$[cg]_8$	d(cgcgcgcgcgcgcgcg)· d(cgcgcgcgcgcgcgcg)	cg/cg, gc/gc
[CGcg]	d(CGCGCGCGcgCGCGCG)· d(CGCGCGcgCGCGCGCG)	Gc/gC, cg/cg
$[cg]_6$	d(cgcgcgcgcgcg)· d(cgcgcgcgcgcg)	gc/gc
[GCgc]	d(CGCGCGCgcGCGCGCG)· d(CGCGCGCgcGCGCGCG)	gc/gc

Table 4.1: Names and sequences of DNA oligos used in simulation. Lower case c and g mean mCYT and the complementary GUA, respectively.

with a spring constant of 2 kcal/(mol·Å^2). This constraint was reduced to 1 kcal/(mol·Å^2) during the 30-ps heating (0 K to 300 K), and then to 0.2 kcal/(mol·Å^2) during the 70-ps equilibration runs. Heating and equilibration were done using the constant pressure (1 atm) and temperature (CPT) method [45]. Production runs were at 300 K under constant volume (NVT) without any restraint on DNA. They lasted 100 ns except for [GCgc] that run for 50 ns. We used the SHAKE [176] algorithm to fix the length of covalent bonds for hydrogen atoms, and used a 2-fs integration time step. The particle-mesh Ewald (PME) summation method was used to account for long-range electrostatic interactions [44]. Simulation systems typically had about 61,000 atoms, and the domain decomposition (DOMDEC) module of CHARMM was used to achieve efficient parallelization of simulation [76]. Coordinates were saved every 10 ps (10^4 coordinate frames for 100 ns). Analysis of the coordinate trajectory was done mostly for the last 50 ns (5000 frames). For [GCgc], the last 25 ns was used (2500 frames). Only the middle 8 base pairs of oligos were used for analysis, to avoid end effects.

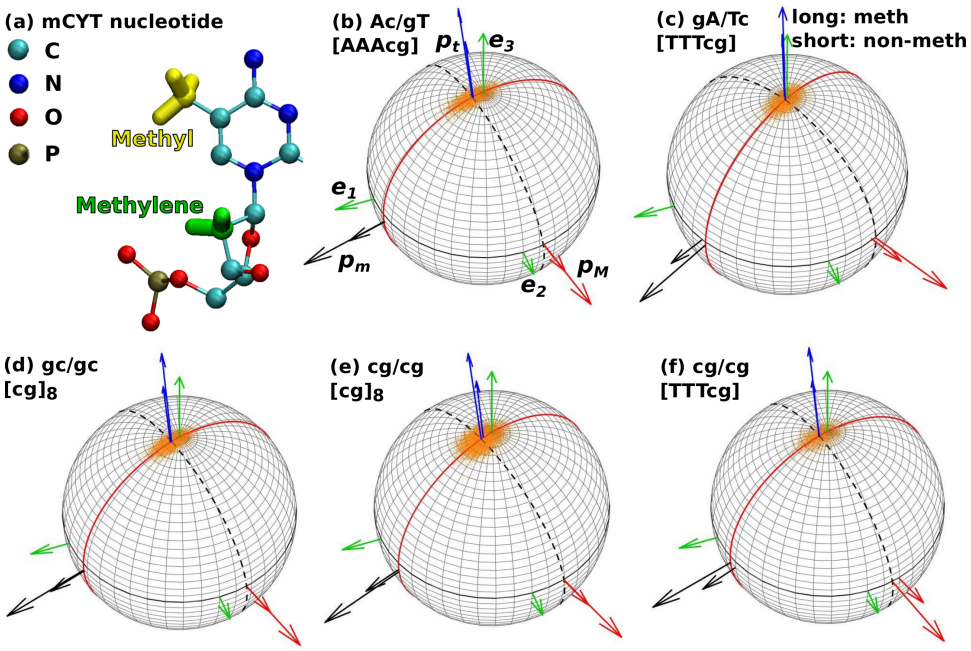

Figure 4.1: Effect of cytosine methylation on dinucleotide step motion. (a) Structure of a methylated cytosine. Yellow: methyl group. Green: methylene group of the deoxyribose ring. (b-f) Principal axis-based analysis of dinucleotide step motion (see Methods). $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$: Reference triad assigned to the first base pair of a step. For example, for (b), it is assigned to the A-T pair. $\{\boldsymbol{p_m}, \boldsymbol{p_M}, \boldsymbol{p_t}\}$: equilibrium triad for the next base pair in a dinucleotide step ($\boldsymbol{p_m}$/$\boldsymbol{p_M}$: minor/major principal axes). Long/short arrows: equilibrium triads for methylated/un-methylated steps. Name of the oligo from which the analysis was performed is indicated in each panel. Equilibrium triads for the cg/cg steps in [CGcg] and [AAAcg] are very similar to that in $[cg]_8$ (panel (e)), and the equilibrium triad for the Gc/gC step of [CGcg] is also similar to that for gc/gc in $[cg]_8$, hence they are not shown. Refer to Table 4.2 for detail values. Changes in principal axes and the equilibrium axial vector $\boldsymbol{p_t}$ are minimal except for [TTTcg].

Calculation of standard helicoidal parameters, such as roll, tilt, and helicoidal twist, was done using X3DNA [119]. Molecular structures in figures were rendered mostly using VMD [74]. For Fig. 4.12, UCSF Chimera [158] was used to generate the water density map stored as the electron microscopy mrc format files.

Step (name)	s	$\boldsymbol{p_M}$	$\boldsymbol{p_m}$	θ_t	$\sigma(s)$	$\sigma(\theta_M)$	$\sigma(\theta_m)$
Ac/gT [AAAcg]	3.6184	(-0.1878,0.9821,0.0157)	(0.9767,0.1885,-0.1023)	19.44±4.50	0.32	6.22	2.42
Gc/gC [CGcg]	3.5245	(-0.2145,0.9767,0.0076)	(0.9727,0.2143,-0.0891)	17.82±6.26	0.30	5.29	2.39
gA/Tc [TTTcg]	3.7112	(-0.4580,0.8889,-0.0103)	(0.8885,0.4574,-0.0369)	6.04±5.18	0.35	5.46	2.86
gc/gc $[cg]_8$	3.5599	(-0.2572,0.9664,-0.0009)	(0.9630,0.2563,-0.0830)	14.52±5.33	0.28	4.92	2.18
cg/cg [AAAcg]	3.3760	(-0.2946,0.9556,0.0001)	(0.9524,0.2936,-0.0825)	22.28±4.75	0.40	6.11	3.47
cg/cg [CGcg]	3.4435	(-0.3276,0.9448,0.0028)	(0.9410,0.3265,-0.0893)	19.22±5.39	0.37	5.82	3.15
cg/cg [TTTcg]	3.7837	(-0.2350,0.9720,-0.0005)	(0.9685,0.2341,-0.0853)	18.49±6.34	0.40	6.32	3.70
cg/cg $[cg]_8$	3.3268	(-0.3032,0.9529,-0.0028)	(0.9467,0.3009,-0.1150)	21.43±4.83	0.37	5.70	3.61

Table 4.2: Equilibrium conformations of dinucleotide steps regarding methylation. s: average distance between centroids (Å); $\boldsymbol{p_M}$ and $\boldsymbol{p_m}$: Coordinates of the major and minor principal axes relative to the triad of the reference base pair. θ_t: Twist angle (degrees). $\sigma(s)$: standard deviation of the distance between centroids (Å). $\sigma(\theta_M)$ and $\sigma(\theta_m)$: standard deviations in the major and minor bending angles (degrees).

4.2.3 Principal axis-based analysis of DNA mechanics

Details of this method is explained in Ref. [212]. Briefly, local triads (orthogonal coordinate bases) are assigned to individual base pairs. Denote the triads for the two base pairs forming a dinucleotide step as $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ and $\{\boldsymbol{e_1'}, \boldsymbol{e_2'}, \boldsymbol{e_3'}\}$, respectively. The trajectory of $\boldsymbol{e_3'}$ in the space spanned by $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ forms an ellipsoidal set of dots (orange dots in Fig. 4.1). Its centroid vector $\boldsymbol{p_t}$ corresponds to the equilibrium bending direction of the dinucleotide step. The long and short axes of the ellipsoid form the major and minor bending directions (red and black dashed circles in Fig. 4.1). Axes perpendicular to these directions are respectively the major and minor principal axes of bending, $\boldsymbol{p_M}$ and $\boldsymbol{p_m}$. The set $\{\boldsymbol{p_m}, \boldsymbol{p_M}, \boldsymbol{p_t}\}$ forms a right-handed orthogonal system that we call as the equilibrium triad. The orientation of the equilibrium triad relative to $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ represents the equilibrium conformation of a given dinucleotide step. Note that, unlike helicoidal parameters that are based on static DNA structures, the principal axes are determined by local bending motion, and indicate the most/least flexible bending directions. The ma-

jor/minor bending angle of a step at each time frame can be measured as the the projection angles of $\boldsymbol{e}_3'$ onto the major and minor directions (circles in Fig. 4.1). The reference direction for the bending angle measurement was $\boldsymbol{p}_t$, so that a zero bending angle means that the dinucleotide step takes an equilibrium curvature at a given time frame. The twist angle of a step at each frame was obtained by measuring the Euler angle between the current triad ($\{\boldsymbol{e}_1', \boldsymbol{e}_2', \boldsymbol{e}_3'\}$ expressed in the basis of $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$) and its equilibrium triad $\{\boldsymbol{p}_m, \boldsymbol{p}_M, \boldsymbol{p}_t\}$ in axial direction. Extension of a dinucleotide step was calculated from the distance between the centroids of neighboring triads. The stiffness of each deformational mode was calculated by applying the equipartition theorem [167].

4.2.4 Water density map

Average distribution of water near DNA was obtained by using the solvation map calculation method that we previously developed [164]. Briefly, the space within 15 Å from a given set of atoms was divided into cubic cells of size 0.7 Å. The fraction of coordinate frames that a water oxygen atom visits this cell divided by the volume of the cell (0.7^3 Å^3) yields the local water density. Since the oligo moves in space, coordinate frames were aligned relative to the given set of atoms. Thus the density map accounts for surface water molecules moving with the given atoms. In Fig. 4.12, regions with water density 0.054 A^{-3} (1.5 times higher than the bulk density) are displayed.

4.3 Results and Discussion

4.3.1 Effect of methylation on the stiffness and equilibrium conformations of dinucleotide steps

For most dinucleotide steps, the equilibrium triad $\{\boldsymbol{p}_m, \boldsymbol{p}_M, \boldsymbol{p}_t\}$ changes its orientation very little upon methylation, suggesting that the equilibrium conformation of DNA is largely unaffected by methylation. [TTTcg] has the largest discrepancy in methylated and un-methylated equilibrium triads (Fig. 4.1c,f), whose structural origin is discussed below.

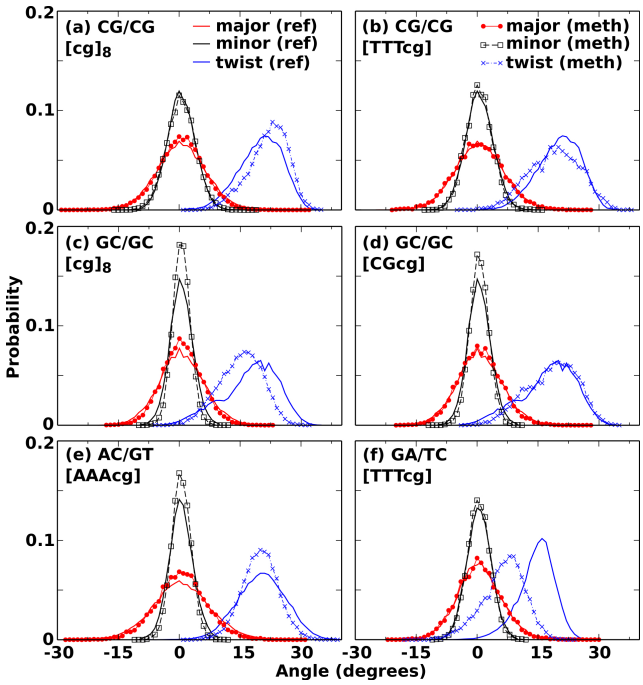

Figure 4.2: Distributions of angles with and without methylation (lines with and without symbols, respectively). References for the major and minor bending angles (along the red solid and black dashed circles in Fig. 4.1) are the equilibrium direction $\boldsymbol{p}_t$, so that their distributions are peaked at zero degrees.

Similar to the un-methylated case [212], distributions of the major and minor bending angles for the methylated cases are Gaussian, which suggests that their bending motion can be described by linear elasticity. However, the twist distribution is more strongly affected in several cases (Fig. 4.2c,e,f).

We calculated the stiffness associated with major/minor bending and twist, and compared with those for the un-methylated ones from our previous study [212] (Fig. 4.3, values are in Table 4.3). Dinucleotide steps that are away from the methylation site have stiffness that are overall similar to those from our earlier study for un-methylated oligos, though twist stiffness varies more (yellow region in Fig. 4.3a,b,d). Similarly, for the $^{\mathrm{Me}}$CpG step, both major (κ_M) and minor (κ_m) bending stiffness seldom changes upon methylation (star

Step (name)	κ_M ($\times 10^4$)	κ_m ($\times 10^4$)	κ_t ($\times 10^4$)	κ_E
Ac/gT [AAAcg]	1.27 (0.84)	8.40 (5.44)	2.43 (1.38)	416 (303)
Gc/gC [CGcg]	1.71 (1.47)	8.38 (5.84)	1.22 (1.02)	450 (384)
gA/Tc [TTTcg]	1.69 (1.52)	6.15 (4.94)	1.88 (2.30)	338 (360)
gc/gc $[cg]_8$	2.00 (1.47)	10.14 (5.84)	1.70 (1.02)	521 (384)
cg/cg [AAAcg]	1.23 (1.24)	3.82 (3.95)	2.04 (1.66)	262 (308)
cg/cg [CGcg]	1.29 (1.24)	3.76 (3.95)	1.61 (1.66)	301 (308)
cg/cg [TTTcg]	1.38 (1.24)	4.73 (3.95)	1.28 (1.66)	254 (308)
cg/cg $[cg]_8$	1.39 (1.24)	3.48 (3.95)	1.94 (1.66)	305 (308)

Table 4.3: Stiffness of steps related to methylation. κ_M, κ_m, and κ_t are in units of pN·Å^2, and κ_E in pN/Å. Reference values from previous study are given in parenthesis [212].

in Fig. 4.3). The largest deviation is for κ_m in [TTTcg], increasing by 20% (Fig. 4.3b). In comparison, a previous study reported that the roll and tilt stiffness (helicoidal parameters) increase for the $^{\mathrm{Me}}$CpG step [155], which we did not observe in our calculation (Fig. 4.4).

More prominent increase in bending stiffness is observed for steps neighboring the $^{\mathrm{Me}}$CpG step (marked 'N' in Fig. 4.3). The major bending stiffness increases up to 56% ([AAAcg]), and the minor bending stiffness increases up to 74% ($[cg]_8$). The twist stiffness κ_t also increases, except that a consistent decrease was observed for [TTTcg], where κ_t for the $^{\mathrm{Me}}$CpG step itself was also lower (Fig. 4.3b).

Taken together, methylation generally makes the neighboring steps stiffer rather than the methylated step itself, and twist becomes either stiffer or more flexible, depending on the sequence. The structural origin for this behavior is explained below. It is also worth noting that, as the density of methylation increases (Fig. 4.3c,d; $[cg]_8$ vs [CGcg]), the stiffness is further enhanced. This indicates that hyper-methylation of the CpG island greatly alters the mechanical property of DNA, which may have implications in processes such as chromosome inactivation [153, 207], cancer pathology [82, 84, 10], and the requirement

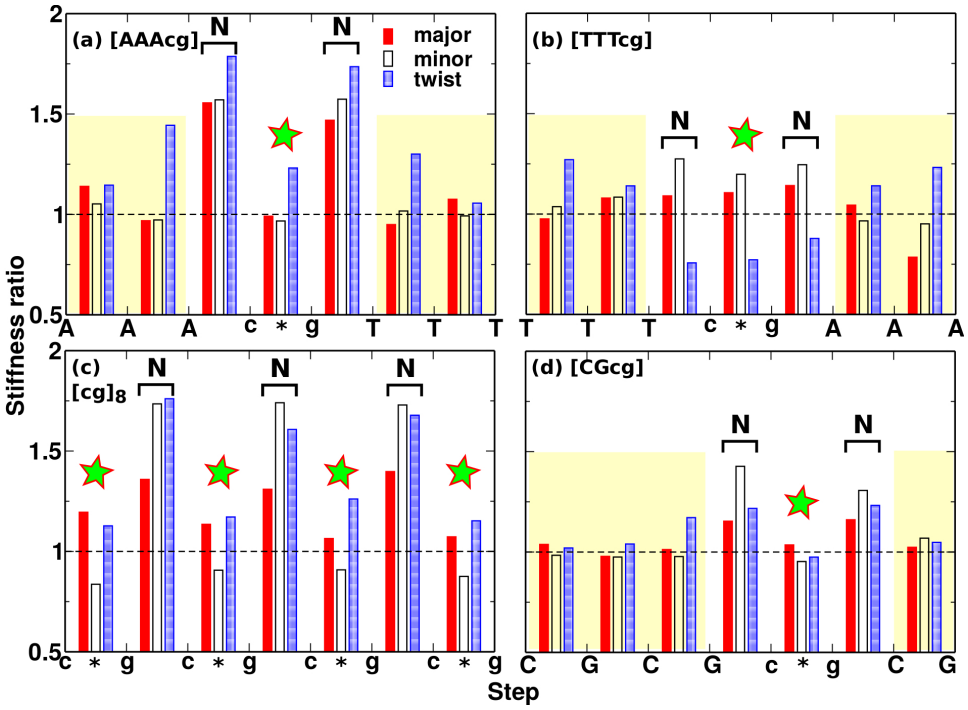

Figure 4.3: The stiffness of dinucleotide steps divided by the values from their un-methylated counterpart from our previous study [212]. The ^{Me}CpG step is marked by a star. The neighboring steps are marked by 'N,' and the un-methylated regions are in yellow background.

of 'methylation threshold' for inactivating certain genes [71].

For comparison, we measured the helicoidal parameters that describe the DNA structure closer to the atomistic level (Table 4.4). Overall, greater changes upon methylation were observed for [AAAcg] and [TTTcg], suggesting that they underwent greater deformation than other oligos tested. We used fluctuations of roll, tilt, and helicoidal twist, to calculate the corresponding stiffness for each step (Fig. 4.4). These three helicoidal parameters correspond approximately to the two bending angles and twist in our principal axis-based approach. Although their stiffness are also somewhat enhanced for steps neighboring the methylated step, the extent of change is much smaller than the stiffness calculated based on our approach (Fig. 4.3 vs. 4.4). This shows that our analysis based on

Step [oligo]	Shift	Slide	Rise
Ac/gT [AAAcg]	0.0059±0.6733(0.4006±0.6864)	-0.0570±0.5573(-0.0399±0.6650)	3.2076±0.3524(3.2322±0.3869)
gA/Tc [TTTcg]	-0.0196±0.7048(-0.3509±0.5559)	0.5219±0.6041(0.3808±0.5193)	3.2803±0.3312(3.3775±0.3094)
gc/gc $[cg]_8$	-0.0152±0.5017(-0.0261±0.6787)	-0.1740±0.4050(-0.0720±0.4672)	3.1380±0.3351(3.2504±0.3569)
Gc/gC [CGcg]	-0.0300±0.6080(-0.0261±0.6787)	-0.1338±0.4243(-0.0720±0.4672)	3.1996±0.3481(3.2504±0.3569)
cg/cg [AAAcg]	0.0621±0.6200(0.0101±0.6474)	0.7139±0.5599(0.3575±0.5470)	3.6309±0.3789(3.4869±0.3485)
cg/cg $[cg]_8$	0.0000±0.6385(0.0101±0.6474)	0.4201±0.5378(0.3575±0.5470)	3.5867±0.3492(3.4869±0.3485)
cg/cg [CGcg]	0.0084±0.6498(0.0101±0.6474)	0.5230±0.5326(0.3575±0.5470)	3.5277±0.3454(3.4869±0.3485)
cg/cg [TTTcg]	-0.0108±0.6597(0.0101±0.6474)	0.8567±0.7005(0.3575±0.5470)	3.3632±0.3540(3.4869±0.3485)
Step [oligo]	Tilt	Roll	Helicoidal Twist
Ac/gT [AAAcg]	-0.2922±4.6122(0.6023±5.0852)	2.8108±5.1950(3.2681±6.6852)	31.200±4.870(31.575±6.170)
gA/Tc [TTTcg]	-0.0821±5.0135(-0.8318±4.6911)	5.4023±5.6451(1.3880±5.8302)	34.095±5.536(40.030±4.720)
gc/gc $[cg]_8$	-0.0740±4.2847(-0.1138±5.1440)	2.3072±4.8469(2.8347±5.7120)	31.790±5.519(34.240±6.803)
Gc/gC [CGcg]	-0.2540±4.8264(-0.1138±5.1440)	2.6504±5.2308(2.8347±5.7120)	32.648±6.314(34.240±6.803)
cg/cg [AAAcg]	0.6715±6.2185(0.1416±5.8246)	8.4821±7.4975(8.7749±7.3696)	36.767±4.966(34.317±5.391)
cg/cg $[cg]_8$	0.0619±5.8176(0.1416±5.8246)	10.591±6.800(8.7749±7.3696)	36.388±4.992(34.317±5.391)
cg/cg [CGcg]	-0.0410±5.6787(0.1416±5.8246)	9.1531±7.2039(8.7749±7.3696)	35.526±5.523(34.317±5.391)
cg/cg [TTTcg]	0.0743±5.9956(0.1416±5.8246)	10.369±7.197(8.7749±7.3696)	29.330±6.285(34.317±5.391)

Table 4.4: Average helicoidal parameters. Values for the un-methylated reference (in parenthesis) are from Ref. [212].

orthogonal order parameters yields more sensitive measure of the changes in mechanical properties of DNA.

The extensional stiffness (κ_E) ratio compared with un-methylated is given in the Fig. 4.5. In [AAAcg], the $^{\mathrm{Me}}$CpG step's κ_E decreases, and that of all other steps increases. For [TTTcg], κ_E of $^{\mathrm{Me}}$CpG and its neighbors decreases, with an increase in the region beyond. In [CGcg] and $[cg]_8$, κ_E of $^{\mathrm{Me}}$CpG step changes little, with neighbor increased but beyond region unaffected. This trend is quite similar to the twist, which could be due to a coupling between twist and extension [124, 54].

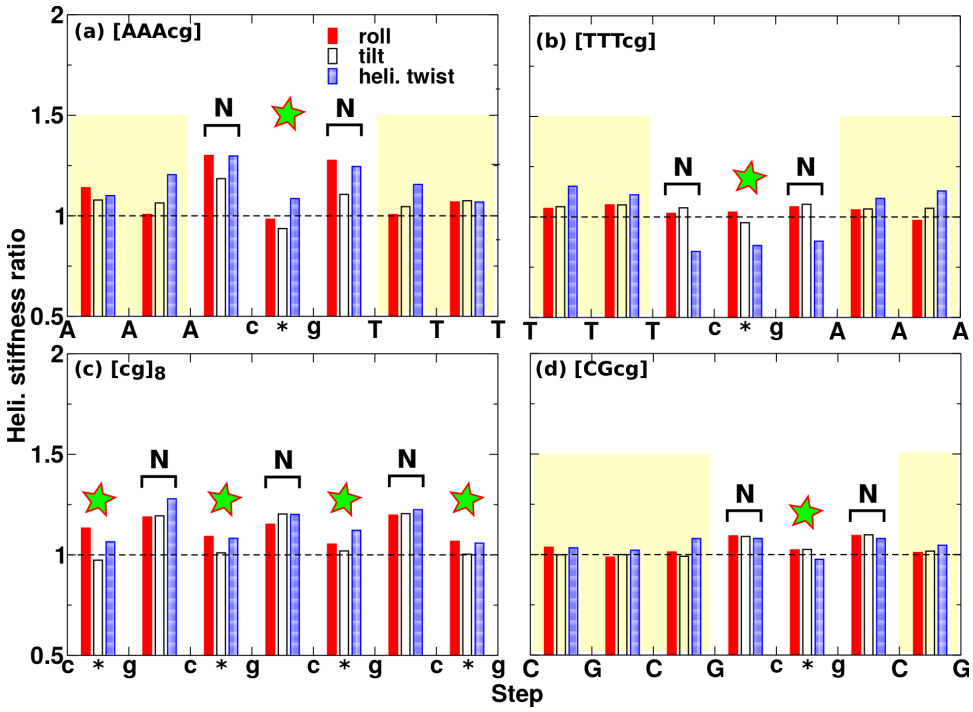

Figure 4.4: Stiffness ratio of helicoidal roll, tilt, and twist for dinucleotide steps measured between methylated and un-methylated oligos. The same data used for Fig. 4.3 were used for analysis. Note that the vertical axis has the same range as in Fig. 4.3.

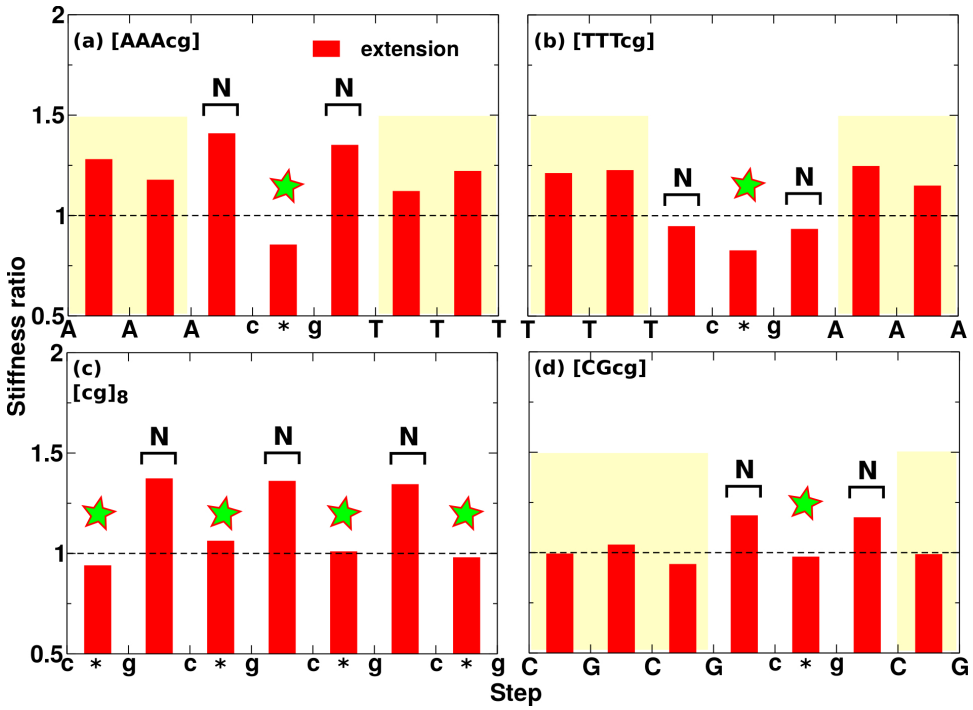

Figure 4.5: Extensional stiffness ratio, methylated over un-methylated. The trend shows some similarity to twist.

4.3.2 Structural basis for the altered stiffness

The sequence-dependent changes in the stiffness can be explained mostly in terms of the interaction between the methyl group of mCYT and other groups of the neighboring nucleotide, in particular the C2' methylene of the deoxyribose ring (Fig. 4.6a). Due to the right-handed structure of B-DNA, nucleotide on the 5' side of ^{Me}CpG may interact with the mCYT-methyl group. We first consider bending stiffness of steps that do not involve THY, whose methyl group introduces additional interaction. The steric repulsion between the mCYT-methyl group and the methylene group of neighboring bases suppresses both major and minor bending motions (Fig. 4.6a,b). This leads to an increase in bending stiffness of the steps neighboring ^{Me}CpG. For the ^{Me}CpG step itself, the methyl groups do not interact and its bending motion is unaffected (Fig. 4.6b; Fig. 4.3a,b,d). Since the steric

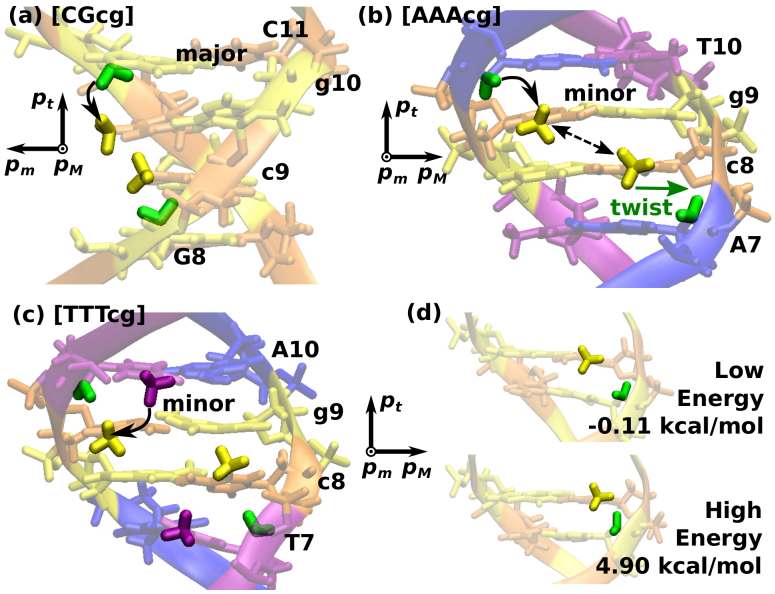

Figure 4.6: Interaction between the mCYT-methyl group and adjacent nonpolar groups (arrows with solid line). Yellow: mCYT-methyl, green: C2' methylene, purple: THY-methyl groups. (a,b) Examples of the steric interaction between mCYT-methyl and C2' methylene groups, leading to the suppression of bending. The equilibrium triad for the methylated step is shown in each panel.(a) [CGcg] viewed along $\boldsymbol{p}_M$. Curved arrow: steric clash between mCYT-methyl and C2' methylene groups upon major bending. (b) [AAAcg] viewed along $\boldsymbol{p}_m$. Curved (solid) arrow: steric clash of the methyl groups upon minor bending (twist). Dashed arrow denotes lack of interaction between methyl groups within a $^{\mathrm{Me}}$CpG step. (c) [TTTcg]. Curved arrow: steric clash between methyl groups of mCYT and the neighboring THY upon minor bending. Similar effects are present for major bending, but the minor bending stiffness is affected more than the major bending stiffness does (Fig. 4.3b). (d) Snapshots of conformations with high and low interaction energy between methyl and methylene groups in [CGcg].

effect occurs only upon bending, the equilibrium conformation of the steps are largely unaffected by methylation (Fig. 4.1).

In [TTTcg], the methyl group of THY that is bulkier and more protruding compared to the C2' methylene, has a greater influence. The equilibrium conformation of the dinucleotide steps differ more significantly compared to steps without THY on the 5' side (Fig. 4.1c,f). The THY-methyl group may also restrict the bending motion, for which the

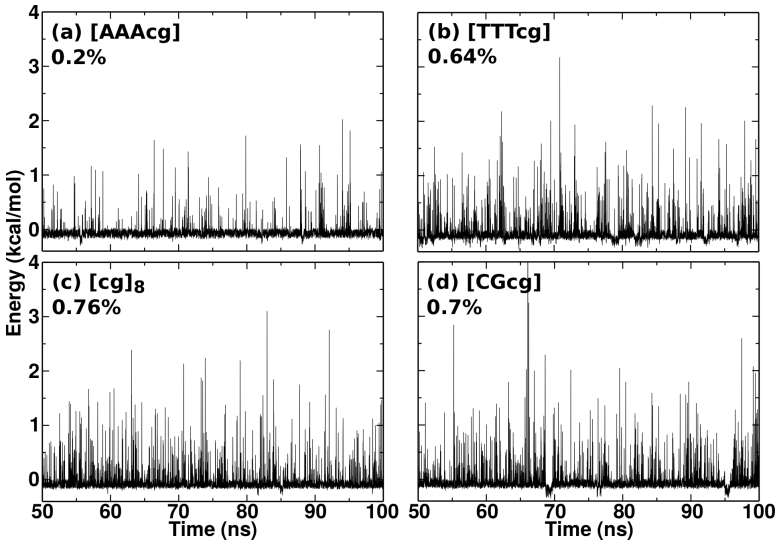

Figure 4.7: Interaction energy (electrostatic and van der Waals) between one mCYT methyl group and its neighbor C2' methylene group (Fig. 4.6). In $[cg]_8$, it is calculated for a single pair of groups in the middle of the oligo. Percentage of frames where the interaction energy exceeds 1.2 kcal/mol (two times the thermal energy at 300 K) are indicated in each panel. These are cases when steric repulsion occurs via large deformation (*cf.*, Fig. 4.6d).

minor bending appears to be more affected than the major bending (rotation about $\boldsymbol{p_m}$ in Fig. 4.6c), which results in the increase of κ_m in both $^{\mathrm{Me}}$CpG and its neighboring steps of [TTTcg] (Fig. 4.3b). The steric repulsion effect can be seen by the occasional presence of large positive interaction energy (Figs. 4.7 and 4.8). Whereas the mCYT-methyl and C2'-methylene groups also have negative interaction energies (Fig. 4.7), mCYT-methyl and THY-methyl groups always have positive interaction energies (Fig. 4.8), which supports that the presence of THY on the 5' side leads to the deformation of the step (Fig. 4.1c,f).

Twist stiffness κ_t shows more complex sequence dependence. For [AAAcg], $[cg]_8$, and [CGcg], the increase in κ_t for the neighboring steps of $^{\mathrm{Me}}$CpG (Fig. 4.3a,c,d) is again due to the steric repulsion with the methylene group (Fig. 4.6b). However, for [TTTcg], κ_t decreases for both $^{\mathrm{Me}}$CpG and its neighboring steps. In the former case, since the equilib-

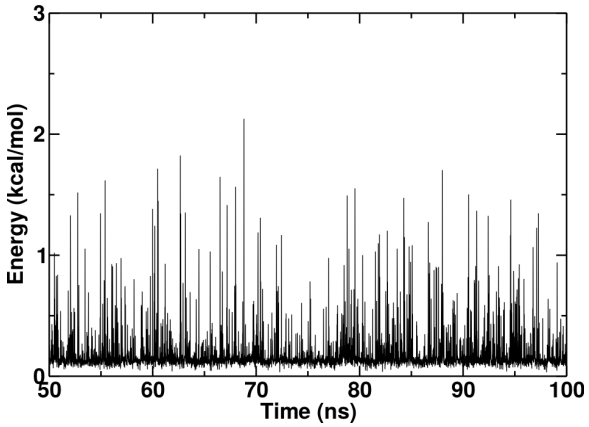

Figure 4.8: Interaction energy between methyl groups of THY and mCYT. This energy is always positive, indicating its repelling nature.

rium conformation is unaffected, the interaction with the neighboring C2' methylene group may only restrict conformational motion including twist. In the latter case, the steric clash between the two methyl groups leads to distortion of the structure, to which the twisting motion may have nontrivial dependence. The distortion can be gleaned from changes in the equilibrium triad (Fig. 4.1c,f) and also in helicoidal parameters (Table 4.4). Since the methyl groups in [TTTcg] line up along the major groove, suppression of bending motion is expected (Fig. 4.9). This lining up of nonpolar methyl groups also alter the hydration shell along the major groove, which likely contributes to its reduced twist stiffness (see below).

Although steric interactions involving mCYT-methyl with C2' methylene and THY-methyl groups are the basis for the overall increase in the stiffness of steps neighboring a methylated step, there are variations for which structural mechanism is unclear. As noted above, the twist stiffness of the neighboring steps in [TTTcg] decreases. In contrast, it increases for the AA/TT step that is next neighbor to ^{Me}CpG in [AAAcg] (Fig. 4.6a; Fig. 4.10). Increased presence of nonpolar methyl groups in THY and mCYT affect local hydration structure, which may contribute to enhancing the next-neighbor effect [154].

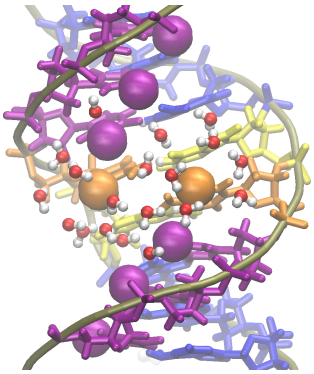

Figure 4.9: Water molecules around ^{Me}CpG of [TTTcg] on the major groove side. Methyl groups next to ^{Me}CpG squeeze the water distribution and hence the elongated water blob in other oligos disappears in this oligo.

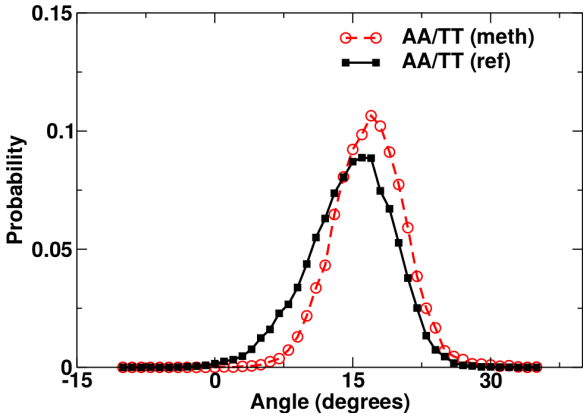

Figure 4.10: Twist distribution of AA/TT step in next neighbor of ^{Me}CpG of [AAAcg] and reference. The distribution becomes narrower and shifted.

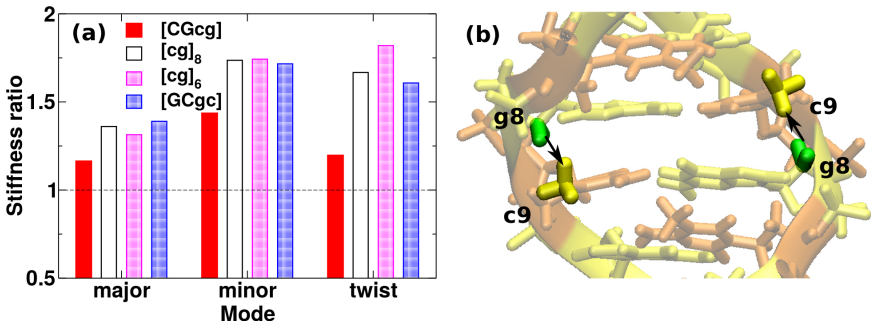

Figure 4.11: (a) (methylated) GC/GC stiffness ratio comparison of four oligos: [CGcg], $[cg]_8$, $[cg]_6$, and [GCgc]. Oligo $[cg]_6$ has six repeated ^{Me}CpG steps, and [GCgc] has one non-physiological Gp^{Me}C step. (b) Illustration of interaction in Gp^{Me}C step. There are repulsion from both 5'-sides.

Another intriguing aspect is the greater increase in stiffness for $[cg]_8$, which has consecutive ^{Me}CpG steps. A possible reason for this increase is suppression of local deformation by the viscous drag acting on the oligo, which may increase as the oligo becomes stiffer and longer. In a limiting case, if only a single dinucleotide step in the middle can deform with other parts of the oligo being rigid, deformation of the step would require a rigid-body motion of the rest of the oligo. Viscous drag on the oligo will suppress conformational fluctuation of the step, resulting in a higher apparent stiffness of the step. In this scenario, the apparent stiffness is expected to increase with the length of the oligo, since it will experience greater viscous drag. However, stiffness measured for $[cg]_6$ were nearly the same as those for $[cg]_8$ (Fig. 4.11a). We instead found the stiffness enhancement to be a more local effect: The two mCYT within a gc/gc step have steric interaction with C2' methylene groups on both DNA strands (Fig. 4.6b vs. Fig. 4.11b). To further test this, we used [GCgc] that contains a Gp^{Me}C step (not physiologically relevant [62, 186]). Despite having a single Gp^{Me}C step, its stiffness are similar to those of $[cg]_8$ (Fig. 4.11a), which confirms that the stiffness enhancement is mainly due to interaction between two consecutive methylated steps.

4.3.3 Methylation-induced changes in surface water structure

In addition to intra-DNA interactions, surface water molecules that form hydrogen bonds with DNA and among themselves, may influence the conformational motion of DNA. While ions also form contacts with DNA and affect its conformation and dynamics, in our simulation, only monovalent ions were used, which had at most 7% contact occupancy with DNA, thus they were not considered. We calculated the water density map for individual ^{Me}CpG steps and its neighbor base pairs (Fig. 4.12).

There is a concentrated hydration shell at the center of ^{Me}CpG step on the major groove side in the 4 methylated oligos, whereas in the reference case without methylation (Fig. 4.12e), the hydration distribution at the same position is discrete. This finding supports methylation can bring about more connected water interaction, and hence makes DNA stiffer. Within the 4 methylated oligos, this hydration shell is elongated except for [TTTcg] (dark red in Fig. 4.12b), which is mainly due to the restraint of methyl groups of THY on 5'-side repelling the water molecules. This comparably smaller hydration shell has less restriction on twist, and consequently results in the flexible twist motion of ^{Me}CpG and neighbor steps in [TTTcg]. It can be found water density near GUA is bulky in all the methylated oligos. Since hydration shell also forms near non-polar groups [164], these densities arise as a part of the hydration shell surrounding the mCYT methyl group (yellow in Fig. 4.12).

On the minor groove side the water distribution again shows sequence-dependence. In [AAAcg], there is a distinctive doughnut shape water layer in the middle, probably due to the narrow groove that forces water molecules to connect. For the other 4 oligos (including reference), there are two hydration shells parallel along the backbone of DNA, and among them, the reference one has the smallest shell, indicating the effect of methylation even the modification is not on the minor groove side.

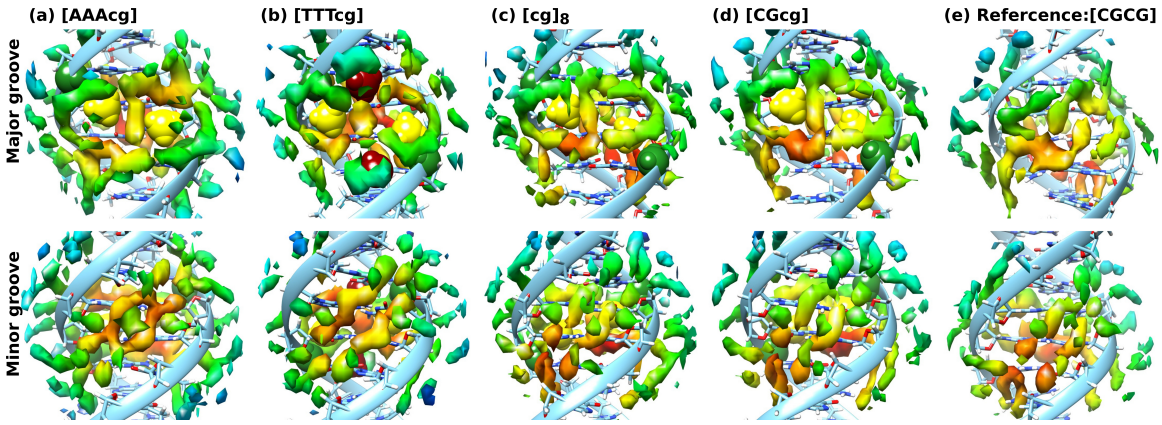

Figure 4.12: Water density map around ^{Me}CpG of (a) [AAAcg], (b) [TTTcg], (c) $[cg]_8$, (d) [CGcg], (e) reference CG/CG step. The bases used to orient are ^{Me}CpG or CG/CG and its immediate neighbor base pairs. Methyl groups from ^{Me}CpG are colored in yellow, methylene prior to mCYT in green, and methyl groups of THY in [TTTcg] in dark red.

It is known hydration force is a fatal factor in biomolecular interaction [201, 26, 164]. Now that water distribution around methylated step changes, it may favor different proteins. This binding affinity change is one of the main hypotheses to explain methylation induced genomic regulation [220, 208, 72, 18, 136, 86, 4, 202]. The hydration distribution varies by virtue of methylation found in this study can provide a substantial support to this explanation.

4.4 Conclusions

In this study, we found the methylation of CpG can generally enhance the stiffness of neighbor dinucleotide steps, with the ^{Me}CpG step itself less affected. This is by reason of the interaction of methyl group with hydrogen atoms on the 5'-side deoxyribose. If the 5'-side base of mCYT is THY, the twist stiffness around would decrease. High density of methylation can further increase the stiffness, and may relate to diseases development. The hydration distribution is affected upon methylation, which could influence the stiffness, and may favor different proteins to be bound and therefore affect the initiation of

transcription and realize the gene regulation.

5. CONCLUSION

Our analysis shows collagen cleavage site indeed has a transition in bending stiffness, from a rigid region to a flexible region. One of the heterotrimer isomers behaves more dynamic than the other two, making it readily cleavable and the most possible candidate present *in vivo*. We demonstrated mechanical environment acts as a central role in regulating the conformation and vulnerability to hydrolysis of collagen monomer. This finding resolves controversial results from collagen extension experiments, indicating the validity of our study. With the knowledge on the relationship between mechanical environment and collagen cleavage, researchers will have a deeper insight into tissue turnover and remodeling. The homotrimer's resistance is because of the Arg residues downstream to the cleavage site, which can stabilize the triple-helical structure. This resistance mechanism could help people understanding homotrimer's role in fetal development and cancer progress.

Limitation of this work is it doesn't include the effect of MMPs. Therefore, simulation with MMPs could be a future direction. Currently there is a crystal structure available [123] that can serve as a good starting point. Key interactions between collagen and MMP-1 can be identified. Also the reaction coordinate from collagen unwinding to the state of one collagen chain located at the active site of MMP-1 will be appealing, since we can understand the atomistic details of the process of cleavage. Further work can consider the role of tension applied on collagen during this process of conformational change.

DNA study shows sequence-dependent mechanical properties of DNA, which will be much useful in DNA packaging and binding studies, as well as DNA engineering (DNA origami). Results indicate that our principal-axes based analysis is more appropriate to describe DNA as an elastic rod, which makes the construction of DNA mesoscale model

possible. Traditionally used helicoidal parameters set is degenerate and not suitable for such a purpose. The coarse-grained model built based on our data replicates the experimental data pretty well. We can further applied it to the DNA looping and packaging studies. This newly developed method can be applied to any filamentous structure.

We also investigated the effect of methylation. We found the pattern of stiffness change and hydration distribution variation. These findings can provide more specific information on the DNA behavior changes after decoration, and could ultimately aid in elucidating the mechanism of its role in gene regulation.

A future work could be studying influence of other epigenetic modifications, such as oxidation of mCYT. It is also worth simulating DNA-protein complexes before and after methylation, to quantify the changes interaction energy and hydration shell. Histones and promoters are the most relevant proteins to look into.

Our current project is focused on kinesin and microtubule (MT) interaction. By applying force at different pulling direction of kinesin, at different adenosine binding states, we will identify the key residues in the pulling process as well as how the conformational change upon different adenosine binding affects the affinity between kinesin and MT. This will be helpful to understand the walking cycle of kinesin on MT.

REFERENCES

[1] A. S. Adhikari, J. Chai, and A. R. Dunn. Mechanical Load Induces a 100-Fold Increase in the Rate of Collagen Proteolysis by MMP-1. *J. Am. Chem. Soc.*, 133:1686–1689, 2011.

[2] A. S. Adhikari, E. Glassey, and A. R. Dunn. Conformational dynamics accompanying the proteolytic degradation of trimeric collagen I by collagenases. *J. Am. Chem. Soc.*, 134(32):13259–13265, 2012.

[3] D. Anderson. Collagen self-assembly: A complementary experimental and theoretical perspective. *PhD Thesis, University of Toronto*, 2005.

[4] E. Ballestar and A. P. Wolffe. Methyl-cpg-binding proteins. *Eur. J. Biochem.*, 268(1):1–6, 2001.

[5] N. K. Banavali and B. Roux. Free energy landscape of a-dna to b-dna conversion in aqueous solution. *J. Am. Chem. Soc.*, 127(18):6866–6876, 2005.

[6] C. Bartels and M. Karplus. Multidimensional adaptive umbrella sampling: applications to main chain and side chain peptide conformations. *J. Comp. Chem.*, 18(12):1450–1462, 1997.

[7] F. Battistini, C. A. Hunter, E. J. Gardiner, and M. J. Packer. Structural mechanics of dna wrapping in the nucleosome. *J. Mol. Biol.*, 396(2):264–279, 2010.

[8] C. G. Baumann, S. B. Smith, V. A. Bloomfield, and C. Bustamante. Ionic effects on the elasticity of single dna molecules. *Proc. Natl. Acad. Sci. USA*, 94(12):6185–6190, 1997.

[9] J. Bella, M. Eaton, B. Brodsky, and H. M. Berman. Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science*, 266:75–81, 1994.

[10] Y. Bergman and H. Cedar. Dna methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.*, 20(3):274–281, 2013.

[11] I. Bertini, M. Fragai, C. Luchinat, M. Melikian, M. Toccafondi, J. L. Lauer, and G. B. Fields. Structural basis for matrix metalloproteinase 1-catalyzed collagenolysis. *J. Am. Chem. Soc.*, 134:2100–2110, 2011.

[12] D. L. Beveridge, T. E. Cheatham III, and M. Mezei. The abcs of molecular dynamics simulations on b-dna, circa 2012. *J. Biosci.*, 37(3):379–397, 2012.

[13] A. P. Bhole, B. P. Flynn, M. Liles, N. Saeidi, C. A. Dimarzio, and J. W. Ruberti. Mechanical strain enhances survivability of collagen micronetworks in the presence of collagenase: implications for load-bearing matrix growth and stability. *Philos. Trans. R. Soc.*, 367(1902):3339–3362, 2009.

[14] A. Bird. Dna methylation patterns and epigenetic memory. *Gene. Dev.*, 16(1):6–21, 2002.

[15] A. P. Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067):209–213, 1985.

[16] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11):789–796, 2009.

[17] B. Bouvier, K. Zakrzewska, and R. Lavery. Protein–dna recognition triggered by a dna conformational switch. *Angew. Chem. Intl. Ed.*, 50(29):6516–6518, 2011.

[18] J. Boyes and A. Bird. Dna methylation inhibits transcription indirectly via a methyl-cpg binding protein. *Cell*, 64(6):1123–1134, 1991.

[19] B. Brodsky and A. V. Persikov. Molecular structure of the collagen triple helix. *Adv. Protein Chem.*, 70:301–339, 2005.

[20] T. H. C. Brondijk, D. Bihan, R. W. Farndale, and E. G. Huizinga. Implications for collagen i chain registry from the structure of the collagen von willebrand factor a3 domain complex. *Proceedings of the National Academy of Sciences*, 109(14):5253–5258, 2012.

[21] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[22] N. Bruant, D. Flatters, R. Lavery, and D. Genest. From atomic to mesoscopic descriptions of the internal dynamics of dna. *Biophys. J.*, 77(5):2366–2376, 1999.

[23] A. Brunet, C. Tardin, L. Salomé, P. Rousseau, N. Destainville, and M. Manghi. Dependence of dna persistence length on ionic strength of solutions with monovalent and divalent salts: A joint theory–experiment study. *Macromolecules*, 48:3641–3652, 2015.

[24] R. J. Camp, M. Liles, J. Beale, N. Saeidi, B. Flynn, E. Moore, S. K. Murthy, and J. W. Ruberti. Molecular Mechanochemistry: Low Force Switch Slows Enzymatic Cleavage of Human Type I Collagen Monomer. *J. Am. Chem. Soc.*, 133:4073–4078, 2011.

[25] S.-W. Chang, B. P. Flynn, J. W. Ruberti, and M. J. Buehler. Molecular mechanism of force induced stabilization of collagen against enzymatic breakdown. *Biomaterials*, 33(15):3852 – 3859, 2012.

[26] M. Chaplin. Do we underestimate the importance of water in cell biology? *Nat. Rev. Mol. Cell Biol.*, 7(11):861–866, 2006.

[27] S. D. Chipman, H. O. Sweet, D. J. McBride, Jr., M. T. Davisson, S. C. Marks, Jr., A. R. Shuldiner, R. J. Wenstrup, D. W. Rowe, and J. R. Shapiro. Defective pro $\alpha 2(I)$ collagen synthesis in a recessive mutation in mice: A model of human osteogenesis imperfecta. *Proc. Natl. Acad. Sci. USA*, 90(5):1701–1705, 1993.

[28] E. Y. D. Chua, D. Vasudevan, G. E. Davey, B. Wu, and C. A. Davey. The mechanics behind dna sequence-dependent properties of the nucleosome. *Nucleic Acids Res.*, 40(13):6338–6352, 2012.

[29] L. Chung, D. Dinakarpandian, N. Yoshida, J. L. Lauer-Fields, G. B. Fields, R. Visse, and H. Nagase. Collagenase unwinds triple-helical collagen prior to peptide bond hydrolysis. *EMBO J.*, 23(15):3020–3030, 2004.

[30] B. D. Coleman, E. H. Dill, M. Lembo, Z. Lu, and I. Tobias. On the dynamics of rods in the theory of kirchhoff and clebsch. *Arch. Rational Mech. Anal.*, 121(4):339–359, 1993.

[31] C. K. Collings, P. J. Waddell, and J. N. Anderson. Effects of dna methylation on nucleosome stability. *Nucleic Acids Res.*, 41(5):2918–2931, 2013.

[32] J. A. Cuesta-Seijo and G. M. Sheldrick. Structures of complexes between echinomycin and duplex dna. *Acta Crystallogr. D Biol. Crystallogr.*, 61(4):442–448, 2005.

[33] P. D. Dans, A. Pérez, I. Faustino, R. Lavery, and M. Orozco. Exploring polymorphisms in b-dna helical conformations. *Nucleic Acids Res.*, 40(21):10668–10678, 2012.

[34] P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano. A coarse grained model for atomic-detailed dna simulations with explicit electrostatics. *J. Chem. Theory Comp.*, 6(5):1711–1725, 2010.

[35] S. Derreumaux, M. Chaoui, G. Tevanian, and S. Fermandjian. Impact of cpg methylation on structure, dynamics and solvation of camp dna responsive element. *Nucleic Acids Res.*, 29(11):2314–2326, 2001.

[36] H. DeVoe and I. Tinoco Jr. The stability of helical polynucleotides: base contributions. *J. Mol. Biol.*, 4(6):500–517, 1962.

[37] R. E. Dickerson. Definitions and nomenclature of nucleic acid structure parameters. *J. Biomol. Struct. Dyn.*, 6(4):627–634, 1989.

[38] K. C. Dong and J. M. Berger. Structural basis for gate-dna recognition and bending by type iia topoisomerases. *Nature*, 450(7173):1201–1205, 2007.

[39] T. Dršata, N. Špačková, P. Jurečka, M. Zgarbová, J. Šponer, and F. Lankaš. Mechanical properties of symmetric and asymmetric dna a-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Res.*, pages 7383–7394, 2014.

[40] Q. Du, Z. Wang, and V. L. Schramm. Human dnmt1 transition state structure. *Proc. Natl. Acad. Sci. USA*, page 201522491, 2016.

[41] M. A. El Hassan and C. R. Calladine. The assessment of the geometry of dinucleotide steps in double-helical dna; a new local calculation scheme. *J. Mol. Biol.*, 251(5):648–664, 1995.

[42] B. Eslami-Mossallam and M. R. Ejtehadi. Stretching an anisotropic dna. *J. Chem. Phys.*, 128(12):125106, 2008.

[43] I. Faustino, A. Pérez, and M. Orozco. Toward a consensus view of duplex rna flexibility. *Biophys. J.*, 99(6):1876–1885, 2010.

[44] S. E. Feller, R. W. Pastor, A. Rojnuckarin, S. Bogusz, and B. R. Brooks. Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.*, 100(42):17011–17020, 1996.

[45] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: the langevin piston method. *J. Chem. Phys.*, 103(11):4613–4621, 1995.

[46] G. B. Fields. A model for interstitial collagen catabolism by mammalian collagenases. *J. Theor. Biol.*, 153(4):585–602, 1991.

[47] G. B. Fields. Interstitial collagen catabolism. *J. Biol. Chem.*, 288(13):8785–8793, 2013.

[48] B. P. Flynn, A. P. Bhole, N. Saeidi, M. Liles, C. A. DiMarzio, and J. W. Ruberti. Mechanical Strain Stabilizes Reconstituted Collagen Fibrils against Enzymatic Degradation by Mammalian Collagenase Matrix Metalloproteinase 8 (MMP-8). *PLoS ONE*, 5(8):e12337, 2010.

[49] N. Foloppe and A. D. MacKerell Jr. Intrinsic conformational properties of deoxyribonucleosides: implicated role for cytosine in the equilibrium among the a, b, and z forms of dna. *Biophys. J.*, 76(6):3206–3218, 1999.

[50] P. Fratzl. Cellulose and collagen: from fibres to tissues. *Curr. Opin. Colloid In.*, 8(1):32–39, 2003.

[51] P. Fratzl, K. Misof, I. Zizak, G. Rapp, H. Amenitsch, and S. Bernstorff. Fibrillar structure and mechanical properties of collagen. *J. Struct. Biol.*, 122:119–122, 1997.

[52] A. Gautieri, M. J. Buehler, and A. Redaelli. Deformation rate controls elasticity and unfolding pathway of single tropocollagen molecules. *J. Mech. Behav. Biomed.*

Mater., 2(2):130–137, 2009.

[53] S. Geggier and A. Vologodskii. Sequence dependence of dna bending rigidity. *Proc. Natl. Acad. Sci. USA*, 107(35):15421–15426, 2010.

[54] J. Gore, Z. Bryant, M. Nöllmann, M. U. Le, N. R. Cozzarelli, and C. Bustamante. Dna overwinds when stretched. *Nature*, 442(7104):836–839, 2006.

[55] A. Goriely and M. Tabor. Nonlinear dynamics of filaments I. Dynamical instabilities. *Physica D*, 105:20–44, 1997.

[56] K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, P. L. Paris, D. C. Tahmassebi, and E. T. Kool. Experimental measurement of aromatic stacking affinities in the context of duplex dna. *J. Am. Chem. Soc.*, 118(34):8182–8183, 1996.

[57] C. Gullekson, L. Lucas, K. Hewitt, and L. Kreplak. Surface-sensitive raman spectroscopy of collagen i fibrils. *Biophy. J.*, 100(7):1837–1845, 2011.

[58] M. F. Hadi, E. A. Sander, J. W. Ruberti, and V. H. Barocas. Simulated remodeling of loaded collagen networks via strain-dependent enzymatic degradation and constant-rate fiber growth. *Mech. Mater.*, 44:72–82, 2012.

[59] P. J. Hagerman. Flexibility of dna. *Annu. Rev. Biophys. Biophys. Chem.*, 17(1):265–286, 1988.

[60] S. Han, E. Makareeva, N. V. Kuznetsova, A. M. DeRidder, M. B. Sutter, W. Losert, C. L. Phillips, R. Visse, H. Nagase, and S. Leikin. Molecular mechanism of type I collagen homotrimer resistance to mammalian collagenases. *J. Cell Biol.*, 285(29):22276–22281, 2010.

[61] S. Han, D. J. McBride, W. Losert, and S. Leikin. Segregation of type I collagen homo-and heterotrimers in fibrils. *J. Mol. Biol.*, 383(1):122–132, 2008.

[62] J. T. Hare and J. H. Taylor. One role for dna methylation in vertebrate cells is strand discrimination in mismatch repair. *Proc. Natl. Acad. Sci. USA*, 82(21):7350–7354, 1985.

[63] R. Harley, D. James, and A. Miller. Phonons and the elastic moduli of collagen and muscle. *Nature*, 267:285–287, 1977.

[64] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, and A. D. MacKerell Jr. Optimization of the charmm additive force field for dna: Improved treatment of the bi/bii conformational equilibrium. *J. Chem. Theory Comp.*, 8(1):348–362, 2012.

[65] M. E. Hogan and R. H. Austin. Importance of dna stiffness in protein–dna binding specificity. *Nature*, 329:263–266, 1987.

[66] M. E. Hogan, J. LeGrange, and B. Austin. Dependence of dna helix flexibility on base composition. *Nature*, 304:752–754, 1983.

[67] M. E. Hogan, M. W. Roberson, and R. H. Austin. Dna flexibility variation may dominate dnase i cleavage. *Proc. Natl. Acad. Sci. USA*, 86(23):9273–9277, 1989.

[68] E. Hohenester, T. Sasaki, C. Giudici, R. W. Farndale, and H. P. Bächinger. Structural basis of sequence-specific collagen recognition by sparc. *Proceedings of the National Academy of Sciences*, 105(47):18273–18277, 2008.

[69] D. S. Horowitz and J. C. Wang. Torsional rigidity of dna and length dependence of the free energy of dna supercoiling. *J. Mol. Biol.*, 173(1):75–91, 1984.

[70] J. Howard. *Mechanics of motor proteins and the cytoskeleton*. Sinauer Associates Sunderland, MA, 2001.

[71] C.-L. Hsieh. Stability of patch methylation and its impact in regions of transcriptional initiation and elongation. *Mol. Cell. Biol.*, 17(10):5897–5904, 1997.

[72] S. Hu, J. Wan, Y. Su, Q. Song, Y. Zeng, H. N. Nguyen, J. Shin, E. Cox, H. S. Rho, and C. Woodard. Dna methylation presents distinct binding sites for human transcription factors. *Elife*, 2:e00726, 2013.

[73] C. Huang and I. V. Yannas. Mechanochemical studies of enzymatic degradation of insoluble collagen fibers. *J. Biomed. Mater. Res.*, 11:137–154, 1977.

[74] W. Humphrey, A. Dalke, and K. Schulten. Vmd: Visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.

[75] W. Hwang. Calculation of conformation-dependent biomolecular forces. *J. Chem. Phys.*, 127:175104, 2007.

[76] A.-P. Hynninen and M. F. Crowley. New faster charmm molecular dynamics engine. *J. Comput. Chem.*, 35(5):406–413, 2014.

[77] W. Im, M. S. Lee, and C. L. Brooks. Generalized born model with a simple smoothing function. *J. Comput. Chem.*, 24(14):1691–1702, 2003.

[78] A. Jain, N. Vaidehi, and G. Rodriguez. A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys.*, 106(2):258–268, 1993.

[79] A. A. Jalan and J. D. Hartgerink. Pairwise interactions in collagen and the design of heterotrimeric helices. *Curr. Opin. Chem. Biol.*, 17(6):960–967, 2013.

[80] I. Jimenez-Useche, J. Ke, Y. Tian, D. Shim, S. C. Howell, X. Qiu, and C. Yuan. Dna methylation regulated nucleosome dynamics. *Sci. Rep.*, 3, 2013.

[81] I. Jimenez-Useche, D. Shim, J. Yu, and C. Yuan. Unmethylated and methylated cpg dinucleotides distinctively regulate the physical properties of dna. *Biopolymers*, 101(5):517–524, 2014.

[82] P. A. Jones. Dna methylation errors and cancer. *Cancer Res.*, 56(11):2463–2467, 1996.

[83] P. A. Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492, 2012.

[84] P. A. Jones and S. B. Baylin. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.*, 3(6):415–428, 2002.

[85] P. A. Jones and D. Takai. The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070, 2001.

[86] P. L. Jones, G. C. J. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis, and A. P. Wolffe. Methylated dna and mecp2 recruit histone deacetylase to repress transcription. *Nat. Genet.*, 19(2):187–191, 1998.

[87] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[88] C. G. Kalodimos, N. Biris, A. M. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. Structure and flexibility adaptation in nonspecific and specific protein-dna complexes. *Science*, 305(5682):386–389, 2004.

[89] E. K. W. Karla, J. W. Bourne, and P. A. Torzilli. Deformation-dependent enzyme mechanokinetic cleavage of type I collagen. *J. Biomech. Eng.*, 131:051004, 2009.

[90] A. Karolak and A. Vaart. Enhanced sampling simulations of dna step parameters. *J. Comput. Chem.*, 35(32):2297–2304, 2014.

[91] R. Z. Kramer, J. Bella, P. Mayville, B. Brodsky, and H. M. Berman. Sequence dependent conformational variations of collagen triple-helical structure. *Nat. Struct. Biol.*, 6(5):454–457, 1999.

[92] R. Z. Kramer, M. G. Venugopal, J. Bella, P. Mayville, B. Brodsky, and H. M. Berman. Staggered molecular packing in crystals of a collagen-like peptide with a single charged pair. *J. Mol. Biol.*, 301(5):1191–1205, 2000.

[93] R. Z. Kramer, L. Vitagliano, J. Bella, R. Berisio, L. Mazzarella, B. Brodsky, A. Zagari, and H. M. Berman. X-ray crystallographic determination of a collagen-like peptide with the repeating sequence (Pro-Pro-Gly). *J. Mol. Biol.*, 280:623–638, 1998.

[94] S. K. Lakkaraju and W. Hwang. Critical buckling length versus persistence length: what governs biofilament conformation? *Phys. Rev. Lett.*, 102(11):118102, 2009.

[95] S. K. Lakkaraju and W. Hwang. Modulation of elasticity in functionally distinct domains of the tropomyosin coiled-coil. *Cell. Mol. Bioeng.*, 2:57–65, 2009.

[96] L. D. Landau and E. M. Lifshitz. *Statistical Physics Part I*. Pergamon Press, Oxford, 3 edition, 1980.

[97] L. D. Landau and E. M. Lifshitz. *Course of Theoretical Physics, Vol. 7: Theory of Elasticity*. Butterworth-Heinemann, 3 edition, 1986.

[98] M. Langecker, A. Ivankin, S. Carson, S. R. Kinney, F. C. Simmel, and M. Wanunu. Nanopores suggest a negligible influence of cpg methylation on nucleosome packaging and stability. *Nano Lett.*, 15(1):783–790, 2014.

[99] F. Lankaš. Dna sequence-dependent deformabilityinsights from computer simulations. *Biopolymers*, 73(3):327–339, 2004.

[100] F. Lankaš, J. Šponer, P. Hobza, and J. Langowski. Sequence-dependent elastic properties of dna. *J. Mol. Biol.*, 299(3):695–709, 2000.

[101] F. Lankaš, J. Šponer, J. Langowski, and T. E. Cheatham. Dna basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, 85(5):2872–2883, 2003.

[102] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nuc. Acids Res.*, 38(1):299–313, 2010.

[103] T. T. Le and H. D. Kim. Measuring shape-dependent looping probability of dna. *Biophys. J.*, 104(9):2068–2076, 2013.

[104] J. Y. Lee, J. Lee, H. Yue, and T.-H. Lee. Dynamics of nucleosome assembly and effects of dna methylation. *J. Biol. Chem.*, 290(7):4291–4303, 2015.

[105] J. Y. Lee and T.-H. Lee. Effects of dna methylation on the structure of nucleosomes. *J. Am. Chem. Soc.*, 134(1):173–175, 2012.

[106] S. Lee, S.-R. Jung, K. Heo, J. A. W. Byl, J. E. Deweese, N. Osheroff, and S. Hohng. Dna cleavage and opening reactions of human topoisomerase iiα are regulated via mg^{2+}-mediated dynamic bending of gate-dna. *Proc. Natl. Acad. Sci. USA*, 109(8):2925–2930, 2012.

[107] M. Leijon and A. Gräslund. Effects of sequence and length on imino proton exchange and base pair opening kinetics in dna oligonucleotide duplexes. *Nucleic Acids Res.*, 20(20):5339–5343, 1992.

[108] S. Leikin, V. A. Parsegian, D. C. Rau, and R. P. Rand. Hydration forces. *Annu. Rev. Phys. Chem.*, 44(1):369–395, 1993.

[109] S. Leikin, D. C. Rau, and V. A. Parsegian. Temperature-favoured assembly of collagen is driven by hydrophilic not hydrophobic interactions. *Nat. Struct. Biol.*, 2:205–210, 1995.

[110] E. Leikina, M. V. Mertts, N. Kuznetsova, and S. Leikin. Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci. USA*, 99:1314–1318, 2002.

[111] E. Li, C. Beard, and R. Jaenisch. Role for dna methylation in genomic imprinting. *Nature*, 366(6453):362–365, 1993.

[112] Y. Li, X. Mo, D. Kim, and S. M. Yu. Template-tethered collagen mimetic peptides for studying heterotrimeric triple-helical interactions. *Biopolymers*, 95(2):94–104, 2011.

[113] T. Lionnet, S. Joubaud, R. Lavery, D. Bensimon, and V. Croquette. Wringing out dna. *Phys. Rev. Lett.*, 96(17):178102, 2006.

[114] J. Lipfert, J. W. Kerssemakers, T. Jager, and N. H. Dekker. Magnetic torque tweezers: measuring torsional stiffness in dna and reca-dna filaments. *Nat. Methods*, 7(12):977–980, 2010.

[115] H. D. Loof, L. Nilsson, and R. Rigler. Molecular dynamics simulation of galanin in aqueous and nonaqueous solution. *J. Am. Chem. Soc.*, 114(11):4028–4035, 1992.

[116] A. C. Lorenzo and E. R. Caffarena. Elastic properties, young's modulus determination and structural stability of the tropocollagen molecule: a computational study by steered molecular dynamics. *J. Biomech.*, 38(7):1527–1533, 2005.

[117] J. C. Lotz, T. Hadi, C. Bratton, K. M. Reiser, and A. H. Hsieh. Anulus fibrosus tension inhibits degenerative structural changes in lamellar collagen. *Eur. Spine J.*, 17(9):1149–1159, 2008.

[118] H. H. Lovelady, S. Shashidhara, and W. G. Matthews. Solvent specific persistence length of molecular type I collagen. *Biopolymers*, 101(4):329–335, 2014.

[119] X.-J. Lu and W. K. Olson. 3dna: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, 3(7):1213–1227, 2008.

[120] N. Ma and A. van der Vaart. Anisotropy of b-dna groove bending. *J. Am. Chem. Soc.*, 138(31):9951–9958, 2016.

[121] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.*, 102(18):3586–3616, 1998.

[122] E. Makareeva, S. Han, J. C. Vera, D. L. Sackett, K. Holmbeck, C. L. Phillips, R. Visse, H. Nagase, and S. Leikin. Carcinomas Contain a Matrix Metalloproteinase–Resistant Isoform of Type I Collagen Exerting Selective Support to Invasion. *Cancer Res.*, 70(11):4366–4374, 2010.

[123] S. W. Manka, F. Carafoli, R. Visse, D. Bihan, N. Raynal, R. W. Farndale, G. Murphy, J. J. Enghild, E. Hohenester, and H. Nagase. Structural insights into triple-helical collagen cleavage by matrix metalloproteinase 1. *Proc. Natl. Acad. Sci. USA*, 109(31):12461–12466, 2012.

[124] J. F. Marko. Stretching must twist dna. *Europhys. Lett.*, 38(3):183, 1997.

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

[126] A. Matsumoto and N. Go. Dynamic properties of double-stranded dna by normal mode analysis. *J. Chem. Phys.*, 110(22):11070–11075, 1999.

[127] A. K. Maunakea, R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. DSouza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, and Y. Zhao. Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature*, 466(7303):253–257, 2010.

[128] A. K. Mazur. Evaluation of elastic properties of atomistic dna models. *Biophys. J.*, 91(12):4507–4518, 2006.

[129] D. J. McBride Jr., V. Choe, J. R. Shapiro, and B. Brodsky. Altered collagen structure in mouse tail tendon lacking the α2(I) chain. *J. Mol. Biol.*, 170:275–284, 1997.

[130] D. J. McBride Jr., K. E. Kadler, Y. Hojima, and D. J. Prockop. Self-assembly into fibrils of a homotrimer of type I collagen. *Matrix*, 12(4):256–263, 1992.

[131] M. J. McCauley and M. C. Williams. Optical tweezers experiments resolve distinct modes of dna-protein binding. *Biopolymers*, 91(4):265–282, 2009.

[132] A. R. McDonald, E. J. Denning, and A. D. MacKerell Jr. Impact of geometry optimization on base–base stacking interaction energies in the canonical a-and b-forms of dna. *J. Phys. Chem. A*, 117(7):1560–1568, 2013.

[133] C. A. Miles and A. J. Bailey. Thermally labile domains in the collagen molecule. *Micron*, 32(3):325–332, 2001.

[134] D. Min, H. Li, G. Li, B. A. Berg, M. O. Fenley, and W. Yang. Efficient sampling of ion motions in molecular dynamics simulations on DNA: variant Hamiltonian replica exchange method. *Chem. Phys. Lett.*, 454(4):391–395, 2008.

[135] Y. Nabeshima, E. S. Grood, A. Sakurai, and J. H. Herman. Uniaxial tension inhibits tendon collagen degradation by collagenase in vitro. *J. Orthop. Res.*, 14:123–130, 1996.

[136] X. Nan, H.-H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman, and A. Bird. Transcriptional repression by the methyl-cpg-binding protein mecp2 involves a histone deacetylase complex. *Nature*, 393(6683):386–389, 1998.

[137] S. Neidle. *Nucleic acid structure and recognition*. Oxford University Press, 2002.

[138] P. S. Nerenberg and C. M. Stultz. Differential Unfolding of α1 and α2 Chains in Type I Collagen and Collagenolysis. *J. Mol. Biol.*, 382(1):246–256, 2008.

[139] T. T. Ngo, J. Yoo, Q. Dai, Q. Zhang, C. He, A. Aksimentiev, and T. Ha. Effects of cytosine modifications on dna flexibility and nucleosome mechanical stability. *Nat. Commun.*, 7, 2016.

[140] T. T. Ngo, Q. Zhang, R. Zhou, J. G. Yodh, and T. Ha. Asymmetric unwrapping of nucleosomes under tension directed by dna local flexibility. *Cell*, 160(6):1135–1144, 2015.

[141] Y. Nishi, S. Uchiyama, M. Doi, Y. Nishiuchi, T. Nakazawa, T. Ohkubo, and Y. Kobayashi. Different effects of 4-hydroxyproline and 4-fluoroproline on the stability of collagen triple helix. *Biochemistry*, 44(16):6034–6042, 2005.

[142] A. Noy and R. Golestanian. Length scale dependence of dna mechanical properties. *Phys. Rev. Lett.*, 109(22):228101, 2012.

[143] A. Noy, A. Pérez, C. A. Laughton, and M. Orozco. Theoretical study of large conformational transitions in dna: the b a conformational change in water and ethanol/water. *Nucleic Acids Res.*, 35(10):3330–3338, 2007.

[144] T. M. Okonogi, S. C. Alley, A. W. Reese, P. B. Hopkins, and B. H. Robinson. Sequence-dependent dynamics of duplex dna: the applicability of a dinucleotide model. *Biophys. J.*, 83(6):3446–3459, 2002.

[145] K. Okuyama, K. Miyama, K. Mizuno, and H. P. Bächinger. Crystal structure of (gly-pro-hyp) 9: Implications for the collagen molecular model. *Biopolymers*, 97(8):607–616, 2012.

[146] L. E. R. O'Leary, J. A. Fallas, E. L. Bakota, M. K. Kang, and J. D. Hartgerink. Multi-hierarchical self-assembly of a collagen mimetic peptide from triple helix to nanofibre and hydrogel. *Nature Chem.*, 3(10):821–828, 2011.

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

[148] W. K. Olson, D. Swigon, and B. D. Coleman. Implications of the dependence of the elastic properties of dna on nucleotide sequence. *Phil. Trans. R. Soc. A*, 362(1820):1403–1422, 2004.

[149] J. P. R. O. Orgel, T. C. Irving, A. Miller, and T. J. Wess. Microfibrillar structure of type I collagen *in situ*. *Proc. Natl. Acad. Sci. USA*, 103:9001–9005, 2006.

[150] M. Orozco, A. Noy, and A. Pérez. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, 18(2):185–193, 2008.

[151] C. O. Pabo and R. T. Sauer. Protein-dna recognition. *Ann. Rev. Biochem.*, 53(1):293–321, 1984.

[152] M. J. Packer, M. P. Dauncey, and C. A. Hunter. Sequence-dependent dna structure: dinucleotide conformational maps. *J. Mol. Biol.*, 295(1):71–83, 2000.

[153] B. Panning and R. Jaenisch. Rna and the epigenetic regulation of x chromosome inactivation. *Cell*, 93(3):305–308, 1998.

[154] M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Pérez, D. Petkeviciut, N. Spackova, J. Sponer, K. Zakrzewska, and R. Lavery. μabc: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in b-dna. *Nucleic Acids Res.*, 42:12272–12283, 2014.

[155] A. Pérez, C. L. Castellazzi, F. Battistini, K. Collinet, O. Flores, O. Deniz, M. L. Ruiz, D. Torrents, R. Eritja, and M. Soler-López. Impact of methylation on the physical properties of dna. *Biophys. J.*, 102(9):2140–2148, 2012.

[156] A. Pérez, F. J. Luque, and M. Orozco. Frontiers in molecular dynamics simulations of dna. *Acc. Chem. Res.*, 45(2):196–205, 2012.

[157] J. P. Peters, S. P. Yelgaonkar, S. G. Srivatsan, Y. Tor, and L. J. Maher. Mechanical properties of dna-like polymers. *Nucleic acids Res.*, 41(22):10593–10604, 2013.

[158] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. Ucsf chimeraa visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605–1612, 2004.

[159] W. H. Press, S. A. Teukolsky, W. Vetterling, and B. P. Flannery. *Numerical recipes: The art of scientific computing*. Cambridge university press, 3 edition, 2007.

[160] J. K. Rainey and M. C. Goh. An interactive triple-helical collagen builder. *Bioinformatics*, 20(15):2458–2459, 2004.

[161] G. N. Ramachandran, M. Bansal, and R. S. Bhatnagar. A hypothesis on the role of hydroxyproline in stabilizing collagen structure. *Biochim. Biophys. Acta*, 322:166–171, 1973.

[162] K. M. Ravikumar, J. D. Humphrey, and W. Hwang. Spontaneous unwinding of a labile domain in a collagen triple helix. *J. Mech. Mater. Struct.*, 2(6):999–1010, 2007.

[163] K. M. Ravikumar and W. Hwang. Region-specific role of water in collagen unwinding and assembly. *Proteins: Struct. Funct. Bioinf.*, 72(4):1320–1332, 2008.

[164] K. M. Ravikumar and W. Hwang. Role of hydration force in the self-assembly of collagens and amyloid steric zipper filaments. *J. Am. Chem. Soc.*, 133:11766–11773, 2011.

[165] N. Rawat and P. Biswas. Shape, flexibility and packing of proteins and nucleic acids in complexes. *Phys Chem. Chem. Phys.*, 13(20):9632–9643, 2011.

[166] A. Razin and A. D. Riggs. Dna methylation and gene function. *Science*, 210(4470):604–610, 1980.

[167] F. Reif. *Fundamentals of statistical and thermal physics*. McGraw-Hill, 1965.

[168] S. Ricard-Blum. The collagen family. *Cold Spring Harb. Perspect. Biol.*, 3(1):a004978, 2011.

[169] A. Rich and F. H. C. Crick. The molecular structure of collagen. *J. Mol. Biol.*, 3(5):483–506, 1961.

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

[171] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-dna recognition. *Annu. Rev. Biochem.*, 79:233, 2010.

[172] E. Rosta, H. L. Woodcock, B. R. Brooks, and G. Hummer. Artificial reaction coordinate tunneling in free-energy calculations: The catalytic reaction of rnase h. *J. Comput. Chem.*, 30(11):1634–1641, 2009.

[173] P. W. Rothemund. Folding dna to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.

[174] J. W. Ruberti and N. J. Hallab. Strain-controlled enzymatic cleavage of collagen in loaded matrix. *Biochem. Biophys. Res. Comm.*, 336:483–489, 2005.

[175] L. E. Russell, J. A. Fallas, and J. D. Hartgerink. Selective assembly of a high stability aab collagen heterotrimer. *J. Am. Chem. Soc.*, 132(10):3242–3243, 2010.

[176] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.*, 23(3):327–341, 1977.

[177] R. Rzehak, A. Arend, D. Kienle, and W. Zimmermann. *Polymer and Cell Dynamics* (Alt *et al.*, eds.), pages 49–68. Birkhäuser, Basel, Switzerland, 2003.

[178] B. Saccà, C. Renner, and L. Moroder. The chain register in heterotrimeric collagen peptides affects triple helix stability and folding kinetics. *J. Mol. Biol.*, 324(2):309–318, 2002.

[179] R. Salsas-Escat and C. M. Stultz. Conformational selection and collagenolysis in type III collagen. *Proteins: Struct, Funct, Bioinf.*, 78(2):325–335, 2010.

[180] K. Sanderson. Bioengineering: What to make with dna origami. *Nature*, 464(7286):158–159, 2010.

[181] J. SantaLucia, H. T. Allawi, and P. A. Seneviratne. Improved nearest-neighbor parameters for predicting dna duplex stability. *Biochemistry*, 35(11):3555–3562, 1996.

[182] A. Sarai and H. Kono. Protein-dna recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, 34:379–398, 2005.

[183] S. K. Sarkar, B. Marmer, G. Goldberg, and K. C. Neuman. Single-molecule tracking of collagenase on native type I collagen fibrils reveals degradation mechanism. *Curr. Biol.*, 22(12):1047–1056, 2012.

[184] N. Sasaki and S. Odajima. Stress-strain curve and Young's modulus of a collagen molecule as determined by the x-ray diffraction technique. *J. Biomech.*, 29:655–658, 1996.

[185] J. A. Schellman and S. C. Harvey. Static contributions to the persistence length of DNA and dynamic contributions to DNA curvature. *Biophys. Chem.*, 55(1):95–114, 1995.

[186] D. Schübeler. Function and information content of dna methylation. *Nature*, 517(7534):321–326, 2015.

[187] A. Scipioni, C. Anselmi, G. Zuccheri, B. Samori, and P. De Santis. Sequence-dependent dna curvature and flexibility from scanning force microscopy images. *Biophys. J.*, 83(5):2408–2418, 2002.

[188] E. Segal and J. Widom. Poly(dA: dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, 19(1):65–71, 2009.

[189] P. M. Severin, X. Zou, H. E. Gaub, and K. Schulten. Cytosine methylation alters dna mechanical properties. *Nucleic Acids Res.*, 39(20):8740–8751, 2011.

[190] M. Y. Sheinin and M. D. Wang. Twist–stretch coupling and phase transition during dna supercoiling. *Phys. Chem. Chem. Phys.*, 11(24):4800–4803, 2009.

[191] M. D. Shoulders and R. T. Raines. Collagen structure and stability. *Annu. Rev. Biochem.*, 78:929–958, 2009.

[192] M. D. Shoulders, K. A. Satyshur, K. T. Forest, and R. T. Raines. Stereoelectronic and steric effects in side chains preorganize a protein main chain. *Proc. Natl. Acad. Sci. USA*, 107(2):559–564, 2010.

[193] T. Siggers and R. Gordân. Protein–dna binding: complexities and multi-protein codes. *Nucleic Acids Res.*, pages 2099–2111, 2014.

[194] L. Sivakumara and G. Agarwal. The influence of discoidin domain receptor 2 on the persistence length of collagen type I fibers. *Biomaterials*, 31(18):4802–4808, 2010.

[195] S. Sivaramakrishnan, B. J. Spink, A. Y. Sim, S. Doniach, and J. A. Spudich. Dynamic charge interactions create surprising rigidity in the ER/K α-helical protein motif. *Proc. Natl. Acad. Sci. USA*, 105(36):13356–13361, 2008.

[196] R. F. Sommese, S. Sivaramakrishnan, R. L. Baldwin, and J. A. Spudich. Helicity of short ER/K peptides. *Protein Sci.*, 19(10):2001–2005, 2010.

[197] J. Spiriti, H. Kamberaj, A. M. R. de Graff, M. F. Thorpe, and A. van der Vaart. Dna bending through large angles is aided by ionic screening. *J. Chem. Theory Comp.*, 8(6):2145–2156, 2012.

[198] J. Spiriti and A. van der Vaart. Dna bending through roll angles is independent of adjacent base pairs. *J. Phys. Chem. Lett.*, 3(20):3029–3033, 2012.

[199] J. A. Spudich and S. Sivaramakrishnan. Myosin VI: an innovative motor that challenged the swinging lever arm hypothesis. *Nat. Rev. Mol. Cell Biol.*, 11(2):128–137, 2010.

[200] R. Stein, Y. Gruenbaum, Y. Pollack, A. Razin, and H. Cedar. Clonal inheritance of the pattern of dna methylation in mouse cells. *Proc. Natl. Acad. Sci. USA*, 79(1):61–65, 1982.

[201] F. H. Stillinger. Water revisited. *Science*, 209(4455):451–457, 1980.

[202] R. Straussman, D. Nejman, D. Roberts, I. Steinfeld, B. Blum, N. Benvenisty, I. Simon, Z. Yakhini, and H. Cedar. Developmental programming of cpg island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.*, 16(5):564–571, 2009.

[203] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. Doye, and A. A. Louis. Sequence-dependent thermodynamics of a coarse-grained DNA model. *J. Chem. Phys.*, 137(13):135101, 2012.

[204] R. S. Sundar, R. Gopalakrishnan, R. C. Wade, and V. Subramanian. Structural basis for the varying propensities of different amino acids to adopt the collagen conformation. *J. Phy. Chem. B*, 115(11):2593–2607, 2011.

[205] S. M. Sweeney, J. P. R. O. Orgel, A. Fertala, J. D. McAuliffe, K. R. Turner, G. A. Di Lullo, S. Chen, O. Antipova, S. Perumal, L. Ala-Kokko, A. Forlino, W. A. Cabral, A. M. Barnes, J. C. Marini, and J. D. S. Antonio. Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates. *J. Biol. Chem.*, 283(30):21187–21197, 2008.

[206] S. Taddese, M. C. Jung, C. Ihling, A. Heinz, R. H. Neubert, and C. E. Schmelzer. Mmp-12 catalytic domain recognizes and cleaves at multiple sites in human skin collagen type i and type iii. *Biochim. Biophys. Acta, Proteins Proteomics*, 1804(4):731–739, 2010.

[207] D. Takai and P. A. Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA*, 99(6):3740–3745, 2002.

[208] P. H. Tate and A. P. Bird. Effects of dna methylation on dna-binding proteins and gene expression. *Curr. Opin. Genet. Dev.*, 3(2):226–231, 1993.

[209] N. A. Temiz, D. E. Donohue, A. Bacolla, B. T. Luke, and J. R. Collins. The role of methylation in the intrinsic dynamics of b-and z-dna. *PLoS One*, 7(4):e35558, 2012.

[210] X. Teng and W. Hwang. Chain registry and load-dependent conformational dynamics of collagen. *Biomacromolecules*, 15(8):3019–3029, 2014.

[211] X. Teng and W. Hwang. Structural and dynamical hierarchy of fibrillar collagen. In R. R. Kaunas and A. Zemel, editors, *Cell and Matrix Mechanics*, chapter 4, pages 101–118. CRC Press, 2014.

[212] X. Teng and W. Hwang. Elastic energy partitioning in dna deformation and binding to proteins. *ACS nano*, 10(1):170–180, 2016.

[213] A. A. Travers. The structural basis of dna flexibility. *Phil. Trans. R. Soc. A*, 362(1820):1423–1438, 2004.

[214] R. Vafabakhsh and T. Ha. Extreme bendability of dna less than 100 base pairs long revealed by single-molecule cyclization. *Science*, 337(6098):1097–1101, 2012.

[215] A. van der Vaart. Coupled binding–bending–folding: The complex conformational dynamics of protein-dna binding studied by atomistic molecular dynamics simulations. *BBA-Gen. Subjects*, 1850(5):1091–1098, 2015.

[216] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 3rd edition, 2007.

[217] S. Vesentini, C. F. Fitié, F. M. Montevecchi, and A. Redaelli. Molecular assessment of the elastic properties of collagen-like homotrimer sequences. *Biomech. Model. Mechanobiol.*, 3(4):224–234, 2005.

[218] M. Vologodskaia and A. Vologodskii. Contribution of the intrinsic curvature to measured dna persistence length. *J. Mol. Biol.*, 317(2):205–213, 2002.

[219] A. V. Vologodskii, V. V. Anshelevich, A. V. Lukashin, and M. D. Frank-Kamenetskii. Statistical mechanics of supercoils and the torsional stiffness of the dna double helix. *Nature*, 280(5720):294–298, 1979.

[220] F. Weih, D. Nitsch, A. Reik, G. Schütz, and P. B. Becker. Analysis of cpg methylation and genomic footprinting at the tyrosine aminotransferase gene: Dna methylation alone is not sufficient to prevent protein binding in vivo. *EMBO J.*, 10(9):2559, 1991.

[221] T. J. Wess. Collagen fibril form and function. *Adv. Protein Chem.*, 70:341–374, 2005.

[222] J. Xiao, R. M. Addabbo, J. L. Lauer, G. B. Fields, and J. Baum. Local conformation and dynamics of isoleucine in the collagenase cleavage site provide a recognition signal for matrix metalloproteinases. *J. Biol. Chem.*, 285:34181–34190, 2010.

[223] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic Acids Res*, 34(2):564–574, 2006.

[224] S. M. Yu, Y. Li, and D. Kim. Collagen mimetic peptides: progress towards functional applications. *Soft Matter.*, 7(18):7927–7938, 2011.

[225] C. Yuan, H. Chen, X. W. Lou, and L. A. Archer. Dna bending stiffness on small length scales. *Phys. Rev. Lett.*, 100(1):018102, 2008.

[226] M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš, and P. Jurečka. Base pair fraying in molecular dynamics simulations of dna and rna. *J. Chem. Theory Comp.*, 10(8):3177–3189, 2014.

[227] V. B. Zhurkin, Y. P. Lysov, and V. I. Ivanov. Anisotropic flexibility of dna and the nucleosomal structure. *Nucleic Acids Res.*, 6(3):1081–1096, 1979.

APPENDIX A

MATLAB code to calculate principal axes

```
matrix=load('triads.dat');   % triad already built and stored in this file
opfile1=fopen('./bend/major_int1.dat','w');
opfile2=fopen('./bend/minor_int1.dat','w');
opfile3=fopen('./bend/info_int1','w');

kb=0.13806;
Temp=300;

format long

tn=16;  % number of triad in each frame
an=64;  % number of triad in each frame * 4

int=1;

for i=1:5000   % from first to last frame
for j=3:tn-int-2   % from first to last triad want to calculate

E1=matrix((i-1+5000)*an+(j-1)*4+1,:);   % read triad
E2=matrix((i-1+5000)*an+(j-1)*4+2,:);
E3=matrix((i-1+5000)*an+(j-1)*4+3,:);
F3=matrix((i-1+5000)*an+(j-1+int)*4+3,:);

o1=matrix((i-1+5000)*an+(j)*4,:);   % read origin
o2=matrix((i-1+5000)*an+(j+int)*4,:);
D(i,j)=norm(o2-o1);   % calculate distance

px(i,j)=dot(E1,F3);   % calculate projection
py(i,j)=dot(E2,F3);
pz(i,j)=dot(E3,F3);
```

```
end
end

for j=3:tn-int-2

dots=[px(:,j) py(:,j) pz(:,j)];

xmin=min(px(:,j));   % range to find minimum
xmax=max(px(:,j));
ymin=min(py(:,j));
ymax=max(py(:,j));

reso=0.002;      % resolution for increment of Cart coor
dist=10000;      % arbitary set an initial great circle distance summation
reso2=0.001;     % resolution for increment of angle
dist2=1000000;   % arbitary set an initial great circle distance square summation

for tempx=xmin:reso:xmax
for tempy=ymin:reso:ymax

tempz=sqrt(1-tempx^2-tempy^2);
tempCen=[tempx tempy tempz];   % temporary centroid

dotprod=tempx*dots(:,1)+tempy*dots(:,2)+tempz*dots(:,3);
%%%% central angle vector of all dots with candidate centroid

tempDist=sum(acos(dotprod));   % sum of the distance

if imag(tempDist) > 0.002
fprintf('imaginary part in 1st calculation at triad %i is %f,...
might be an error!\n',j,imag(tempDist))
end

if tempDist < dist
dist=tempDist;   % new minimum
Cen(j,:)=tempCen;   % new 'centroid'
end
```

```
end
end

%%%%%%%%%%% Spherical Centroid Found %%%%%%%%%%

init_norm=cross(Cen(j,:),[1,0,0]);
normal0=init_norm/norm(init_norm);   % normal of temporary major circle
norm_norm=Cen(j,:)          %%%% the normal of the great circle of the
                            %%%% cluster of normal for candidate major spherical axis

for deg=0:reso2:pi

tempNorm=cos(deg)*normal0+sin(deg)*cross(norm_norm,normal0);
%%%% normal of temporary major circle

w=cross(tempNorm,norm_norm);
u=norm_norm;

dotprod2=u(1)*dots(:,1)+u(2)*dots(:,2)+u(3)*dots(:,3);
dotprod3=w(1)*dots(:,1)+w(2)*dots(:,2)+w(3)*dots(:,3);

tempDist2=norm(acos(sqrt(dotprod2.^2+dotprod3.^2)))^2;
%%%% sum of square of distance to candidate major great circle axis

if tempDist2 < dist2
dist2=tempDist2;
Norm_maj(j,:)=tempNorm;
end

end

var_min(j)=dist2/5000;   % distance (angle) variance to major circle

%%%%%%%%%%% Major Spherical Axis Found %%%%%%%%%%

Norm_min(j,:)=cross(Norm_maj(j,:),norm_norm);   % normal of minor bending circle

w2=cross(Norm_min(j,:),norm_norm);
dotprod4=u(1)*dots(:,1)+u(2)*dots(:,2)+u(3)*dots(:,3);
```

```
dotprod5=w2(1)*dots(:,1)+w2(2)*dots(:,2)+w2(3)*dots(:,3);

var_maj(j)=norm(acos(sqrt(dotprod4.^2+dotprod5.^2)))^2/5000;
%%%% distance (angle) variance to minor circle

%%%%%%%%%%% Minor Spherical Axis Found %%%%%%%%%%

fprintf(opfile3,' For case interval %i triad %i,\n centroid is (%f,%f,%f),...
\n major axis norm is (%f,%f,%f),\n minor axis norm is ...
(%f,%f,%f)\n\n',int,j,Cen(j,:),Norm_maj(j,:),Norm_min(j,:));

end


for k=3:tn-int-2

kf_maj(k)=kb*Temp*mean(D(:,k))/var_maj(k);   % stiffness calculation
kf_min(k)=kb*Temp*mean(D(:,k))/var_min(k);

fprintf(opfile1,'%i   %.6f\n',k,kf_maj(k));
fprintf(opfile2,'%i   %.6f\n',k,kf_min(k));

end
```

APPENDIX B

Average elastic energy per base pair in PDB

Table B.1: Total elastic energy E_T per base pair (pN·Å) in PDB structures analyzed. Structures are listed in an increasing order of E_T in each column, as in Fig. 3.18b.

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
2BCQ	9.12	2R2U	16.85	3PML	23.62	1JKQ	29.17	3FSI	31.46
2PFN	11.59	2BCR	17.35	3HX0	23.66	3HDD	29.17	1RZT	31.55
2PFQ	11.98	3PNC	17.57	3CO6	24.56	3RZL	29.34	3VEB	31.65
2BCS	12.92	3PMN	17.66	2BCU	24.57	9ICH	29.40	1JKO	31.73
2PFO	13.05	2IHM	17.68	3MGI	26.03	1KU7	29.48	8ICC	31.99
3HW8	13.06	3AAF	18.88	3RZM	26.36	3GLF	30.31	1ZQN	32.02
1XSP	13.65	3ZVN	19.31	1TRO	27.34	2ZCJ	30.60	8ICF	32.63
4K4G	14.44	1YFL	19.48	2R2R	27.55	2GWS	30.61	9ICG	32.72
3V72	14.94	2R2T	19.64	3OD8	27.56	1ZQR	30.69	3GLG	32.79
3V7K	15.58	3L2C	20.08	2G1P	28.11	2VOA	30.95	1Z9C	32.85
4K4I	15.84	3MGH	20.09	1ZQI	28.21	1NLW	31.11	2I13	32.96
1DNK	15.97	3TED	20.44	3LDY	28.50	2Z6U	31.15	1KBU	33.13
4KB1	15.99	1D1U	21.08	2C6Y	28.53	3GFI	31.23	1XSL	33.26
3V7J	16.64	3G73	23.40	3MHT	29.04	3BQ1	31.36	1FJX	33.37

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
8ICJ	33.39	4IRI	35.76	4ECX	37.53	1REP	39.15	4IHS	40.64
8ICO	33.66	1S9F	35.80	8ICI	37.56	1APL	39.24	4G82	40.68
1ZAA	33.82	1DU0	35.89	2FJV	37.64	4OSK	39.26	1BC7	40.69
2R9L	33.85	4AUW	35.96	2WBU	37.69	3F2C	39.26	3US0	40.69
1ZQM	34.08	1XO0	35.98	1A1H	37.70	3BQ0	39.41	2Z3X	40.74
3MLN	34.17	1ZQH	36.06	1V15	37.70	4ECU	39.53	1ZTT	40.84
1ZQP	34.18	4J9R	36.19	3PR5	37.77	1KSY	39.65	2EUZ	40.89
1C0W	34.20	1ZQC	36.19	4ED6	38.01	4OSH	39.72	2C5R	40.90
4E9G	34.24	2EVG	36.20	1G2F	38.02	3ODA	39.74	3CO7	40.99
3COA	34.54	2AS5	36.28	4ECQ	38.17	1A1L	39.84	1ZQO	40.99
3A01	34.67	1JK2	36.38	2O49	38.30	4A12	39.84	9ICJ	41.12
8ICK	35.00	4HQE	36.48	4ED1	38.67	1ZQF	39.87	3FYL	41.17
4J9S	35.04	4JBM	36.81	4ECY	38.77	4ED2	39.92	2WBS	41.48
2ETW	35.11	2R2S	36.98	1A1F	38.79	1SXQ	40.02	4ED3	41.52
2EUV	35.28	4ED7	37.32	4H0E	38.82	9ICA	40.03	3BQ2	41.60
1ZQS	35.30	1U8R	37.36	1DRG	38.83	1RH6	40.15	2HOS	41.70
4ITQ	35.47	1AAY	37.46	1JX4	38.94	4ECR	40.39	2EVF	41.73
3JSO	35.57	1G2D	37.48	4ECW	38.97	4FZX	40.41	3GLI	41.79
4EOT	35.66	3OA6	37.50	8ICZ	38.98	1FJL	40.41	4HF1	41.81
2IEF	35.72	4KB0	37.50	4ECZ	39.01	4HP3	40.58	1S97	41.82
1ZQB	35.72	1ZQT	37.52	3DSD	39.02	4ECV	40.61	2DRP	41.88

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
4ED8	41.90	8ICM	43.23	3LNQ	44.86	1H6F	46.76	3PVP	48.29
1JK1	41.90	2PI5	43.34	3TQ1	44.91	4NE1	46.82	2A07	48.30
4ED0	41.99	3M4A	43.36	1A1K	44.94	4KLG	46.87	1PP7	48.30
2BZF	42.10	1MNN	43.36	1PUE	45.00	1ZG1	47.05	1OCT	48.31
1NKP	42.18	1CEZ	43.45	2ERE	45.02	4OSJ	47.20	4KNY	48.36
3DSC	42.37	1QAI	43.47	3PVV	45.16	4J9P	47.20	2X6V	48.48
1TRR	42.39	2PZS	43.70	4JWM	45.35	2EUW	47.24	3MLO	48.58
2HOT	42.41	4ECT	43.73	2EVI	45.49	1LE8	47.25	2C9N	48.73
4ECS	42.54	3QYM	43.79	1E3O	45.75	4CRX	47.25	1IMH	48.74
3ZVK	42.54	3A5T	43.85	1PVP	45.76	3JTG	47.50	4OSM	48.79
2EVJ	42.68	4J19	43.89	1AKH	45.90	2DNJ	47.53	3QOQ	48.84
3ZQL	42.69	3M7K	43.90	1OWR	45.93	1ZTW	47.63	3G0R	48.90
1A1G	42.70	1D5Y	43.98	1A1I	46.15	3Q8K	47.66	4OSV	48.93
1ZG5	42.74	3SSE	44.05	1UUT	46.18	3BRF	47.69	1NJX	49.00
1IGN	42.91	2O4A	44.15	2ISZ	46.23	2HR1	47.79	1S10	49.08
1P47	42.95	1XC9	44.23	9ANT	46.40	4EVV	47.90	2O8B	49.16
3U3W	42.96	2YVH	44.48	4KAZ	46.51	3O9X	47.93	1PVR	49.17
4EGZ	43.00	2R0Q	44.53	2C9L	46.51	2IT0	47.95	3G6Q	49.27
3G6U	43.03	1HLO	44.63	4GFB	46.56	2VWJ	47.95	4KLF	49.31
3Q5F	43.09	1BDT	44.65	2EUX	46.60	2D5V	48.02	3OSG	49.53
3QYN	43.09	3OGU	44.68	1KSX	46.71	1B8I	48.14	4R65	49.55

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
2I9K	49.56	1J1V	50.50	4OSS	52.67	1YRN	53.42	2HDD	54.91
3UFD	49.67	3GV8	50.58	1P71	52.74	2FMS	53.60	2XSD	54.96
3PR4	49.70	4OSZ	50.98	1YA6	52.76	4KLO	53.62	3KET	55.05
2QL2	49.72	1F2I	51.00	4J9Q	52.84	4KLH	53.69	3GZ6	55.08
4KLJ	49.83	1ODH	51.15	3N6S	52.91	3WU1	53.73	3SI6	55.23
1IG7	49.92	4IBU	51.25	4KLE	52.95	2Z9O	53.74	3D0A	55.25
4L0Y	49.92	4OT0	51.36	1H88	52.95	1HJC	53.83	1PDN	55.26
1ZQD	49.96	3MKW	51.55	1I6J	52.98	4OT3	53.83	4AIJ	55.28
1ZQL	50.04	1F44	51.64	2DPU	53.01	4IWR	53.86	3MX4	55.36
8ICB	50.09	3BEP	51.72	1S9K	53.06	2Z6Q	53.90	4FJH	55.39
3QMC	50.11	1HDD	51.73	3FDQ	53.07	1B72	54.05	3C25	55.39
3G9M	50.16	3JR9	51.97	1S0O	53.14	4OSW	54.07	3MU6	55.40
1PVQ	50.17	4KLI	52.38	4E0J	53.17	4J2X	54.20	3G9I	55.44
4AV1	50.20	1ZQE	52.40	1L5U	53.17	1OZJ	54.31	4J2D	55.54
1T2T	50.21	3LAP	52.41	1PZU	53.19	2Z6A	54.31	4OST	55.57
1ZQK	50.25	2XY7	52.44	3MLP	53.22	1XBR	54.39	1F5T	55.61
1A1J	50.30	1MA7	52.45	4OSQ	53.22	1U8B	54.42	3M9N	55.62
4KHQ	50.30	1B01	52.45	2IS4	53.26	1XNS	54.47	4GZ0	55.62
2PRT	50.33	2EVH	52.53	4J2B	53.27	8ICH	54.68	3JSP	55.66
3RMP	50.39	2FIO	52.58	1HCR	53.39	3JRD	54.76	4AAB	55.68
1XPX	50.50	2BPF	52.66	1MDY	53.42	3LAJ	54.81	2CRX	55.78

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
2FJX	55.79	2H8R	57.11	4FJK	57.97	3JXD	58.96	1NJW	60.37
3ERE	55.82	1NK4	57.11	2ERG	58.02	9ICN	58.96	3C0W	60.38
1ZQG	55.91	3MKY	57.17	1GU5	58.06	2EFW	59.02	1ECR	60.40
4DSK	55.91	3JXC	57.20	4K1M	58.10	1H89	59.17	4HN6	60.45
1HF0	55.93	3JRH	57.23	3VEA	58.11	1K79	59.30	3CLC	60.50
1DDN	55.99	3CRX	57.23	1B3T	58.18	1OH8	59.35	3CQ8	60.54
3S8Q	56.00	3JRC	57.23	3EXL	58.19	3UXP	59.56	4QTJ	60.56
1GD2	56.16	2R1J	57.24	4F4X	58.26	8ICG	59.60	1P51	60.56
3CBB	56.26	2FMQ	57.37	1AM9	58.37	2H8C	59.62	4OSR	60.60
1IC8	56.29	3M9E	57.43	3L1P	58.49	2V6E	59.66	3T72	60.65
4IHT	56.33	1RYS	57.51	4AAG	58.49	1RPE	59.73	2GEQ	60.78
3SJJ	56.47	1H9D	57.56	1L3V	58.52	1A02	59.82	3QMD	60.79
4KLM	56.49	3QWS	57.59	2NNY	58.57	4A04	59.83	1KB4	60.85
1BDV	56.49	4J2A	57.69	4OSL	58.58	7ICH	59.87	3G6T	60.85
3QRF	56.83	4KLL	57.75	1JE8	58.62	4EEY	59.89	4FTH	60.86
9ICM	56.83	1HWT	57.77	1L3S	58.67	4FZZ	59.91	4HRI	60.90
3BRD	56.87	2EZV	57.80	3JXB	58.73	2B9S	59.96	4AIK	60.96
3D1N	56.88	4J2E	57.84	1BL0	58.74	2ER8	59.96	4FJ9	60.97
1MNM	56.91	3IKT	57.87	1JEY	58.74	4EGY	59.99	3UKG	61.02
4FJL	56.93	4OTO	57.88	3V6T	58.77	3G8U	59.99	3ZP5	61.06
1ZJN	56.93	3RN5	57.93	3JXY	58.91	3MVA	60.26	1PP8	61.22

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
1NK7	61.30	1UA1	62.38	1TQE	63.44	2O8F	65.21	7ICS	66.30
3QMG	61.37	4JCX	62.39	1V14	63.52	1R8D	65.28	3K4X	66.31
3VD1	61.39	3UIQ	62.46	1NK0	63.71	4KIS	65.30	1MJM	66.33
4FJX	61.44	3G6P	62.51	1PAR	63.73	2HZV	65.30	4IZZ	66.36
3KMD	61.49	1WBB	62.56	3MVB	63.77	1Z63	65.35	3N7Q	66.56
2OG0	61.62	3DZU	62.65	4DSL	63.91	1R49	65.39	3QSV	66.61
2QHB	61.67	4JWN	62.75	3JRA	64.02	1VTN	65.39	2PFJ	66.70
2XE0	61.72	3RKQ	62.81	4FJ7	64.05	3Q05	65.45	2HAN	66.79
3GV7	61.74	1VAS	62.84	3JRE	64.07	1ZNS	65.53	2E42	66.92
4HF2	61.75	3QE9	62.90	1ZS4	64.10	3HQF	65.56	4KFC	66.94
3MKZ	61.77	1YSA	62.90	4E0G	64.14	7ICQ	65.62	1I3J	67.18
3Q8L	61.79	1NK6	62.95	3U4Q	64.30	1FOS	65.63	4IHW	67.24
2RBF	61.82	4OSI	62.98	1UA0	64.31	2XMA	65.63	3E6C	67.35
3VD0	61.83	3JRI	63.05	3EZ5	64.40	3G9P	65.76	2E43	67.41
2O8E	61.86	2VS7	63.08	4F4Z	64.44	3SJM	65.87	3P57	67.49
1JNM	61.95	1NK5	63.10	3SCX	64.47	3GV5	65.97	1OH7	67.63
4QTK	62.20	4DSJ	63.12	6PAX	64.55	1K6O	66.05	3KMP	67.64
1EGW	62.26	1WB9	63.24	3UGM	64.78	4AAD	66.06	2H27	67.65
2A66	62.28	4GCT	63.27	2QSH	64.79	4G83	66.08	2O8D	67.66
2DPD	62.36	3C2K	63.30	1P78	64.89	3QMB	66.09	1GU4	67.69
1GXP	62.37	1TTU	63.37	1ZQJ	64.98	1LLI	66.27	1DUX	67.70

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
1MJ2	67.76	3L2Q	68.75	7ICF	70.33	2NTC	71.31	3PTA	72.91
2H1K	67.83	3ODH	68.83	8MHT	70.34	3TS8	71.40	4ASS	72.97
3IV5	68.01	3E00	68.89	3IH7	70.38	4GUO	71.42	3EXJ	73.02
9ICW	68.03	2P5L	69.13	1MDM	70.45	1EJ9	71.46	4IVZ	73.03
4F4W	68.08	9ICS	69.20	1CMA	70.51	2FO1	71.49	1MJO	73.06
2BDP	68.08	1NWQ	69.38	4H10	70.56	4BDP	71.52	3TAR	73.28
1L3U	68.14	1NJY	69.49	4IHY	70.63	3SPD	71.57	1JGG	73.39
3MFK	68.15	7MHT	69.52	1U47	70.65	2NRA	71.65	1PER	73.50
1JKR	68.22	2EX5	69.65	4DM0	70.76	2O93	71.75	2OST	73.57
2VLA	68.28	4HLY	69.73	3JRF	70.78	1OH6	71.84	3S3O	73.81
7ICM	68.29	3US2	69.81	1D0E	70.78	1ZX4	71.91	1R7M	73.82
2UZK	68.32	1HCQ	69.85	3QZ7	70.84	3L2R	71.92	3MXA	73.83
3G8X	68.33	4KLD	69.87	2AC0	70.88	2A3V	71.92	2QSG	73.92
9ICL	68.37	1K78	69.90	1N3E	70.90	2R5Y	72.06	1YF3	73.93
1NK8	68.38	1HUO	69.91	3KK3	70.92	1PUF	72.20	1LMB	73.96
7ICI	68.38	1SA3	70.00	1MUH	70.98	1S0N	72.37	4EFJ	74.05
2GB7	68.53	4GDF	70.02	4F6M	71.04	2WIW	72.48	3CWA	74.07
1OUP	68.62	1KB2	70.20	3POV	71.12	1TF6	72.61	3ECP	74.08
1KB6	68.64	1WBD	70.20	1YFJ	71.15	4BQA	72.66	4NDY	74.16
3OS0	68.66	1YO5	70.21	2XO6	71.19	3L2U	72.66	3ODE	74.23
1W7A	68.70	7ICV	70.23	2I3Q	71.25	3G99	72.74	3K0S	74.30

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
1GM5	74.43	3WTV	75.34	1TSR	76.67	3DZY	77.60	3OYL	79.18
4IQR	74.54	3S3M	75.38	3OYC	76.68	1TUP	77.61	2R8K	79.20
1CF7	74.56	3UXW	75.45	1DH3	76.69	3L2W	77.73	1JKP	79.28
1IO4	74.56	2H7G	75.50	3OYN	76.72	1CA6	77.82	2IS2	79.29
3OOR	74.59	2BAM	75.78	3V79	76.78	3OYF	77.84	1LQ1	79.30
2E52	74.60	2I9T	75.93	1A36	76.78	1F4K	77.98	4HJE	79.35
4IHV	74.65	1YFI	76.00	1LE9	76.85	1NK9	78.00	3OYB	79.36
4KLQ	74.65	2IS6	76.03	2IS1	76.87	9ICR	78.04	1K7A	79.40
1HBX	74.68	3EYZ	76.06	4F4Y	76.87	3OYD	78.09	9ICQ	79.53
3MX9	74.72	3OYG	76.09	3C0X	76.89	2VBO	78.25	1BHM	79.58
1R4O	74.72	7ICT	76.19	3WGI	76.94	1LLM	78.33	3TAQ	79.68
3Q8M	74.79	4F50	76.21	4G3I	76.95	3ZQC	78.33	4B1P	79.72
3QZ8	74.83	2VBN	76.26	3QQY	77.03	3WTW	78.45	3OYE	79.74
4AAE	74.87	3OYK	76.27	2PI4	77.07	4AA6	78.51	2VBL	79.76
3OYA	74.93	7ICR	76.31	3OYI	77.12	3OYM	78.51	4B1O	79.84
2WTF	74.93	4K97	76.34	1AU7	77.15	4KLT	78.84	2I06	79.87
4E7H	74.98	3QMH	76.37	1MUS	77.40	1NFK	79.00	4ATK	79.91
4E7I	75.20	1JT0	76.47	2AHI	77.45	1BPZ	79.11	3HOS	79.97
2VBJ	75.20	3S3N	76.51	3OY9	77.47	3OYJ	79.14	1HJB	80.00
1U78	75.23	1HLZ	76.59	4K1G	77.50	4PXI	79.16	1GLU	80.04
2FQZ	75.33	7ICE	76.66	3WTX	77.52	1ZR4	79.17	1OH5	80.16

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
1CIT	80.22	3OYH	82.21	3PT6	84.25	2I3P	86.24	4KLU	88.00
4HN5	80.33	4BAC	82.30	3OOL	84.36	3C2M	86.33	2O6G	88.11
4BE2	80.36	1RIO	82.37	3U44	84.40	2WWY	86.45	3WTY	88.11
1NG9	80.45	4R66	82.39	8ICE	84.78	1ZR2	86.59	1BP7	88.15
3WTS	80.47	4DQY	82.47	2F8X	84.81	2QNC	86.61	1LAT	88.29
3BRG	80.48	3C2L	82.93	3EBC	84.92	3LJA	86.77	1EWQ	88.44
9ICO	80.71	1R4I	82.93	3G9J	85.09	4F5P	86.79	1ESG	88.44
4BDY	80.71	1RTD	82.93	1SAX	85.10	2O6M	86.84	1U49	88.53
3FHZ	80.93	3OSF	82.99	2CAX	85.13	1TC3	86.88	2R5Z	88.66
9ICX	81.21	3JRB	83.04	4FCY	85.33	4JL3	86.95	1DFM	88.83
4BE1	81.24	1U3E	83.13	4GCK	85.33	3CRO	86.98	1C8C	88.86
1NKE	81.41	4FB3	83.16	1MJQ	85.43	3W2A	87.01	2OR1	89.38
4BE0	81.48	1YNW	83.26	4GCL	85.44	1EA4	87.01	3RI4	89.46
1L3L	81.51	4BDZ	83.38	1D2I	85.58	4KLS	87.01	4FGN	89.65
1QAJ	81.54	2I05	83.39	3A4K	85.62	3R7P	87.16	1N3F	89.70
4DS5	81.70	3COQ	83.50	1ZJM	85.70	3MGR	87.47	1NKC	89.71
3JRG	81.75	4FZY	83.63	3ISB	85.85	4AAF	87.51	1A73	89.80
1DSZ	81.79	3G97	83.68	2HAP	85.89	3WTT	87.56	1S0M	89.85
2VS8	81.88	1ZQQ	83.72	1ZLK	85.96	1BPX	87.72	3KOV	90.05
3Q06	81.96	1GJI	83.89	1GDT	85.97	1PYI	87.91	2XM3	90.14
1T9I	81.99	2HHX	83.93	2ADY	86.01	7ICL	87.92	3WTU	90.50

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
1JJ4	90.52	2QOJ	93.21	3THW	95.35	1G9Y	99.86	3MIS	102.92
3TAP	90.60	2H7H	93.38	2W42	95.65	3IGM	100.16	4FLW	103.09
1RAM	90.71	4JCY	93.41	1ZBB	96.24	1P34	100.35	4FLX	103.23
7ICP	90.75	1YQK	93.43	9ICE	96.42	4FM2	100.37	3AV1	103.26
1MM8	90.98	3W3C	93.52	2BGW	96.74	2QBY	100.65	1FOK	103.27
1A74	91.04	7ICN	93.77	3L2V	96.84	2GM4	100.92	1H0M	103.28
1KX5	91.09	1WD0	93.78	4HC7	96.97	3T79	100.92	4HC9	103.28
3IGC	91.09	1CZ0	93.88	1IPP	97.12	3AZG	100.93	3EH8	103.54
3G6R	91.23	2AGO	93.93	3MGS	97.30	4FLV	101.07	3DFX	103.63
4I6Z	91.26	2CV5	94.27	7ICU	97.32	3W6V	101.15	3THZ	103.71
3RN2	91.40	4R63	94.31	4E7K	97.36	3TU4	101.24	1P3L	103.76
4HYK	91.45	2FJW	94.35	4R64	97.65	4FLU	101.24	4JBK	104.09
3QEB	91.46	1P8K	94.39	3A6N	98.05	1G9Z	101.36	4FLY	104.10
3UTB	91.78	2VE9	94.45	3OS2	98.20	3E54	101.44	4I2O	104.14
4G7H	91.80	2BNW	94.57	9ICT	98.52	4FLT	101.67	1H9T	104.49
1T2K	91.93	3QEA	94.63	4HCA	98.56	3REK	101.99	2HOI	104.84
2PYO	91.96	2WTU	94.78	3REJ	98.62	2O61	102.12	1P3O	104.84
3C28	92.10	2IHN	94.84	1HLV	98.63	3UTA	102.22	3AFA	104.97
1CYQ	92.87	1UBD	94.90	3SQI	98.94	3AZL	102.34	2RVE	105.02
2IVH	93.07	1T9J	95.07	2EWJ	98.97	1P7H	102.66	4J00	105.15
3VW4	93.12	3MVD	95.08	2C7A	98.98	2BPG	102.72	4DWI	105.24

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
2OH2	105.47	1LE5	107.21	4J8U	111.29	3ODC	114.10	1NKB	118.45
1P3G	105.56	3FC3	107.49	1WVL	111.60	4K96	114.35	1A3Q	118.49
1RZ9	105.61	1P3K	107.58	3FBD	111.73	1W36	114.42	3WA9	118.54
3REL	105.73	3WAA	107.69	3AZK	111.83	2WJ0	114.65	3LEL	119.01
4FM0	105.84	1QP9	107.90	3N97	112.01	1IF1	114.87	4J8X	119.30
9ICU	106.09	3IAG	107.91	3VD6	112.09	3KXB	115.04	3THX	119.41
3MGP	106.16	3PJR	107.93	4J8W	112.21	1M18	115.14	1KX4	119.43
2PI0	106.19	4L0Z	107.93	4KGC	112.24	1F66	115.26	1P3M	119.59
3MGQ	106.21	3UT9	108.05	1IU3	112.28	2V1U	115.34	1P3B	120.03
1P3P	106.26	3AZJ	108.20	3TAN	112.29	4J8V	115.43	2R8J	120.63
2IBK	106.36	3QFQ	108.40	3DFV	112.34	1R0N	115.75	3KUY	120.80
2BNZ	106.53	1VRR	108.56	1KLN	112.62	4IX7	115.83	3W96	120.84
2QNF	106.63	4GLX	108.87	1S32	112.79	3REH	116.12	1D66	121.07
3AV2	106.63	4JJN	109.23	4E7J	112.87	1VKX	116.23	1LWT	121.24
1P3A	106.66	1RYR	109.27	3ZKC	113.31	1P3I	116.74	1XYI	121.33
3AZN	106.73	1MJP	109.29	1P3F	113.46	7ICJ	116.84	1AOI	121.50
4IKF	106.73	1EQZ	109.30	1CQT	113.55	2Y9Z	117.28	3W99	121.54
1IAW	106.88	3WKJ	109.54	1KX3	113.59	3UVF	117.72	3B6G	121.74
1M19	106.99	2D45	109.83	3VD2	113.77	4FLZ	117.78	3MIP	121.78
4FM1	107.02	2HOF	110.36	7ICO	113.95	3GOX	117.86	2GLI	122.42
3VEK	107.16	1SXP	110.49	2NZD	114.01	9ICP	118.30	3GYH	122.66

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
3THY	122.66	3NM9	128.83	3AZE	135.87	1RV5	144.42	4I8T	156.78
4KB6	122.96	1ZLA	129.15	1M5X	136.61	3MQ6	144.71	1OUZ	157.10
1R0O	123.36	3MQY	129.40	3V21	137.25	1Z1G	145.33	2XRO	157.54
4IS1	123.39	1M1A	129.68	3RNU	137.28	1G3X	146.65	4G92	158.34
1Q0T	124.57	3KWQ	129.94	4ATI	137.36	3FD2	146.91	2AOQ	162.29
3M9M	124.78	3REI	130.01	1NJZ	137.39	1BY4	146.98	2P0J	165.31
3W98	124.83	3B6F	130.28	4L18	138.34	2IVK	148.11	4AWL	165.31
4HT4	124.88	1BSS	130.46	2FLD	139.48	3ORC	148.33	3LZ1	167.04
4ASO	125.12	3AYW	130.84	6CRO	139.69	1EOP	149.54	1IHF	167.76
3UK3	125.26	1CLQ	132.03	1FW6	140.40	3BDN	150.44	1OWF	169.10
1L3T	125.45	3LZ0	132.33	1HW2	140.57	1RXW	151.01	2OAA	170.22
2V2T	125.73	1U35	132.33	3GXQ	140.62	2IIF	151.21	3AZF	170.37
3DW9	125.88	3DO7	132.51	1K61	140.83	1GT0	151.24	3TMM	170.62
1EVW	125.98	4GZ1	132.65	2IIE	140.90	2B0D	151.49	3NBN	170.63
2P6R	126.16	3AZH	132.99	2HMI	141.31	3W97	153.11	3BM3	170.78
3C1B	126.52	2NQB	133.21	1NNE	141.37	3C1C	153.80	3O62	171.92
1AN4	127.05	2AYB	133.41	4J01	141.51	3AZM	154.21	1D3U	172.35
4D8J	127.26	3DVO	133.63	1LV5	142.71	1STX	154.37	1L2B	175.01
3KO2	127.79	1R4R	133.91	3K5N	142.78	1OWG	154.56	3N7B	175.49
4GZ2	127.80	2WIZ	134.24	4KUD	144.18	3FOF	154.86	1EYU	175.79
2NTZ	128.67	3GUT	134.57	2F8N	144.41	2FJ7	156.34	1FIU	176.60

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
4I3H	176.80	1EOO	197.72	4FM9	232.76	1C9B	264.23	2FKC	301.63
1DC1	177.03	3FOE	201.40	1QRV	233.03	2GIG	268.07	1WTW	306.67
1AZ0	177.17	3RZG	201.83	2RGR	233.49	2GIH	269.11	1WTX	307.23
1SX5	177.44	1RVB	207.26	1B96	236.32	3LTN	269.18	1SUZ	307.79
2AYG	177.54	1A0A	208.94	1VOL	238.24	1NGM	272.39	3E45	308.18
2NOI	179.59	3S57	210.11	1RM1	241.63	3E43	272.67	1WTR	313.26
3F27	180.29	1TGH	210.30	3SWM	245.25	2FL3	273.29	3KSB	314.34
2NP2	180.36	1J5O	211.29	1PT3	246.69	1JFI	274.46	2F5P	314.82
1LWS	181.65	3DPG	211.59	1WTV	249.01	2FLC	277.11	4INM	317.25
1RVA	182.73	1B95	215.36	2F5N	249.30	3F22	280.90	1QNE	320.06
1ID3	182.81	3IMB	215.59	3E44	250.40	1RVC	281.45	1BNZ	320.20
3S5A	183.02	3AN2	216.41	1YTB	250.65	1TAU	281.88	1VTO	320.38
4RVE	184.52	1B97	216.64	2Q10	250.76	3E40	284.37	2PE5	323.90
2GE5	185.76	3GPX	218.57	2ODI	250.91	3RAD	284.85	4JUO	325.43
1YTF	187.87	2GII	219.01	3E42	252.18	1CDW	288.52	1VTL	329.52
1BGB	188.71	1B94	220.46	2HT0	254.80	1NVP	290.13	4EUW	329.98
1F0O	188.98	2ATA	221.39	2GIJ	256.89	1JWL	290.27	3GQ5	334.99
3N78	189.56	1EFA	221.67	1QN4	259.03	3E41	291.38	3LWH	335.46
3U2B	192.24	2F5O	224.13	3RAF	259.33	3K9F	295.79	3QX3	342.27
3PVI	192.88	3RAE	224.89	3G38	261.36	3KSA	296.94	1CA5	343.72
3AZI	195.92	1PVI	227.69	3SWP	261.40	3E3Y	297.51	3OQG	345.33

Continued on the next page

Table B.1 – Continued from the previous page

PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T	PDB ID	E_T
1WTQ	352.91	2FKH	383.96	1QN8	473.24	4G0U	530.85	316D	642.08
1SX8	353.18	1AZP	384.63	1TW8	479.12	1QBJ	532.08	3IRR	648.56
1WTO	358.88	1KC6	393.60	2ACJ	483.80	3EYI	554.85	3IRQ	661.92
4HIV	359.79	1WD1	398.59	2GIE	485.94	1QNC	559.72	3UBT	728.74
1LBG	360.67	1QN7	407.47	1QN9	486.54	1QNB	560.51	2D55	732.10
1AZQ	368.00	1QN6	415.04	1XHV	495.96	1QNA	562.01	209D	788.85
3KXT	368.86	3VH0	442.08	2HEO	521.65	4G0W	590.71	4KA4	792.34
1D02	369.63	1QN5	452.68	3NDH	522.78	4G0V	600.73	2ADW	1416.07
4J3N	376.37	1XHU	462.59	1QN3	523.49	2XCT	620.84	3GO3	1577.16
3OR3	378.48	1TX3	466.64	1SFU	530.73	3F21	622.72	1XVR	1938.14
3LWI	383.46								

The end