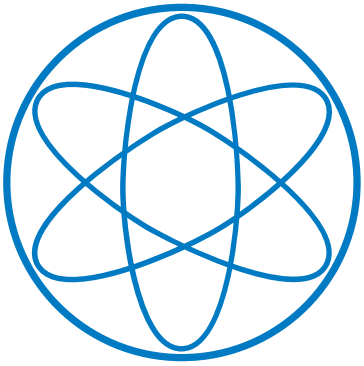
PHYSIK-DEPARTMENT



The Mechanics of DNA: Theoretical and Computational Studies

Dissertation by
Korbinian Bernhard Matthias Liebl

TECHNISCHE UNIVERSITÄT MÜNCHEN

Physik Department der Technischen Universität München
Lehrstuhl für Theoretische Biophysik

The Mechanics of DNA: Theoretical and Computational Studies

Korbinian Bernhard Matthias Liebl

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Franz Pfeiffer

Prüfer der Dissertation:

1. Prof. Dr. Martin Zacharias
2. Prof. Dr. Ulrich Gerland

Die Dissertation wurde am 30.09.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 20.11.2020 angenommen.

Contents

1 Introduction **1**
- 1.1 Cellular Life in the Context of Physical Laws . . . 1
- 1.2 Overview . . . 3
- 1.3 Bibliography . . . 5

2 The Structure of DNA **7**
- 2.1 From the Chemical Structure to the Double Helix . . . 7
- 2.2 Base-Pair Geometry . . . 9
- 2.3 The Backbone Structure of Double-Stranded DNA . . . 9
- 2.4 Sugar Puckering . . . 11
- 2.5 Bibliography . . . 13

3 Theoretical Background **17**
- 3.1 Principles of Statistical Mechanics . . . 17
- 3.2 Molecular Dynamics Simulations . . . 18
  - 3.2.1 General Remarks . . . 18
  - 3.2.2 Force Fields . . . 19
  - 3.2.3 Integrating Newton's Equations of Motion . . . 20
  - 3.2.4 Temperature and Pressure Coupling . . . 20
  - 3.2.5 Periodic Boundary Conditions and Electrostatic Interactions . . . 21
  - 3.2.6 Umbrella Sampling and Hamiltonian Replica Exchange . . . 23
  - 3.2.7 Weighted Histogram Analysis Method . . . 24
  - 3.2.8 Free Energy Perturbation and Thermodynamic Integration . . . 25
- 3.3 Quantum Chemistry . . . 26
  - 3.3.1 Hartree-Fock Method . . . 26
  - 3.3.2 Møller-Plesset Perturbation Theory . . . 28
  - 3.3.3 Density Functional Theory . . . 29
  - 3.3.4 Technical Concepts . . . 31
- 3.4 Bibliography . . . 33

4 Unwinding-Induced Melting of Double-Stranded DNA **39**
- 4.1 Introduction . . . 39
- 4.2 Materials and Methods . . . 40
  - 4.2.1 DNA Structures and Equilibration . . . 40
  - 4.2.2 Cylindrical Restraints . . . 40
  - 4.2.3 HREUS Simulations . . . 41

4.3 Results and Discussion . . . 43
4.3.1 Free Energy Changes related to DNA Unwinding and Comparison to Harmonic Models . . . 43
4.3.2 Free Energy Change upon DNA Unwinding beyond the Harmonic Regime . . . 46
4.3.3 Torque Induced Helical Conformations . . . 47
4.3.4 Changes in the Backbone Structure . . . 49
4.3.5 Relaxation of Denatured DNA to Intact B-DNA . . . 52
4.4 Conclusion . . . 53
4.5 Bibliography . . . 55

5 How Methyl-Sugar Interactions Determine DNA Structure and Flexibility 59
5.1 Introduction . . . 59
5.2 Materials and Methods . . . 60
5.2.1 Force Field Modifications . . . 60
5.2.2 Simulation Setup and Equilibration . . . 61
5.3 Results and Discussion . . . 62
5.3.1 How the Methyl Group Affects DNA's Backbone . . . 62
5.3.2 Why Methyl-Sugar Clashes Trigger BI Promiscuity . . . 63
5.3.3 Methyl-Sugar Clashes Influence DNA's Global Structure and Flexibility . . . 68
5.4 Conclusion . . . 70
5.5 Bibliography . . . 73

6 How Global DNA Unwinding Causes Non-uniform Stress Distribution and Melting of DNA 77
6.1 Introduction . . . 77
6.2 Materials and Methods . . . 78
6.2.1 Simulation Setup . . . 78
6.2.2 Calculation of Stiffness and Covariance Matrices . . . 79
6.2.3 Computation of Geometric Parameters through Rigid-Body Transformation . . . 79
6.2.4 Specification of Base-Pair Fraying . . . 80
6.3 Results and Discussion . . . 80
6.3.1 DNA as a Set of Harmonic Oscillators . . . 80
6.3.2 Local Bimodality in the Twist Distribution of dsDNA . . . 81
6.3.3 Structural Changes in dsDNA upon Global Unwinding . . . 82
6.3.4 Redistribution of Elastic Energy upon Global Unwinding . . . 86
6.3.5 Calculation of Free Energy Changes . . . 89
6.3.6 Modeling Melting Behavior upon Unwinding by an Ising Model . . . 89
6.4 Conclusion . . . 92
6.5 Bibliography . . . 95

7 Coarse-graining DNA: A Multivariate Ising Model 101
7.1 Introduction . . . 101
7.2 Materials and Methods . . . 102
7.2.1 MD Simulations of the ABC and Test Sequences . . . 102
7.2.2 Parameterization and Implementation of the Ising Model . . . 103
7.2.3 MD Simulations of the Papillomavirus E2 Systems and Nucleosome Complexes . . . 103
7.3 Results . . . 104
7.3.1 Splitting DNA's Mechanical Information into a Subset of Tetranucleotides . . . 104
7.3.2 Inclusion of Multimodality . . . 105
7.3.3 Ising Model . . . 106
7.3.4 Free Energy Profiles . . . 108
7.3.5 Application to DNA-Papillomavirus Complexes . . . 109
7.3.6 Deformation Energies in Nucleosome Complexes - A-tract Depletion 113
7.4 Conclusion . . . 114
7.5 Bibliography . . . 117

8 Tumuc1: Parameterization of a New DNA Force Field 123
8.1 Introduction . . . 123
8.2 Results and Discussion . . . 124
8.2.1 Parameterization . . . 124
8.2.2 Simulation Results . . . 132
8.3 Materials and Methods . . . 148
8.3.1 Parameterization . . . 148
8.3.2 MD Simulations . . . 151
8.3.3 Analysis of MD Trajectories . . . 152
8.4 Conclusion . . . 153
8.5 Bibliography . . . 157

9 The Geometry of Blunt-end and Base-Pair Stacking 165
9.1 Introduction . . . 165
9.2 Materials and Methods . . . 166
9.2.1 MD Protocol . . . 166
9.2.2 QM Calculations . . . 167
9.3 Results and Discussion . . . 167
9.3.1 MD Simulations of Blunt-end Base-Pair Steps . . . 167
9.3.2 QM Calculations of the Potential Energy of Base-Pair Stacking . 172
9.4 Conclusion . . . 174
9.5 Bibliography . . . 177

10 Calculation of the Absolute Binding Free Energy for the Cren7-DNA Complex 181
10.1 Introduction . . . 181

10.2 Materials and Methods 182
10.2.1 Advanced Sampling Method for the Computation of Absolute Binding Free Energies 182
10.2.2 MD Simulations 184
10.3 Results and Discussion 186
10.3.1 Binding Free Energy 186
10.3.2 Deformation Energy of the DNA upon Binding 189
10.4 Conclusion 190
10.5 Bibliography 193

11 Outlook **197**

Acknowledgments **199**

Und jedem Anfang wohnt ein Zauber inne
- Hermann Hesse, Stufen

1 Introduction

1.1 Cellular Life in the Context of Physical Laws

In a human being's lifetime more than a light-year of DNA is replicated. Thereby, a replication fork operates at a surprisingly high speed of ~ 50 nucleotides per second and with an incredibly low error rate of $1 : 10^8$. This happens in all of our $\sim 10^{13}$ cells, more precisely, in the nuclei of our cells. A cell nucleus has a diameter of a few micrometers, but it stores DNA with a total length of 2 meters [1, 2]. Nature overcomes this apparent discrepancy by tightly wrapping the DNA around histone-proteins and condensing the resulting nucleosome complexes into chromosomes. Besides, our hereditary information has to deal with permanent attacks stemming from different sources: The DNA of a single human cell is harmed by approximately 70000 lesions per day [3]. Also here, our cells have developed a remarkable arsenal of repair mechanisms that enable the recognition and repair of damages at high fidelity. We therefore have discovered the fascinating beauty of cellular life with respect to both, architecture and dynamic functionality. Part of this beauty, which drives us as scientists, is our lack of understanding. How can all this be carried out at such a high level of precision and speed? What are the physical rules behind these processes which are basal to life?
What is certain about life, however, is that it is made of atoms, that it happens in a thermodynamic ensemble, and that, in Feynman's words, it can be understood as 'jiggling' and 'wiggling' of these atoms. That we can conceptualize cellular life as concerted motions of atoms is the basic motivation behind my thesis for two reasons:
Firstly, it sets the topic. The double-helical structure of DNA was discovered by Watson and Crick in 1953 [4]. Fifty years later, the entire human genome was sequenced [5]. Though great progress has hence been made in the field of molecular biology, our understanding is limited to a rather static picture. But it is essential to life, that DNA also acts dynamically, that it can be transcribed and replicated. This requires ruptures, local melting of specific DNA sequences to facilitate access to our hereditary information. Moreover, many cellular functions are triggered by the binding of proteins to target DNA sequences. All these dynamic processes cause deformations of the DNA, and thus, we can only begin to understand them by understanding the mechanics of DNA first. An example for my contribution to this topic is the development of a coarse-grained model for DNA, presented in chapter 7. Here, we express the deformation free energy for a given double-stranded DNA structure by an Ising model, in which we consider two distinct conformational subspaces for every base-pair step. These two subspaces represent the BI and BII backbone configurations and are described by a multivariate harmonic

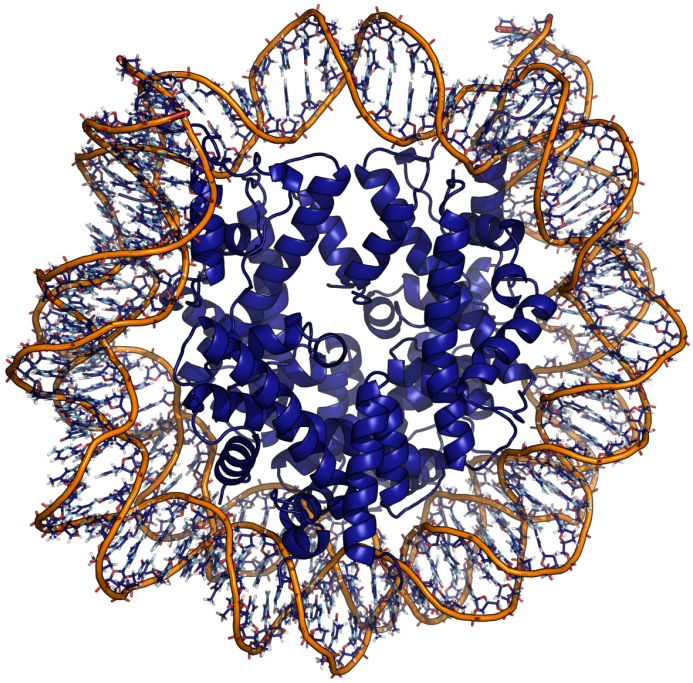


Figure 1.1: Nucleosome complex, consisting of double-stranded DNA (the two backbone strands are highlighted in orange) wrapped around the histone-proteins (shown in blue, cartoon representation) [6]. This complex represents a fundamental mechanism how DNA is compacted and accommodated in the cell nucleus. Formation of this complex requires strong mechanical deformation of the DNA and is hence dependent on DNA's sequence. Note that the figures in this thesis have been created with pymol and VMD [7, 8].

approximation, whereby we account for correlation to adjacent base-pair steps. In this way, we have been able to derive a first coarse-grained DNA model that captures both, nearest-neighbor coupling and multimodality. Our model is parameterized from MD simulations on a set of DNA sequences composed of all possible 136 tetranucleotides. Thus, we can quantify sequence-dependent free energies afforded by proteins to deform the DNA. We have applied our model to compute deformation free energies for DNA in the nucleosome complex (Fig 1.1) in the context of DNA's sequence. This explains for instance why specific sequences, poly(dA/dT)-tracts, avoid nucleosome formation, as they are mechanically too resistant to undergo such deformations.

Secondly, it sets the method. In the past, we have learned a lot about DNA's mechanics from experimental methods, such as optical tweezers and atomic force microscopy [9]. However, these techniques come with limited insight. The only microscopes with which we can observe the dynamics at the atomic level and femtosecond time resolution are those in our computers. The computational microscope utilized in my thesis is Molecular Dynamics simulations, and we have made hard efforts to contribute to the development of this method. From high-level quantum mechanical calculations on DNA subsystems, we

have inferred a description for the interaction between all atoms of the DNA. Subjecting now the description for DNA's potential energy to Newton's laws of motion allows one to propagate a DNA system in time. Basically, this is what we understand with Molecular Dynamics (MD) simulations and how we try to find answers to questions of cellular life. Within my PhD thesis, I have conducted a large number of Molecular Dynamics simulations, and will share in the following my insights on many aspects of the mechanics of DNA.

1.2 Overview

At the beginning of my thesis, I give an overview of DNA's structure and various concepts to describe it. Next, fundamental principles of statistical mechanics are outlined, followed by an introduction to Molecular Dynamics simulations and an explanation of related techniques. We then turn to Quantum Chemistry, where the theories behind our conducted quantum mechanical calculations are summarized.
Chapter 4 has been published in similar form in J. Phys. Chem. B, 121(49), 11019-11030, 2017 [10]. It deals with the question of how local unwinding stress is absorbed by the DNA and highlights the role of global restraining. Furthermore, we characterize resulting phase transitions and discuss their sequence dependence.
In chapter 5, we elucidate sterical effects which impact DNA's local conformational flexibility: Clashes between methyl groups (as present in the thymine or C5-methylated cytosine base) and sugar rings suppress the population of the BII backbone substate. We show that this also causes sequence-dependent deformabilities. This study has been published in Nucleic Acids Res., 47(3):1132-1140, 2018 [11].
Chapter 6 deals with global unwinding of double-stranded DNA, and has been published in similar form in Plos One, 15(5):e0232976, 2020 [12]. Here, we point out that global stress causes a non-uniform distribution of the deformation energy. For lower stress levels, this serves as a protection mechanism for AT-rich sequences. High torsional stress, instead, is fully absorbed by the melting of a promoter-like TATA box segment. By means of our free energy simulations, we give novel atomistic insights into the opening of promoter boxes and draw comparison to an Ising model, which we have established from experimental values.
Derivation of a coarse-grained model for double-stranded DNA is the subject of chapter 7. Thereby, we describe DNA's deformability by combining a multivariate harmonic approximation with an Ising model. In this way, we account for both effects, nearest-neighbor correlation and multimodality, which is a consequence of backbone polymorphism. Our multivariate Ising model is more precise than the conventional, harmonic approximation, as the overestimation of free energies is reduced. Based on our new model, we show that deformation free energies are in excellent correlation with experimentally derived binding affinities for the papillomavirus-E2 protein. Furthermore, we quantify sequence-dependent deformation free energies arising in the nucleosome complex. We demonstrate that the incorporation of poly(dA/dT) sequences is mechanically

unfavorable, which serves as explanation for their reluctance to undergo nucleosome formation.
In chapter 8, we derive an entirely new DNA force field from quantum mechanical calculations. We have subjected our new force field ('Tumuc1') to many different DNA systems: For B-DNA we find an accurate description of our force field, where most parameters are in good agreement with experimental data. Especially for the dihedral angles, we achieve significant improvements over the bsc1 force field. Besides, Tumuc1 is also robust for other DNA systems and excels in hybridization of single-strands and hairpin folding.
Chapter 9 deals with the structure of DNA blunt-ends, which can be thought of as double-strand breaks. We study the structure and conformational flexibility of these motives with MD simulations and QM calculations. Our main point here is that blunt-end stacks adopt a negatively twisted configuration, which is possibly of high biological and technical relevance.
In chapter 10, we quantify the binding free energy for the Cren7/DNA complex with advanced free energy simulations. Our estimation for the binding free energy strongly overestimates experimental values, which may be a result of an insufficient electrostatic description by the bsc1 force field. In addition, we compute the configurational contributions to the binding free energy. Thereby, we emphasize that the Cren7 protein experiences only little structural changes upon binding, but it strongly deforms the DNA.

1.3 Bibliography

[1] Franziska Bleichert, Michael R. Botchan, and James M. Berger. Mechanisms for initiating cellular DNA replication. *Science*, 355(6327), 2017.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. DNA Replication, Repair, and Recombination. In *Molecular Biology of the Cell*, chapter 5, pages 263–328. Garland Science, Taylor & Francis Group, 5$^{\text{th}}$ edition, 2008.

[3] Anthony Tubbs and André Nussenzweig. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell*, 168(4):644–656, 2017.

[4] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.

[5] Francis S. Collins, Michael Morgan, and Aristides Patrinos. The Human Genome Project: Lessons from Large-Scale Biology. *Science*, 300(5617):286–290, 2003.

[6] Yasuo Tsunaka, Naoko Kajimura, Shin-ichi Tate, and Kosuke Morikawa. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res.*, 33(10):3424–3434, 2005.

[7] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version∼1.8. 2015.

[8] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. *J. Molec. Graphics*, 14(1):33–38, 1996.

[9] Carlos Bustamante, Steven B Smith, Jan Liphardt, and Doug Smith. Single-molecule studies of DNA mechanics. *Current Opinion in Structural Biology*, 10(3):279–285, 2000.

[10] Korbinian Liebl and Martin Zacharias. Unwinding Induced Melting of Double-Stranded DNA Studied by Free Energy Simulations. *J. Phys. Chem. B*, 121(49):11019–11030, 2017.

[11] Korbinian Liebl and Martin Zacharias. How methyl-sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.*, 47(3):1132–1140, 2018.

[12] Korbinian Liebl and Martin Zacharias. How global DNA unwinding causes non-uniform stress distribution and melting of DNA. *PLOS ONE*, 15(5):e0232976, 2020.

2 The Structure of DNA

2.1 From the Chemical Structure to the Double Helix

The DNA molecule is composed of nucleotides, which consist of a nucleobase, a deoxyribose sugar and a phosphate backbone (Fig 2.1). In DNA, there naturally exist four different nucleobases, adenine (A), cytosine (C), guanine (G) and thymine (T), which are linked to the sugar ring by a glycosidic bond with the C1' atom. Two nucleotides can be merged through the phosphate group, which forms a phosphodiester bond to the C3' atom of the first and the C5' atom of the second nucleotide. In this way, nucleotides build a polymer, a DNA single strand, with a $5' \rightarrow 3'$ polarity [1].

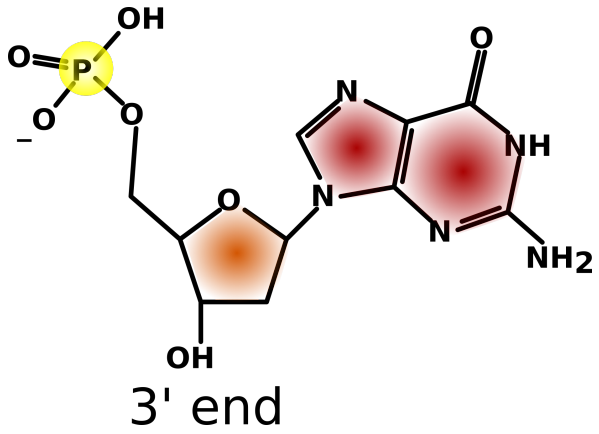


Figure 2.1: Chemical structure of a guanine-nucleotide composed of a phosphate backbone (yellow), a sugar ring (orange) and a nucleobase (red). Ref [2] served as template for this figure.

The DNA double helix emerges from the hybridization of two single strands polarized in opposite direction. Thereby, bases of the opposite strands are paired together in a complementary way: Adenine is generally paired with thymine through two, and cytosine with guanine through three hydrogen bonds (Fig 2.2) [3, 4].

The DNA double helix occurs in different topologies, such as A- and B-DNA. A-DNA has a larger diameter but reduced winding and helical extension than the classical B-DNA, which is depicted in Fig 2.3.

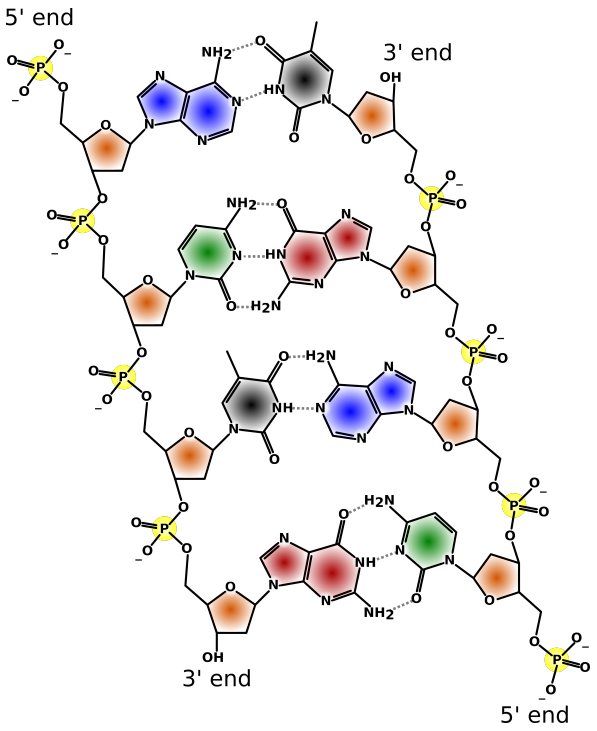


Figure 2.2: Chemical structure of a DNA double strand. Adenine (blue) pairs with thymine (black), and guanine (red) with cytosine (green). The figure has been taken from Ref [2]. Slight modifications have been made.

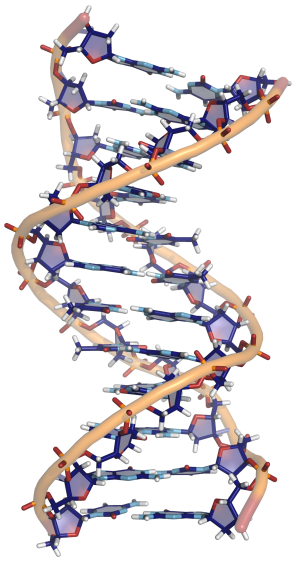


Figure 2.3: Structure of the DNA double helix in B-form, shown in Cartoon-representation. The backbone strands are highlighted as orange spline curves.

B-DNA has a helical extension of $\sim 3.4\,\mathring{A}$ and a winding of $\sim 34.5°$ per base-pair step [5–7]. Another important feature of DNA is its groove asymmetry: The appearance of a minor and major groove has strong mechanical and biological consequences and results from the non-diametrical arrangement of the glycosidic bonds [8–10].
Moreover, DNA's double-helical structure is stabilized through base-stacking, meaning that the aromatic bases arrange in a parallel geometry. In this configuration, the base-pairs experience exchange-repulsion, attractive London-dispersion and electrostatic interactions, with the latter two causing the stability of the complex [11, 12].

2.2 Base-Pair Geometry

DNA's double-helical structure is commonly quantified through a rigid-body model, where each base is represented as a bead [13–15]. The pairing of two bases is then described by three translational (shear, stretch, stagger) and three rotational parameters (buckle, propeller, opening), which map one bead onto the bead corresponding to its paired base. The stacking-geometry between two successive base-pairs is computed by rigid-body transformations of 'mean'-beads of the base-pairs, and yields again three translational (shift, slide, rise) and three rotational parameters (tilt, roll, twist). Note that a DNA double helix with N base-pairs contains $12N - 6$ degrees of freedom in this local frame. Alternatively, DNA's structure can be described by a set of helical parameters (inclination, tip, x-displacement, y-displacement, helical-rise, helical-twist). These parameters are calculated by interpolating a mean-helical axis into DNA's structure, whereby helical-rise and helical-twist are very similar to their local pendants. All local and helical parameters are depicted in Fig 2.4.

2.3 The Backbone Structure of Double-Stranded DNA

DNA's base-pair geometry is strongly correlated with the backbone strands. The configuration of the backbone is determined by the dihedral angles shown in Fig 2.5.

The backbone dihedral angles represent a highly coupled system, which is typically described through ϵ/ζ and α/γ conformers. Note that the ϵ and ζ dihedral angles define the prominent BI and BII substates (Fig 2.6):

$$\epsilon - \zeta < 0 : \text{BI} \tag{2.1}$$

$$\epsilon - \zeta > 0 : \text{BII} \tag{2.2}$$

The population of these states is sequence-dependent, which is elucidated in chapter 5. In general, however, a strand is preferentially in the BI conformation [18]. Base-pair steps with Watson- and Crick-strand in BI configuration usually adopt stacking

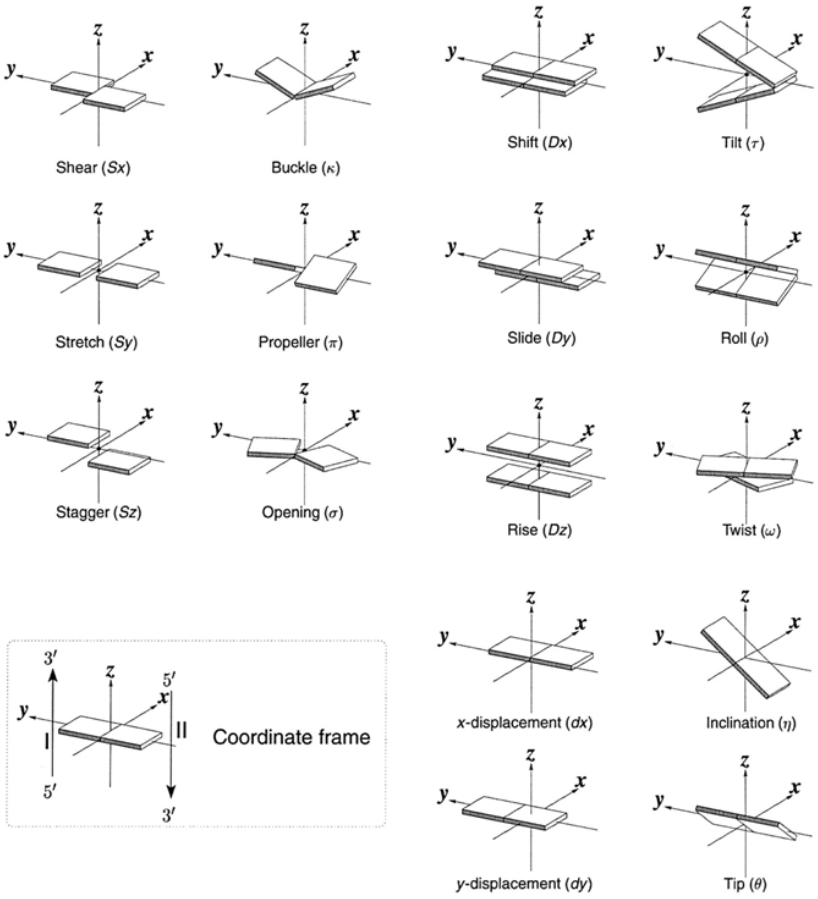


Figure 2.4: Parameters describing DNA's base-pair geometry [16].

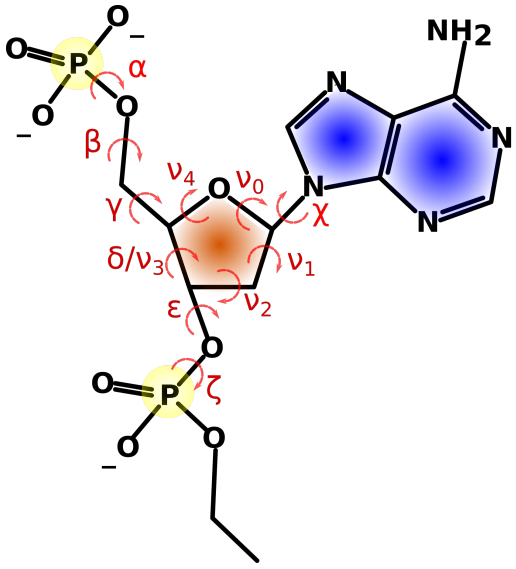


Figure 2.5: Extract of a DNA strand, showing all backbone and sugar dihedral angles. Ref [2] served as template for this figure.

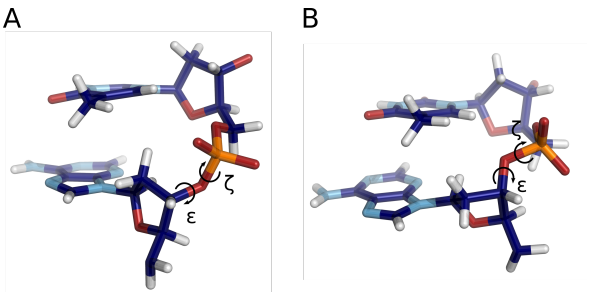


Figure 2.6: Snapshots of an ApT-step in BI- (A) and in BII-conformation (B). These conformations are determined by the ϵ and ζ dihedral angles [17].

geometries with smaller twist and slide but higher roll compared to conformations with one of the two strands in the BII substates (both in BII configuration is very unlikely) [19, 20]. Thus, the BI and BII states give rise to bimodal distributions in some of the base-pair parameters [21]. Furthermore, it has been argued that these substates are also implicated in readout-mechanisms by proteins [22–24].
The remaining backbone dihedral angles are classified in three substates:

$$g+ : 0^\circ - 120^\circ \tag{2.3}$$
$$t : 120^\circ - 240^\circ \tag{2.4}$$
$$g- : 240^\circ - 360^\circ \tag{2.5}$$

The α/γ dihedral angles occupy the $g-/g+$ substate in regular B-DNA. A transition to another substate causes strong local distortions in the DNA structure [25–27]. For the parameterization of a DNA force field, dominant population of g-/g+ for α/γ is therefore essential.

2.4 Sugar Puckering

Sugar puckering is a concept to quantify the non-planarity of DNA's sugar ring. Thereby, the conformational space of the sugar ring is mapped onto a polar coordinate system spanned by a phase angle P [18, 28]:

$$\tan(P) = \frac{\nu_4 + \nu_1 - \nu_3 - \nu_0}{2\nu_2 \cdot (\sin(36^\circ) + \sin(72^\circ))} \tag{2.6}$$

and an amplitude [18, 28]

$$\tau_m = \frac{\nu_2}{\cos(P)}. \tag{2.7}$$

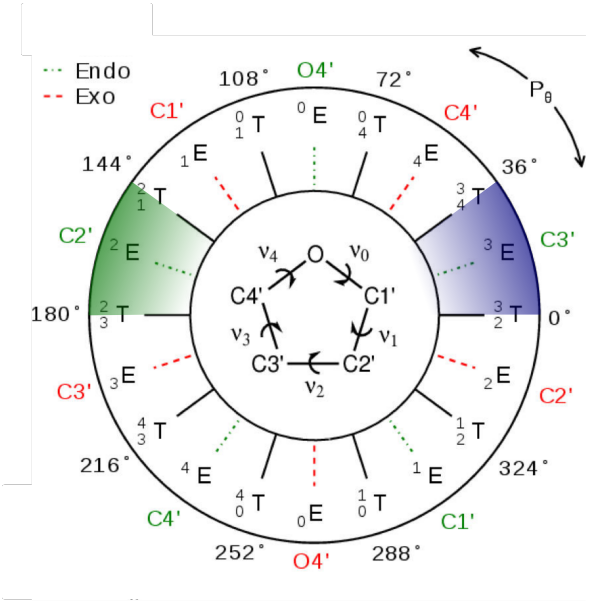


Figure 2.7: Conformational states of the sugar ring, defined by the pseudorotational phase angle P. The C2'-endo and C3'-endo subspaces are colored in green and blue, respectively. The figure has been taken from Ref [29], slight modifications have been made.

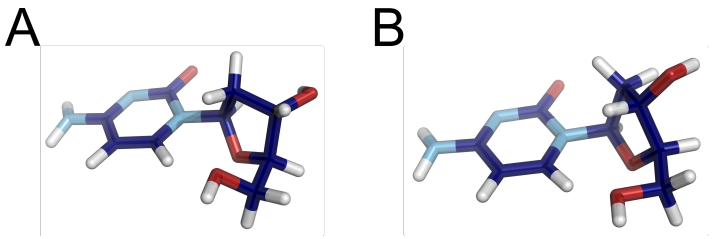


Figure 2.8: Exemplary structures of a cytosine-nucleoside in C2'-endo (A) and in C3'-endo conformation (B).

The state of the sugar ring is determined only by the phase angle P, see Fig 2.7.

The most important states are the C2'-endo ($144° \leq P \leq 180°$, Fig 2.8 A) and the C3'-endo ($0° \leq P \leq 36°$, Fig 2.8 B) states . In this nomenclature, 'endo' represents states with the corresponding atom lying outside of the sugar ring on the side of the C4'-C5' bond, whereas 'exo' denotes opposite positioning. B-DNA usually adopts the C2'-endo and A-DNA the C3'-endo state. Importantly, sugar puckering is correlated to the backbone dihedrals, which is also reflected in the absence of BII-states in A-DNA [18].

2.5 Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Cell Chemistry and Biosynthesis. In *Molecular Biology of the Cell*, chapter 2, pages 45–124. Garland Science, Taylor & Francis Group, 5th edition, 2008.

[2] Madprime. Chemical structure of DNA. https://en.wikipedia.org/wiki/DNA, Accessed: 15.9.2020.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. DNA, Chromosomes and Genomes. In *Molecular Biology of the Cell*, chapter 4, pages 195–262. Garland Science, Taylor & Francis Group, 5th edition, 2008.

[4] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.

[5] Richard Wing, Horace Drew, Tsunehiro Takano, Chris Broka, Shoji Tanaka, Keiichi Itakura, and Richard E. Dickerson. Crystal structure analysis of a complete turn of B-DNA. *Nature*, 287(5784):755–758, 1980.

[6] J C Wang. Helical repeat of DNA in solution. *Proc. Natl. Acad. Sci. USA*, 76(1):200–203, 1979.

[7] M Levitt. How many base-pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations. *Proc. Natl. Acad. Sci. USA*, 75(2):640–644, 1978.

[8] J. M. Berg, J. L. Tymoczko, and L. Stryer. DNA, RNA und der Fluss der genetischen Information. In *Biochemie*, chapter 4, pages 110–141. Spektrum Akademischer Verlag, 6th edition, 2007, corrected reprint: 2010.

[9] J. F. Marko and E. D. Siggia. Bending and twisting elasticity of DNA. *Macromolecules*, 27(4):981–988, 1994.

[10] Stefanos K. Nomidis, Franziska Kriegel, Willem Vanderlinden, Jan Lipfert, and Enrico Carlon. Twist-Bend Coupling and the Torsional Response of Double-Stranded DNA. *Phys. Rev. Lett.*, 118(21):217801, 2017.

[11] J. Šponer, P. Jurečka, I. Marchan, F. J. Luque, M. Orozco, and P. Hobza. Nature of Base Stacking: Reference Quantum-Chemical Stacking Energies in Ten Unique B-DNA Base-Pair Steps. *Chem. Eur. J.*, 12(10):2854–2865, 2006.

[12] Jiří Šponer, Judit E. Šponer, Arnošt Mládek, Petr Jurečka, Pavel Banáš, and Michal Otyepka. Nature and magnitude of aromatic base stacking in DNA and RNA: Quantum chemistry, molecular mechanics, and experiment. *Biopolymers*, 99(12):978–988, 2013.

[13] Wilma K Olson, Manju Bansal, Stephen K Burley, Richard E Dickerson, Mark Gerstein, Stephen C Harvey, Udo Heinemann, Xiang-Jun Lu, Stephen Neidle, Zippora Shakked, Heinz Sklenar, Masashi Suzuki, Chang-Shung Tung, Eric Westhof, Cynthia Wolberger, and Helen M Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology*, 313(1):229–237, 2001.

[14] Christophe Blanchet, Marco Pasi, Krystyna Zakrzewska, and Richard Lavery. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.*, 39(Web Server issue):W68–W73, 2011.

[15] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37(17):5917–5929, 2009.

[16] Xiang-Jun Lu and Wilma K. Olsen. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, 31(17):5108–5121, 2003.

[17] Korbinian Liebl and Martin Zacharias. How methyl-sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.*, 47(3):1132–1140, 2018.

[18] Bohdan Schneider and Helen M. Berman. Basics of Nucleic Acid Structure. In *Computational Studies of RNA and DNA (Challenges and Advances in Computational Chemistry and Physics vol 2)*, chapter 1, pages 1–44. Springer, 2006.

[19] D. Djuranovic and B. Hartmann. DNA fine structure and dynamics in crystals and in solution: The impact of BI/BII backbone conformations. *Biopolymers*, 73(3):356–368, 2004.

[20] Marie Zgarbová, Petr Jurečka, Filip Lankaš, Thomas E. Cheatham, Jiří Šponer, and Michal Otyepka. Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Inf. Model.*, 57(2):275–287, 2017.

[21] Tomáš Dršata, Alberto Pérez, Modesto Orozco, Alexandre V. Morozov, Jiří Šponer, and Filip Lankaš. Structure, stiffness and substates of the dickerson-drew dodecamer. *J. Chem. Theory Comput.*, 9(1):707–721, 2013.

[22] Rashmi S. Hegde, Steven R. Grossman, Laimonis A. Laimins, and Paul B. Sigler. Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DMA-binding domain bound to its DNA target. *Nature*, 359(6395):505–512, 1992.

[23] Fajar R. Wibowo, Christine Rauch, Michael Trieb, and Klaus R. Liedl. M.TaqI facilitates the base flipping via an unusual DNA backbone conformation. *Biopolymers*, 79(3):128–138, 2005.

[24] James C. Robertson and Thomas E Cheatham. DNA Backbone BI/BII Distribution and Dynamics in E2 Protein-Bound Environment Determined by Molecular Dynamics Simulations. *J. Phys. Chem. B*, 119(44):14111–14119, 2015.

[25] Péter Várnai, Dragana Djuranovic, Richard Lavery, and Brigitte Hartmann. α/γ Transitions in the B-DNA backbone. *Nucleic Acids Res.*, 30(24):5398–5406, 2002.

[26] Mahmut Kara and Martin Zacharias. Influence of 8-oxoguanosine on the fine structure of DNA studied with biasing-potential replica exchange simulations. *Biophysical journal*, 104(5):1089–1097, 2013.

[27] Rodrigo Galindo-Murillo, James C. Robertson, Marie Zgarbová, Jiří Šponer, Michal Otyepka, Petr Jurečka, and Thomas E. Cheatham. Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.*, 12(8):4114–4127, 2016.

[28] C. Altona and M. Sundaralingam. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. A New Description Using the Concept of Pseudorotation. *Journal of the American Chemical Society*, 94(23):8205–8212, 1972.

[29] Ming Huang, Timothy J. Giese, Tai-Sung Lee, and Darrin M. York. Improvement of DNA and RNA Sugar Pucker Profiles from Semiempirical Quantum Methods. *J. Chem. Theory Comput.*, 10(4):1538–1545, 2014.

3 Theoretical Background

3.1 Principles of Statistical Mechanics

The phase space of a N-particle system is defined by the generalized coordinates $\mathbf{q}_1, ..., \mathbf{q}_N = q_1, q_2, .., q_{3N-1}, q_{3N}$ and their conjugate momenta $\mathbf{p}_1, ..., \mathbf{p}_N = p_1, p_2, ..., p_{3N-1}, p_{3N}$. The probability to find the system at time t in the surrounding of a certain point $(\mathbf{q}_1, ..., \mathbf{q}_N, \mathbf{p}_1, .., \mathbf{p}_N)$ is given by $\rho(\mathbf{q}_1, ..., \mathbf{q}_N, \mathbf{p}_1, .., \mathbf{p}_N, t) \cdot \mathrm{d}^N\mathbf{q} \cdot \mathrm{d}^N\mathbf{p}$, whereby we have introduced the ensemble distribution function $\rho(\mathbf{q}_1, ..., \mathbf{q}_N, \mathbf{p}_1, .., \mathbf{p}_N, t)$. As stated by the Liouville equation, the ensemble distribution function is constant in time [1]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho = \frac{\partial\rho}{\partial t} + \{\rho, H\} = 0. \tag{3.1}$$

In thermodynamic equilibrium, observables do not explicitly depend on time and hence also not the ensemble distribution function. With $\frac{\partial\rho}{\partial t} = 0$, we then obtain $\{\rho, H\} = 0$. Thus, ρ is a function of H [2]. In the microcanonical ensemble, the system evolves according to Hamilton's equations of motion, implying also conservation of energy: $H(\mathbf{q}_1, .., \mathbf{q}_N, \mathbf{p}_1, ..., \mathbf{p}_N) = E$. This condition is fulfilled by $\rho \sim \delta(H(\mathbf{q}_1, .., \mathbf{q}_N, \mathbf{p}_1, ..., \mathbf{p}_N) - E)$, which also assumes that every microscopic state (defined by a phase space vector $(\mathbf{q}_1, .., \mathbf{q}_N, \mathbf{p}_1, .., \mathbf{p}_N)$) is equally probable [3].
Now, the microcanonical partition sum $\Omega(N, V, E)$ results from integration over the phase space [3, 4]:

$$\Omega(N, V, E) = \frac{\Delta E}{N! h^{3N}} \int \mathrm{d}^N\mathbf{q} \cdot \mathrm{d}^N\mathbf{p}\, \delta(H(\mathbf{q}_1, .., \mathbf{q}_N, \mathbf{p}_1, .., \mathbf{p}_N) - E). \tag{3.2}$$

Thereby, we have accounted for Heisenberg's uncertainty relation $\Delta q \cdot \Delta p = h$. ΔE represents the uncertainty of measuring E and is without effect on thermodynamic properties, whereas the factor 1/N! corrects for the particles being indistinguishable. From $\Omega(E, V, N)$, we can calculate all thermodynamic quantities. In biology, however, most systems do not obey the conservation of energy. Constant temperature ensembles, such as the canonical (NVT) and isothermal-isobaric (NPT) ensemble, are more realistic. The canonical ensemble is derived by considering two microcanonical ensembles, where the larger one acts as a thermal reservoir [4]. The canonical partition sum emerges then as the Laplace-transform of the microcanonical partition sum [5, 6]:

$$Z(N,V,T) = \frac{1}{\Delta E}\int \mathrm{d}E \exp[-\beta E]\cdot\Omega(N,V,E) = \frac{1}{N!h^{3N}}\int \mathrm{d}^N\mathbf{q}\cdot\mathrm{d}^N\mathbf{p}\exp[-\beta H]. \quad (3.3)$$

We hence obtain $\rho \sim \exp[-\beta H]$ in the canonical ensemble, and the free energy is given by $F = -\frac{1}{\beta}\ln(Z)$. Note that the isothermal-isobaric partition sum, in turn, is the Laplace-transform of the canonical partition sum with respect to the volume [7]. The expectation value of any observable A can be calculated by its integration over the phase space weighted by the ensemble distribution function [8]:

$$< A >= \frac{1}{N!h^{3N}}\int \mathrm{d}^N\mathbf{q}\cdot\mathrm{d}^N\mathbf{p}\,\rho(\mathbf{q}_1,..,\mathbf{q}_N,\mathbf{p}_1,...,\mathbf{p}_N)\cdot A(\mathbf{q}_1,..,\mathbf{q}_N,\mathbf{p}_1,...,\mathbf{p}_N). \quad (3.4)$$

Importantly, the ensemble average of an observable is equal to its temporal average (measured for a sufficiently long time), as stated by the ergodic hypothesis [8]:

$$< A >= \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau \mathrm{d}t\, A(t). \quad (3.5)$$

3.2 Molecular Dynamics Simulations

3.2.1 General Remarks

The most basic decision for studying a N-particle system has to be made for the choice of the level of theory. Here, Molecular Dynamics (MD) simulations use a classical formalism, whereby the system is propagated in time by numerically solving Newton's equations of motion [9, 10]. The major reasonings behind this choice are the following: First, a quantum mechanical treatment is computationally too expensive for larger systems [11]. Second, nuclear and electronic motion can be separated according to the Born-Oppenheimer approximation [11]. Third, we neglect motions of the electrons, as they relax quickly upon motion of the nucleus, which we then approximate to move in an averaged electron density [12].
Recently, the classical description of biological systems is under criticism, as especially the classical treatment of hydrogen atoms is contestable. Indeed, Pereyaslavets and coworkers have shown that reproduction of essential properties of alkane systems is highly problematic on the basis of classical mechanics and that nuclear quantum effects play a major role in these systems [13]. However, several MD studies have also revealed excellent agreement with experimental studies, such as accurate folding behavior of proteins [14], the reproduction of elastic properties of DNA and RNA [15], and binding of proteins into the correct binding-site of double-stranded DNA [16]. Furthermore, nuclear

quantum effects can be easily incorporated into classical MD simulations by representing atoms as a polygon composed of pseudo-particles [17, 18].
In the following, we explain the basic principles behind Molecular Dynamics simulations. This includes the description of the potential energy landscape, the integration of the equations of motion and the treatment of boundary conditions and long-range electrostatic interactions. Moreover, we discuss the simulation in the NVT and NPT ensemble as well as advanced sampling methods.

3.2.2 Force Fields

The term 'force field' means the description of the potential energy $U(\mathbf{r}_1, ..., \mathbf{r}_N)$, whose gradients are the forces acting on the respective particles. MD simulations performed within this thesis rely on an additive force field [19]:

$$\begin{aligned} U(\mathbf{r}_1, ..., \mathbf{r}_N) = &\sum_{Bonds} k_d \cdot (d - d_0)^2 + \sum_{Angles} k_\theta \cdot (\theta - \theta_0)^2 + \\ + &\sum_{Dihedrals} \frac{V_n}{2} \cdot [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon \cdot r_{ij}} \right]. \end{aligned} \tag{3.6}$$

The first two terms on the right side model the bonds and angles between covalently bound atoms and use a harmonic form. Here, most force fields rely on empirical values for the parameters k_d, d_0, k_θ and θ_0. Dihedral angles are parameterized through a truncated Fourier series (commonly to third order). Note that this potential term does not have a direct physical justification, it is rather used to correct for local misdescriptions [20]. Inaccuracies in the parameterization of dihedrals lead to obvious structural artifacts in the simulations. Refinements of force fields have therefore mostly addressed the dihedral parameters [20–26]. The last sum models the non-bonded interactions. It consists of a short-range, repulsive (Pauli-exclusion) and a dispersion-related attractive interaction. Eventually, electrostatic interactions are described through a point-charge model.
Major attempts in improving this potential form include multipole expansions for the charge model and especially polarizations terms [27, 28]. Recently, it has been shown that the omission of polarization effects is highly problematic when multivalent ions are involved [29]. Nevertheless, polarizable force fields have not replaced the additive functional form of equation 3.6 from the status quo. Not only are polarizable force fields computationally more expensive, but also harder to parameterize. We find that clear examples, where a polarizable force field systematically outperforms standard force fields are still outstanding. Finally, we emphasize that the general difficulty of parameterizing Molecular Mechanics force fields can already be anticipated from the problems of cheap quantum mechanical methods [20].

3.2.3 Integrating Newton's Equations of Motion

As mentioned in the previous section, our representation of the internal energy $U(\mathbf{r}_1, ..., \mathbf{r}_N)$ enables the calculation of forces through $\mathbf{F}_i = -\nabla_i U(\mathbf{r}_1, ..., \mathbf{r}_N)$. According to Newton's second law, we can obtain the acceleration of the particles:

$$\mathbf{F}_i = m\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} \tag{3.7}$$

The equations of motion are then integrated numerically. In the following, we show the velocity Verlet algorithm [3, 30]. Here, updates in the coordinates are given by a Taylor expansion to second order:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \Delta t\mathbf{v}_i(t) + \frac{\Delta t^2}{2m}\mathbf{F}_i(t) \tag{3.8}$$

From another Taylor expansion, starting from $\mathbf{r}_i(t+\Delta t)$ and going back in time, we find for the update in velocities:

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \frac{\Delta t}{2 \cdot m_i}\left[\mathbf{F}_i(t) + \mathbf{F}_i(t + \Delta t)\right] \tag{3.9}$$

Intriguingly, the velocity Verlet algorithm can also be derived from Trotter expansion of the Liouville operator. Note that the velocity Verlet algorithm is time-reversible and symplectic, i.e. it preserves phase space [3, 31]. Vibrations of bonds between light atoms would require a subfemtosecond time step, however, employing holonomic, time-independent constraints allows a time step of $\Delta t = 2\,\mathrm{fs}$ [12, 32, 33].

3.2.4 Temperature and Pressure Coupling

The theoretical concepts presented so far allow us to simulate biological systems in the microcanonical (NVE) ensemble. Physiologically more sensible, however, are the NVT or NPT ensemble. In the following, we therefore briefly discuss methods to couple a thermostat and a barostat to a simulation system.
One of the most widely used thermostats is the Berendsen thermostat, which scales the velocities, $\mathbf{v}_i \mapsto \lambda \cdot \mathbf{v}_i$, with [34, 35]:

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \cdot \left(\frac{T_0}{T} - 1\right)\right]^{\frac{1}{2}}. \tag{3.10}$$

T_0 represents the reference and T the actual temperature. τ_T is the coupling constant, recommended values are $0.5\,\mathrm{ps} - 5.0\,\mathrm{ps}$. Larger values result in too low kinetic energy

fluctuations, whereas the choice of a lower τ_T risks running into flying ice cube artifacts through transferring kinetic energy from high frequency into low frequency modes [36, 37].

Analogously, pressure coupling is achieved by scaling of the box size and the coordinates: $\mathbf{r}_i \mapsto \mu \cdot \mathbf{r}_i$ and $l(t + \Delta t) = \mu l(t)$, whereby the pressure-scaling factor is expressed as a function of isothermal compressibility β, a coupling constant τ_P and the reference pressure P_0 and actual pressure P [34]:

$$\mu = \left[1 + \frac{\beta \Delta t}{\tau_P} \cdot (P - P_0)\right]^{\frac{1}{3}}. \tag{3.11}$$

From $P = k_B T \left(\frac{\partial \ln(Z)}{\partial V}\right)_{N,T}$ it is deducible, that we can obtain P as ensemble average [38]:

$$P = \frac{1}{3V} \left\langle \sum_i \left[\frac{\mathbf{p}_i^2}{m_i} + \sum_{j<i} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij}\right]\right\rangle. \tag{3.12}$$

Note that the second term is the virial, with $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{F}_{ij}$ the force between corresponding particles.

3.2.5 Periodic Boundary Conditions and Electrostatic Interactions

An important aspect in the simulation of a biomolecular system exposed to explicit solvent is the treatment of boundary conditions. Restricting the system to a finite space by hard boundaries leads to artifacts, as a large number of particles stick at the boundary surface [39]. Most MD simulations therefore include periodic boundary conditions. Thereby, the simulation system represents a unit cell in an infinite lattice of copies of the system. As shown in Fig 3.1, a particle that leaves the system towards one side enters it on the opposite side. Using periodic boundary conditions, particles also interact with particles of other periodic images [39].

In order to reduce computational costs, short-range interactions are usually truncated and shifted so that they vanish at a predefined cut-off distance ($\sim 9\mathring{A}$) [37, 39]. The long-range Coulomb interactions are treated with the Ewald summation method, which improves the scaling of the performance from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^{\frac{3}{2}})$ [40]. In this method, a Gaussian screening charge density is added to each point charge in order to generate short-range electrostatic potentials. Each screening charge density is neutralized by means of a compensating Gaussian charge density (Fig 3.2), whose long-range electrostatic potential is obtained by solving the Poisson equation in Fourier-space. By correcting for self-interactions, the overall electrostatic potential finally becomes [41]:

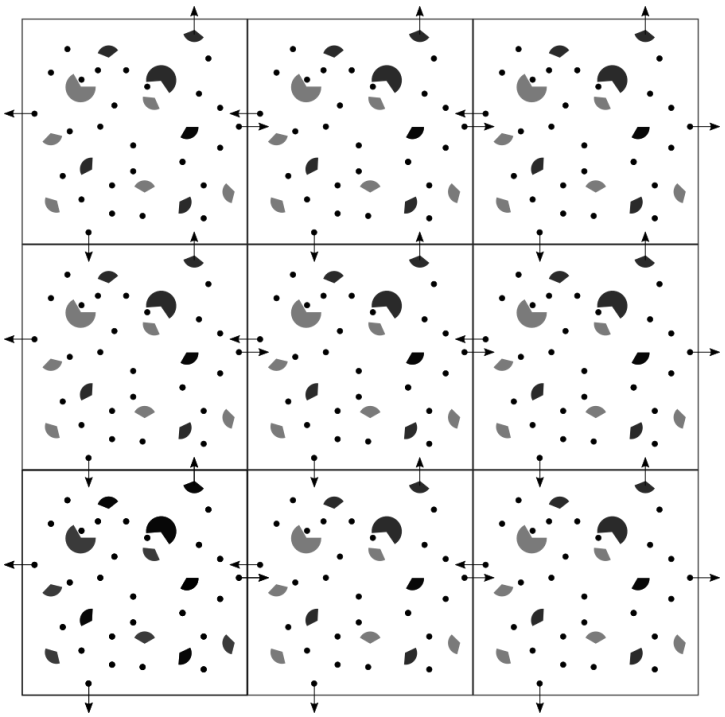


Figure 3.1: Schematic representation of periodic boundary conditions in two dimensions: The central unit cell is the simulation system, which is embedded in an infinite lattice of copies.

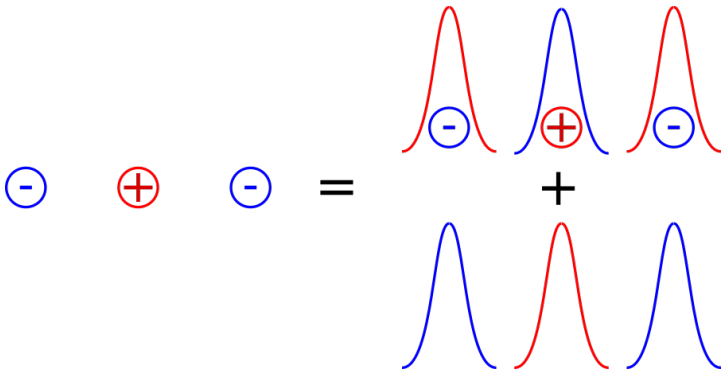


Figure 3.2: Principle of the Ewald summation. The long-range point-charge potentials are split into two terms: The point charges screened by Gaussian distributions of opposite sign plus compensating Gaussian distributions.

$$U_{elec} = \frac{1}{2} \cdot \sum_{i \neq j} \frac{q_i q_j \mathrm{erfc}(\sqrt{\alpha} r_{ij})}{r_{ij}} + \frac{1}{2V} \sum_{k \neq 0} \frac{4\pi}{k^2} |\rho(k)|^2 \exp\left[-\frac{k^2}{4\alpha}\right] - \left(\frac{\alpha}{\pi}\right)^{\frac{1}{2}} \cdot \sum_{i=1} q_i^2, \quad (3.13)$$

whereby α is chosen such, that we achieve best computational efficiency [41].

3.2.6 Umbrella Sampling and Hamiltonian Replica Exchange

Umbrella Sampling and Hamiltonian Replica Exchange Molecular Dynamics simulations are advanced sampling methods, which enable the calculation of Free Energy profiles along a reaction coordinate ξ [42, 43]. In Umbrella Sampling, a system is simulated in consecutive windows, whereby each window is characterized by a specific external potential. The external potential is usually of harmonic form employing for every window i the same spring constant but different equilibrium values ξ_i^{ref} [43, 44]:

$$U_i^{ext} = \frac{k}{2} \cdot (\xi - \xi_i^{ref})^2. \quad (3.14)$$

Note that the set of ξ_i^{ref} defines the reaction path. The recorded probability distribution of the reaction coordinate $P_i(\xi)$ depends on the spring constant k and ξ_i^{ref} [43]. Importantly, they should warrant the probability distributions of juxtaposed windows to overlap sufficiently, as this is essential for the calculation of the Free Energy profile (also denoted as Potential of Mean Force, PMF) with the Weighted Histogram Analysis Method (WHAM) [45].
However, in many cases, the Hamiltonian Replica Exchange Method offers better convergence prospects. Here, simulation of the windows is not performed consecutively but

in parallel, while every window may still have the same external potential as in the standard Umbrella Sampling [46–48]. After a certain number of steps, neighboring windows are allowed to exchange conformations based on the Metropolis criterion [49]:

$$P((X_i, Y_{i+1}) \rightarrow (Y_i, X_{i+1})) = \min\{1, \exp[\beta \cdot (U(X_i) - U(X_{i+1}) + U(Y_{i+1}) - U(Y_i))]\}. \tag{3.15}$$

X_i and Y_i denote individual conformations of the system exposed to window i. This method can be used in various contexts, e.g. by varying the temperature instead of the Hamiltonian.

3.2.7 Weighted Histogram Analysis Method

Using the aforementioned sampling methods, we obtain biased probability distributions $P_i^{bias}(\xi)$. The Weighted Histogram Analysis Method (WHAM) facilitates the determination of the entire unbiased probability distribution $P(\xi)$. Therefore, the latter is expressed as a linear combination of unbiased probability distributions from the Umbrella windows [44, 50]:

$$P(\xi) = \sum_{i=1} C_i(\xi) P_i(\xi). \tag{3.16}$$

The unbiased probability distributions are given by [50, 51]

$$P_i(\xi) = P_i^{bias}(\xi) \cdot \exp[\beta w_i(\xi)] \cdot \exp[-\beta F_i], \tag{3.17}$$

where $w_i(\xi)$ are the Umbrella potentials and F_i are free energy constants we need to solve for.

Minimizing the statistical error, $\frac{\partial \sigma(P(\xi))^2}{\partial C_i(\xi)}$, under the constraint $\sum_{i=1} C_i(\xi) = 1$ eventually yields [44, 45, 50, 51]

$$P(\xi) = \sum_{i=1} \frac{n_i \cdot P_i^{bias}(\xi)}{\sum_{j=1} n_j \exp[-\beta(w_j(\xi) - F_j)]}. \tag{3.18}$$

n_i is the number of conformations sampled in window i, and Boltzmann-factors of the free energy constants obey [44, 50, 51]

$$\exp[-\beta F_i] = \int \mathrm{d}\xi \exp[-\beta w_i(\xi)] \cdot P(\xi). \tag{3.19}$$

Equation 3.18 and 3.19 are solved self-consistently by starting with an initial set for the free energy constants, and these are plugged into equation 3.18. Thus, we obtain an estimate for the probability distribution $P(\xi)$, from which we calculate a new set of free energy constants F_i. This iterative procedure is carried out until convergence is achieved. The free energy profile can then simply be determined by Boltzmann inversion of $P(\xi)$ [44, 50, 51].

3.2.8 Free Energy Perturbation and Thermodynamic Integration

Suppose we want to calculate the free energy difference between two thermodynamic states A and B, which differ with respect to the potential energies U_A and U_B. The free energy difference is then given by:

$$\Delta F_{A,B} = F_B - F_A = -k_B T \cdot \ln\left(\frac{Z_B}{Z_A}\right). \tag{3.20}$$

Here, Z_A and Z_B denote configurational partition functions, as the integration over the momenta cancels out [44],

$$\begin{aligned} Z_A &= \int \mathrm{d}^N\mathbf{r}\, \exp[-\beta U_A(\mathbf{r}_1, ..., \mathbf{r}_N)] \\ Z_B &= \int \mathrm{d}^N\mathbf{r}\, \exp[-\beta U_B(\mathbf{r}_1, ..., \mathbf{r}_N)]. \end{aligned} \tag{3.21}$$

The ratio of the configurational partition functions in equation 3.20 can be reformulated as an average of ensemble A [44]:

$$\begin{aligned} \frac{Z_B}{Z_A} &= \frac{1}{Z_A}\int \mathrm{d}^N\mathbf{r}\, \exp[-\beta U_A(\mathbf{r}_1, ..., \mathbf{r}_N)] \cdot \exp[-\beta(U_B(\mathbf{r}_1, ..., \mathbf{r}_N) - U_A(\mathbf{r}_1, ..., \mathbf{r}_N))] = \\ &= \langle \exp[-\beta(U_B(\mathbf{r}_1, ..., \mathbf{r}_N) - U_A(\mathbf{r}_1, ..., \mathbf{r}_N))] \rangle_A . \end{aligned} \tag{3.22}$$

Plugging equation 3.22 into 3.20, yields the Zwanzig equation [44, 52]:

$$\Delta F_{A,B} = -k_B T \cdot \ln \langle \exp[-\beta(U_B - U_A)] \rangle_A . \tag{3.23}$$

Practically, however, a direct evaluation of equation 3.23 is often problematic. If the configurations we sample for ensemble A are unfavorable in ensemble B, then we underestimate the Boltzmann factor in the Zwanzig equation. A solution to this problem is

to include intermediate ensembles between A and B by introducing a 'control-variable' λ. Therefore, we define a new potential energy function [44]:

$$U(\mathbf{r}_1, .., \mathbf{r}_N, \lambda) = (1 - \lambda)U_A + \lambda U_B. \tag{3.24}$$

Thus, increasing λ from 0 to 1 allows a stepwise transition from A to B.

The free energy difference can be written as

$$\Delta F_{A,B} = \int_0^1 \frac{\partial F}{\partial \lambda} \mathrm{d}\lambda, \tag{3.25}$$

and with

$$\frac{\partial F}{\partial \lambda} = -\frac{k_B T}{Z}\frac{\partial Z}{\partial \lambda} = -k_B T \int \mathrm{d}^N \mathbf{r} \left(-\beta \frac{\partial U}{\partial \lambda}\right) \exp[-\beta U(\mathbf{r}_1, ..., \mathbf{r}_N, \lambda)] = \left\langle \frac{\partial U}{\partial \lambda} \right\rangle, \tag{3.26}$$

we arrive at the thermodynamic integration formula [44, 53]:

$$\Delta F_{A,B} = \int_0^1 \left\langle \frac{\partial U}{\partial \lambda} \right\rangle \mathrm{d}\lambda = \int_0^1 \langle U_B - U_A \rangle_\lambda \, \mathrm{d}\lambda. \tag{3.27}$$

In order to calculate now $\Delta F_{A,B}$, MD simulations or Monte Carlo samplings are carried out for different values of λ within [0,1] and equation 3.27 is integrated numerically.

3.3 Quantum Chemistry

3.3.1 Hartree-Fock Method

Separating electronic and nuclear motion by means of the Born-Oppenheimer approximation, we write the electronic Schrödinger equation as [54, 55]

$$\left[-\frac{1}{2} \cdot \sum_i \Delta_i - \sum_{A,i} \frac{Z_A}{r_{A,i}} + \sum_{i>j} \frac{1}{r_{ij}}\right] \Psi(\mathbf{r}, \mathbf{R}) = E_{el} \Psi(\mathbf{r}, \mathbf{R}), \tag{3.28}$$

with Z_A denoting the nuclear charge, $\mathbf{r}$ the coordinates of the electrons and $\mathbf{R}$ the fixed nuclear positions.

However, the electronic wave function does not only depend on spatial degrees of freedom but also on the spin coordinate. We therefore express the absolute electronic wave

function through spin orbitals $\phi_i(\mathbf{r}, \sigma) = \varphi_i(\mathbf{r}) \cdot \alpha_i(\sigma)$, which are a product of spatial orbitals φ_i and spin functions α_i [54, 56].
Due to the antisymmetric nature of fermions, we make the ansatz of a Slater determinant for the absolute electronic wave function [54, 56]:

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & .... & \phi_1(N) \\ \phi_2(1) & \phi_2(2) & .... & \phi_2(N) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \phi_N(1) & \phi_N(2) & .... & \phi_N(N) \end{vmatrix} \tag{3.29}$$

This satisfies both, the indistinguishability of electrons and the Pauli exclusion principle.

For the rest of this section, we introduce the one-electron operator [54, 56]:

$$\hat{h}(i) = -\frac{1}{2}\Delta_i - \sum_A \frac{Z_A}{r_{A,i}}. \tag{3.30}$$

Furthermore, we make use of the Dirac-notation for the one-electron integral [54, 56]

$$< i|\hat{h}|j >= \int \mathrm{d}\mathbf{x}_1 \phi_i^*(\mathbf{x}_1) h(\mathbf{r}_1) \phi_j(\mathbf{x}_1), \tag{3.31}$$

and the two-electron integral [54, 56] :

$$< ij|kl >= \int \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \phi_i^*(\mathbf{x}_1) \phi_j(\mathbf{x}_1) \frac{1}{r_{12}} \phi_k^*(\mathbf{x}_2) \phi_l(\mathbf{x}_2). \tag{3.32}$$

Note that the Hartree-Fock method is a variational approach, i.e. we seek to find the set of orbitals which minimizes E_{el}. The expression for the energy to be minimized can be rewritten as [54, 56]

$$E_{el} = \sum_i < i|\hat{h}|i > + \frac{1}{2} \sum_{ij} (< ii|jj > - < ij|ji >). \tag{3.33}$$

We carry out variation under the constraint of orthonormality of spinorbitals [54, 56]

$$\delta(E_{el}(\{\phi_i\}) - \sum_{i,j} L_{ij}(< i|j > -\delta_{ij}) = 0, \tag{3.34}$$

where L_{ij} denote undetermined Lagrange multipliers. From this, we can derive the Hartree-Fock equation [54, 56]

$$\hat{F}(\mathbf{x}_1)\phi_i(\mathbf{x}_i) = \epsilon_i\phi_i(\mathbf{x}_i). \tag{3.35}$$

The Fock operator is composed of the one-electron operator, Coulomb operator $\hat{J}(\mathbf{x}_1)$ and the exchange operator $\hat{K}(\mathbf{x}_1)$: $\hat{F}(\mathbf{x}_1) = \hat{h}(\mathbf{x}_1) + 2\hat{J}(\mathbf{x}_1) - \hat{K}(\mathbf{x}_1)$. The Coulomb and exchange operator are given by equation 3.36 and 3.37, respectively [54, 56].

$$\hat{J}(\mathbf{x}_1)\phi_i(\mathbf{x}_1) = \sum_j \int \mathrm{d}\mathbf{x}_2 \frac{1}{r_{12}} \phi_j^*(\mathbf{x}_2)\phi_j(\mathbf{x}_2)\phi_i(\mathbf{x}_1) \tag{3.36}$$

$$\hat{K}(\mathbf{x}_1)\phi_i(\mathbf{x}_1) = \sum_j \int \mathrm{d}\mathbf{x}_2 \frac{1}{r_{12}} \phi_j^*(\mathbf{x}_2)\phi_i(\mathbf{x}_2)\phi_j(\mathbf{x}_1) \tag{3.37}$$

Expanding the molecular orbitals as a linear combination of atomic orbitals, $\phi_i(\mathbf{x}_1) = \sum_s c_{is}\chi_s(\mathbf{x}_1)$, we finally obtain the Hartree-Fock-Roothaan equations:

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{3.38}$$

In the Hartree-Fock-Roothaan equation, $\mathbf{F}$ denotes the Fock-matrix, $\mathbf{S}$ the matrix of overlap integrals of atomic orbitals and ϵ the diagonal matrix of orbital energies. As a starting point in Hartree-Fock calculations, we choose an initial coefficient matrix $\mathbf{C}$. From the coefficients, we then calculate the Fock-matrix. With the estimate for $\mathbf{F}$, we solve the Hartree-Fock-Roothaan equation and hence obtain the molecular orbitals and a new coefficient matrix. This procedure is carried out iteratively until we obtain a self-consistent solution [54, 56].

3.3.2 Møller-Plesset Perturbation Theory

The Møller-Plesset perturbation theory captures electron correlation by introducing it as perturbation potential. Hereby, the Hartree-Fock function $\phi_0^{(0)}$ represents the zero-order approximation of the ground state wave function, and the unperturbed Hamiltonian is the sum of the Fock operators: $\hat{H}^{(0)} = \sum_i \hat{F}(\mathbf{x}_i)$.

For the ground state, we expand the wave function and energy in a power series [57, 58]

$$\phi_0 = \lim_{n\to\infty} \sum_{i=0}^{n} \lambda^i \phi_0^{(i)} \tag{3.39}$$

$$E_0 = \lim_{n\to\infty} \sum_{i=0}^{n} \lambda^i E_0^{(i)} \tag{3.40}$$

and the perturbation potential is $\hat{V} = \hat{H} - \hat{H}^{(0)}$.

Incorporating the expansion into the Schrödinger-equation yields [57, 58]:

$$(\hat{H}_0 + \lambda\hat{V})(|\phi_0^{(0)}> + \lambda|\phi_0^{(1)}> + \lambda^2|\phi_0^{(2)}> + ...) = \tag{3.41}$$
$$= (E_0^{(0)} + \lambda E_0^{(1)} + \lambda^2 E_0^{(2)} + ...)(|\phi_0^{(0)}> + \lambda|\phi_0^{(1)}> + \lambda^2|\phi_0^{(2)}> + ...).$$

We then evaluate equation 3.41 order by order in λ. Here, zeroth and first order correction together amount to the Hartree-Fock energy: $E_0^{(0)} + E_0^{(1)} = E_{HF}$. The correction to the Hartree-Fock energy comes with the second order, after which we usually truncate the expansion [57, 58]:

$$E_{MP2} = E_{HF} + \sum_{k\neq 0} \frac{|<\phi_k^{(0)}|\hat{V}|\phi_0^{(0)}>|^2}{E_0^{(0)} - E_k^{(0)}}. \tag{3.42}$$

The ϕ_k^0 are excited Slater determinants, and it can be shown that only double-excitations (i.e. we substitute occupied orbitals through virtuals : $a \to p, b \to q$) contribute. Eventually, we find for the energy corrected to second order [57, 58]

$$E_{MP2} = E_{HF} + \sum_{a<b}\sum_{p<q} \frac{|<ab|pq> - <ab|qp>|}{\epsilon_a + \epsilon_b - \epsilon_p - \epsilon_q}, \tag{3.43}$$

whereby we used the notation for the two-electron integral from equation 3.32.

3.3.3 Density Functional Theory

The electron density $\rho(\mathbf{r}) = \sum_i n_i|\phi_i(\mathbf{r})|^2$ is the central quantity in the Density Functional Theory (DFT) [59–61]. As shown by Kohn and Hohenberg, the ground state electron density $\rho_0(\mathbf{r})$ fixes an external potential V_{ext}, to which it is also subjected, in an unique way. Furthermore, the level of information stored in the ρ_0 and the ground state wavefunction ψ_0 is equivalent [59–61]. A formal expression for the ground state energy E_0 is then pursued based on the variational principle $E_0 = \min <\psi|\hat{H}|\psi>$ and the Levy-constraint search method [61–63]:

$$E_0 = \min_{\rho,\int\rho dV=N} \min_{\psi\to\rho} <\psi|\hat{T} + \hat{U} + V_{ext}|\psi> . \tag{3.44}$$

In this method, the ground state energy is found in two minimization steps: First, we minimize the integral starting from a set of trial wave functions ψ, which correspond to a certain electron density ρ. Thus, we obtain a set of electron densities. We then calculate

the ground state energy by minimizing the integral from the set of electron densities under the constraint of conserving the number of electrons N. Expressing the external potential through one-electron contribution $V_{ext} = \sum_{i=1}^{N} \upsilon(\mathbf{r}_i)$, we find $< \psi|V_{ext}|\psi >= \int \upsilon(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r}$. Thus, the first minimization step only affects the kinetic-energy operator $\hat{T}$ and the electron-electron interaction operator $\hat{U}$, which motivates the introduction of the auxiliary functional [61–63]:

$$F^{HK}[\rho] = \min_{\psi\to\rho} < \psi|\hat{T} + \hat{U}|\psi > . \tag{3.45}$$

Note that the constraint search method is fairly impractical. However, the above concepts provide the theoretical basis for modern DFT calculations, which rely on the Kohn-Sham method. In the Kohn-Sham method, we rewrite the auxiliary functional from equation 3.45 [60, 61]:

$$F^{HK}[\rho] = \hat{T}[\rho] + \frac{1}{2}\int \frac{\rho(\mathbf{r})\rho(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r'} + E^{XC}[\rho]. \tag{3.46}$$

The exchange correlation functional $E^{XC}[\rho]$ is not known yet. The famous Kohn-Sham equation is then derived by variation of the functional $E^{HK}[\rho] = \int \upsilon(\mathbf{r})\rho[\mathbf{r}]d\mathbf{r} + F^{HK}[\rho]$ with respect to ρ [60, 61]:

$$\left(-\frac{1}{2}\nabla^2 + \int \frac{\rho(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}\mathrm{d}\mathbf{r_2} + v^{XC}(\mathbf{r}_1) + \upsilon(\mathbf{r}_1)\right)\phi_i = \epsilon_i\phi_i. \tag{3.47}$$

Here, we have defined $v^{XC}(\mathbf{r}) = \frac{\delta E^{XC}[\rho]}{\delta\rho(\mathbf{r})}$. Different approaches to $E^{XC}[\rho]$ exist, within the projects of this thesis we used the B3LYP hybrid functionals. These are empirically derived functionals, which are composed by functionals stemming from various methods (Hartree-Fock exchange energy, local density approximation and generalized gradient approximation) [64–66]. Analogously to the Hartree-Fock method, the Kohn-Sham equation can be solved self-consistently now. From a starting set for the orbitals ϕ_i, we calculate the zeroth-iteration density $\rho(\mathbf{r})$ and consequently the operators in the Kohn-Sham equation. These in turn give us new orbitals, and we proceed iteratively until we match the convergence criteria.
The ground state energy in the Kohn-Sham approach then equals [60, 61]

$$E_0 = \sum_j \epsilon_j + E^{XC}[\rho(\mathbf{r})] - \int v^{XC}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} - \frac{1}{2}\int \frac{\rho(\mathbf{r})\rho(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r'}. \tag{3.48}$$

Unfortunately, the Kohn-Sham energy does not capture the long-range London dispersion interactions correctly. Modern DFT corrections are therefore commonly complemented by the Grimme correction. Two- and three-body energies are added to the

Kohn-Sham energy in this correction. Thereby, the two-body terms are dominant [67]:

$$E_2 = \sum_{AB} \sum_{n=6,8,10..} s_n \frac{C_n^{AB}}{r_{AB}^n} f_{d,n}(r_{AB}). \tag{3.49}$$

Summation is carried out over all atom-pairs (indicated by AB), s_n denote scaling factors, C_n^{AB} dispersion coefficients, r_{AB} the internuclear distance and $f_{d,n}(r_{AB})$ damping functions to avoid singularities for small r_{AB} [67]:

$$f_{d,n}(r_{AB}) = \frac{1}{1 + 6 \cdot (\frac{r_{AB}}{s_{r,n} R_0^{AB}})^{-\alpha_n}}. \tag{3.50}$$

This damping function is defined by new scaling parameters $s_{r,n}$ to the cut-off radii R_0^{AB} and steepness-parameters α_n [67].

3.3.4 Technical Concepts

Basis Sets

Remind that the molecular orbitals are a linear combination of atomic orbitals, about whose shape we have not made any statement yet. The most natural choice therefore are Slater-type-orbitals (STOs), as they resemble the eigenfunctions of the hydrogen atom [68]:

$$\chi^{STO} \sim r^{n-1} \exp[-\zeta r] Y_{lm}(\theta, \phi). \tag{3.51}$$

On the downside, the evaluation of integrals including STOs becomes computationally very expensive. Gaussian-type-orbitals (GTOs) benefit strongly from the product rule: The product of two GTOs can be represented by a linear combination of GTOs. Thus, modern QM calculations utilize GTOs [68]:

$$\chi^{GTO} \sim x^l y^m z^n \exp[-\alpha r^2]. \tag{3.52}$$

Moreover, it has also become standard to treat valence electrons with multiple basis functions. In this thesis, we usually have employed the Karlsruhe triple-valence basis set [69].

Resolution of Identity

The Resolution of Identity technique (RI) has matured to a widely used method in order to speed up QM calculations. Hereby, we introduce an auxiliary basis set $\eta_\mu(\mathbf{r})$, which

is defined as the product of two basis functions [70]:

$$\chi_i(\mathbf{r})\chi_j(\mathbf{r}) = \sum_{\mu} c_{ij}^{\mu}\eta_{\mu}(\mathbf{r}). \tag{3.53}$$

The coefficients c_{ij}^{μ} can be obtained by minimizing the RI-induced error in the two-electron integral (eq. 3.32). This yields $c_{ij}^{\mu} = \sum_{\nu} < ij|\nu > V_{\nu\mu}^{-1}$, with $V_{\nu\mu} = \int \frac{\eta_{\nu}(\mathbf{r})\eta_{\mu}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'$. Note that ν and μ indicate auxiliary basis functions.

In this way, the four-center Coulomb integrals (i.e. two-electron integral) reduce to the computationally cheaper calculations of three- and two-center integrals: [70]

$$< ij|kl > \approx \sum_{\mu,\nu} < ij|\mu > V_{\mu\nu}^{-1} < \nu|kl > . \tag{3.54}$$

3.4 Bibliography

[1] Franz Schwabl. Grundlagen. In *Statistische Mechanik*, chapter 1, pages 1–24. Springer Verlag, 3rd edition, 2006.

[2] Mark E. Tuckerman. Theoretical foundations of classical statistical mechanics. In *Statistical Mechanics: Theory and Molecular Simulation*, chapter 2, pages 53–73. Oxford University Press, 2010.

[3] Mark E. Tuckerman. The microcanonical ensemble and introduction to molecular dynamics. In *Statistical Mechanics: Theory and Molecular Simulation*, chapter 3, pages 74–132. Oxford University Press, 2010.

[4] Franz Schwabl. Gleichgewichtsensemble. In Franz Schwabl, editor, *Statistische Mechanik*, chapter 2, pages 25–74. Springer Verlag, 3rd edition, 2006.

[5] Mark E. Tuckerman. The canonical ensemble. In *Statistical Mechanics: Theory and Molecular Simulation*, chapter 4, pages 133–213. Oxford University Press, 2010.

[6] R. K. P. Zia, Edward F. Redish, and Susan R. McKay. Making sense of the Legendre transform. *American Journal of Physics*, 77(7):614–622, 2009.

[7] Mark E. Tuckerman. The isobaric ensembles. In *Statistical Mechanics: Theory and Molecular Simulation*, chapter 5, pages 214–260. Oxford University Press, 2010.

[8] Wolfgang Nolting. Klassische Statistische Physik. In *Grundkurs Theoretische Physik 6*, chapter 1, pages 3–98. Springer, 6th edition, 2007.

[9] Adam Hospital, Josep Ramon Goñi, Modesto Orozco, and Josep L Gelpí. Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry : AABC*, 8:37–47, 2015.

[10] Scott A. Hollingsworth and Ron O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143, 2018.

[11] Stewart A Adcock and J Andrew McCammon. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.*, 106(5):1589–1615, 2006.

[12] J. Meller. Molecular Dynamics. *eLS*, 2001.

[13] Leonid Pereyaslavets, Igor Kurnikov, Ganesh Kamath, Oleg Butin, Alexey Illarionov, Igor Leontyev, Michael Olevanov, Michael Levitt, Roger D. Kornberg, and Boris Fain. On the importance of accounting for nuclear quantum effects in ab initio calibrated force fields in biological simulations. *Proc. Natl. Acad. Sci. USA*, 115(36):8878–8882, 2018.

[14] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How Fast-Folding Proteins Fold. *Science*, 334(6055):517–520, 2011.

[15] Korbinian Liebl, Tomáš Dršata, Filip Lankaš, Jan Lipfert, and Martin Zacharias. Explaining the striking difference in twist-stretch coupling between DNA and RNA: A comparative molecular dynamics analysis. *Nucleic Acids Res.*, 43(21):10143–10156, 2015.

[16] Martin Zacharias. Atomic Resolution Insight into Sac7d Protein Binding to DNA and Associated Global Changes by Molecular Dynamics Simulations. *Angew. Chem. Int. Ed.*, 58(18):5967–5972, 2019.

[17] Thomas E. Markland and Michele Ceriotti. Nuclear quantum effects enter the mainstream. *Nature Reviews Chemistry*, 2(3):1–14, 2018.

[18] Alejandro Pérez, Mark E. Tuckerman, Harold P. Hjalmarson, and O. Anatole von Lilienfeld. Enol Tautomers of Watson–Crick Base Pair Models Are Metastable Because of Nuclear Quantum Effects. *J. Am. Chem. Soc.*, 132(33):11510–11515, 2010.

[19] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[20] Jiří Šponer, Giovanni Bussi, Miroslav Krepl, Pavel Banáš, Sandro Bottaro, Richard A. Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simn Poblete, Petr Jurečka, Nils G. Walter, and Michal Otyepka. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chem. Rev.*, 118(8):4177–4338, 2018.

[21] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiří Šponer, 3rd Cheatham, Thomas E, Charles A Laughton, and Modesto Orozco. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ conformers. *Biophysical Journal*, 92(11):3817–3829, 2007.

[22] Marie Zgarbová, Michal Otyepka, Jiří Šponer, Arnošt Mládek, Pavel Banáš, Thomas E. Cheatham, and Petr Jurečka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*, 7(9):2886–2902, 2011.

[23] Miroslav Krepl, Marie Zgarbová, Petr Stadlbauer, Michal Otyepka, Pavel Banáš, Jaroslav Koča, Thomas E. Cheatham, Petr Jurečka, and Jiří Šponer. Reference Simulations of Noncanonical Nucleic Acids with Different χ Variants of the AMBER Force Field: Quadruplex DNA, Quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.*, 8(7):2506–2520, 2012.

[24] Marie Zgarbová, F. Javier Luque, Jiří Šponer, Thomas E. Cheatham, Michal Otyepka, and Petr Jurečka. Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.*, 9(5):2339–2354, 2013.

[25] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham, Rodrigo Galindo-Murillo, and Petr Jurečka. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.*, 11(12):5723–5736, 2015.

[26] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jrgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep Lluis Gelp, Carlos Gonzlez, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55–58, 2016.

[27] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W. Ponder, and Pengyu Ren. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.*, 14(4):2084–2108, 2018.

[28] Zhifeng Jing, Chengwen Liu, Sara Y. Cheng, Rui Qi, Brandon D. Walker, Jean-Philip Piquemal, and Pengyu Ren. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.*, 48(1):371–394, 2019.

[29] Giuseppe Cassone, Holger Kruse, and Jiří Šponer. Interactions between cyclic nucleotides and common cations: an ab initio molecular dynamics study. *Phys. Chem. Chem. Phys.*, 21(15):8121–8132, 2019.

[30] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, 1982.

[31] Daan Frenkel and Berend Smit. Molecular Dynamics Simulations. In *Understanding Molecular Simulation*, chapter 4, pages 63–107. Academic Press, 2nd edition, 2002.

[32] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[33] Alberto Pérez, F. Javier Luque, and Modesto Orozco. Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.*, 45(2):196–205, 2012.

[34] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[35] Tetsuya Morishita. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *J. Chem. Phys.*, 113(8):2976–2982, 2000.

[36] Stephen C. Harvey, Robert K.-Z. Tan, and Thomas E. Cheatham III. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.*, 19(7):726–740, 1998.

[37] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[38] Enrique de Miguel and George Jackson. The nature of the calculation of the pressure in molecular simulations of continuous models from volume perturbations. *The Journal of Chemical Physics*, 125(16):1–11, 2006.

[39] Daan Frenkel and Berend Smit. Monte Carlo Simulations. In *Understanding Molecular Simulation*, chapter 3, pages 23–62. Academic Press, 2nd edition, 2002.

[40] Mark E. Tuckerman. Evaluation of energies and forces. In *Statistical Mechanics: Theory and Molecular Simulation*, chapter Appendix B, pages 652–661. Oxford University Press, 2010.

[41] Daan Frenkel and Berend Smit. Long-Range Interactions. In *Understanding Molecular Simulation*, chapter 12, pages 291–320. Academic Press, 2nd edition, 2002.

[42] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[43] Johannes Kästner. Umbrella sampling. *WIREs Comput Mol Sci*, 1(6):932–942, 2011.

[44] Mark E. Tuckerman. Free energy calculations. In *Statistical Mechanics: Theory and Molecular Simulation*, chapter 8, pages 312–361. Oxford University Press, 2010.

[45] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

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

[47] Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional replica-exchange method for free-energy calculations. *The Journal of Chemical Physics*, 113(15):6042–6051, 2000.

[48] Jeremy Curuksu and Martin Zacharias. Enhanced conformational sampling of nucleic acids by a new Hamiltonian replica exchange molecular dynamics approach. *The Journal of Chemical Physics*, 130(10):104110, 2009.

[49] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[50] Marc Souaille and Benoit Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40–57, 2001.

[51] Benoit Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282, 1994.

[52] Robert W. Zwanzig. HighTemperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

[53] John G. Kirkwood. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.*, 3(5):300–313, 1935.

[54] C. David Sherrill. An Introduction to Hartree-Fock Molecular Orbital Theory. 2009.

[55] Lucjan Piela. Separation of Electronic and Nuclear Motions. In *Ideas of Quantum Chemistry (Second Edition)*, chapter 6, pages 257–335. Elsevier, Oxford, 2014.

[56] Lucjan Piela. Orbital Model of Electronic Motion in Atoms and Molecules. In *Ideas of Quantum Chemistry (Second Edition)*, chapter 8, pages 389–504. Elsevier, Oxford, 2014.

[57] Chr. Møller and M. S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, 46(7):618–622, 1934.

[58] Lucjan Piela. Correlation of the Electronic Motions. In *Ideas of Quantum Chemistry (Second Edition)*, chapter 10, pages 577–662. Elsevier, Oxford, 2014.

[59] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, 1964.

[60] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965.

[61] Lucjan Piela. Chasing Correlation Dragon: Density Functional Theory (DFT). In *Ideas of Quantum Chemistry (Second Edition)*, chapter 11, pages 663–717. Elsevier, Oxford, 2014.

[62] Mel Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proc. Natl. Acad. Sci. USA*, 76(12):6062, 1979.

[63] Mel Levy. Electron densities in search of Hamiltonians. *Phys. Rev. A*, 26(3):1200–1208, 1982.

[64] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

[65] Axel D. Becke. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J. Chem. Phys.*, 96(3):2155–2160, 1992.

[66] K. Kim and K. D. Jordan. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.*, 98(40):10089–10094, 1994.

[67] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010.

[68] Wolfram Koch and Max C. Holthausen. The Basic Machinery of Density Functional Programs. In *A Chemist's Guide to Density Functional Theory*, chapter 7, pages 93–116. John Wiley & Sons, Ltd, 2001.

[69] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

[70] Xinguo Ren, Patrick Rinke, Volker Blum, Jrgen Wieferink, Alexandre Tkatchenko, Andrea Sanfilippo, Karsten Reuter, and Matthias Scheffler. Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New Journal of Physics*, 14(5):053020, 2012.

4 Unwinding-Induced Melting of Double-Stranded DNA [1]

4.1 Introduction

The elasticity of DNA is central to many biological functions. In particular, DNA unwinding is implicated in processes such as replication, transcription and repair of damaged DNA [1–6]. The local melting of DNA induced by unwinding is thereby of special interest, as it provides direct access to the hereditary information. Experimentally, DNA unwinding has been studied on the kilo-base-pair length scale, leading to a comprehensive description of DNA's global relaxation to torsional stress [4, 7–18]. More recently, Molecular Dynamics simulations have provided valuable insight into DNA's local structure and flexibility [19–26]. Thereby, harmonic stiffness models have been employed to characterize DNA's global and local flexibility at equilibrium [27–32]. Nevertheless, on the base-pair level, many questions have remained open: How does DNA locally absorb torsional stress, and in how far are the structural changes dependent on the sequence? What amount of energy is required to induce separation of the DNA strands and to what level can this be modulated by global restraints? Are such phase transitions fully reversible?

In this study, we have performed Molecular Dynamics (MD) simulations on two 15 base-pair long dsDNA molecules exposed to explicit solvent. We include a torsional reaction coordinate acting on the termini of the DNA, hence enabling us to control unwinding of the central segment and calculate changes in the free energy. From these Hamiltonian Replica Exchange Umbrella Sampling (HREUS) simulations we find a harmonic response of the DNA molecule to torsional stress close to equilibrium. Here, the free energy curve can be reproduced from a harmonic model based on a stiffness description inferred from covariance matrices. In this way, we deduce that coupling to bending degrees plays an important role in absorbing local unwinding stress. The harmonic response remains valid up to a phase transition, which is characterized by denaturation. In case of the AT-rich sequence, this leads to local strand dissociation, whereas backbone deformations and stacking perturbations are predominantly found for the pure GC-sequence. Achieving this phase transition through unwinding requires a free energy of about ~ 1.0 (AT) to 1.2 kcal/mol (GC) per base-pair step. We then have performed another series

[1] This chapter has been previously published in similar form in J. Phys. Chem. B, 121(49), 11019-11030, 2017.

of free energy simulations, during which we restrain the bending flexibility of the DNA molecules. In line with our harmonic model based conception of relaxation mechanisms through coupled bending, we show that enhancing bending suppression leads to earlier onset of local melting. The DNA molecules show only little further resistance to unwinding in the melted phase, and the free energy profiles indicate full reversibility of the unwinding-process. We validate this reversible character through unrestrained MD simulations starting from denatured structures, where we obtain relaxation to intact B-DNA on the nanosecond timescale.

4.2 Materials and Methods

4.2.1 DNA Structures and Equilibration

All MD simulations have been performed on two B-DNA structures. Both duplexes are 15 bp long and either contain a central segment composed of A:T base-pairs d(5'-CGCGCATATACGCGC) or of G:C base-pairs d(5'-CGCGCGCGCGCGCGC), termed centAT and centGC, respectively. The simulations have been carried out with the Amber14 Molecular Dynamics Package [33]. The simulation systems were neutralized by potassium ions [34], and were solvated in explicit solvent (TIP3P water model) with a rectangular box [35]. The minimum distance between DNA and box boundary was set to $10\mathring{A}$. The bsc0 force field was used to describe DNA's interactions [36] and all simulations were performed with the pmemd.cuda module. Prior to simulation, the DNA structures were aligned along the z-axis of the box and the systems were energy minimized in 5000 steps. Afterwards, the systems were heated up to 300 K in three stages including positional restraints on all heavy atoms of the DNA. Each stage entailed a temperature increase of 100 K and was simulated in the NVT ensemble for 100 ps. The positional restraints were subsequently reduced from $25\,\mathrm{kcal/(mol\cdot\mathring{A}^2)}$ to $0.5\,\mathrm{kcal/(mol\cdot\mathring{A}^2)}$ in five consecutive stages simulated in the NPT ensemble (T=300 K, p=1bar). The equilibration was finalized by a 2 ns NPT simulation including cylindrical restraints (discussed in the next paragraph) only on terminal base-pairs.

4.2.2 Cylindrical Restraints

In order to allow translational and rotational motion with respect to the helical z-axis, we applied cylindrical restraints. Thereby, only the distance to the z-axis was restrained, hence also resulting in restraining bending degrees of freedom. Three different scenarios were investigated: In a first case (unrestrained bending: UB), the cylindrical restraints were applied on the two terminal base-pairs of one end. Thus, the DNA retains full bending mobility. For the second setup, weak bending restraints (WBR), cylindrical restraints were activated on the two terminal base-pairs of both ends, which also limits DNA's bending mobility. Finally, restraints were applied on the four terminal base-pairs

of both ends, termed strong bending restraints (SBR). This freezes DNA's bending motion almost completely, whereas other degrees of freedom remain again unaffected. In all cases, the force constant was set to $0.1\,\mathrm{kcal}/(\mathrm{mol}\cdot\mathrm{\AA}^2)$.

4.2.3 HREUS Simulations

Controlled unwinding of the DNA molecules was induced using the Umbrella Sampling (US) method. Thereby, a torque was applied on the C1' atoms of the 4th base-pair and its symmetric counterpart, the 12th base-pair. A quadratic potential was assigned on the torsional (dihedral) reaction coordinate:

$$V = k \cdot (\xi - \xi_i^{ref})^2. \tag{4.1}$$

The force constant k was set to $150\,\frac{\mathrm{kcal}}{\mathrm{molrad}^2} \cong 0.0457\,\frac{\mathrm{kcal}}{\mathrm{moldeg}^2}$. ξ denotes the reaction coordinate as depicted in Fig 4.1 . Unwinding is induced by changes in the reference value ξ_i^{ref}, which was decreased from 70° to −35° in 5° intervals during the US simulations, resulting in 22 US windows. Sampling along the reaction coordinate was improved by allowing exchanges of the conformations between adjacent US windows (Hamiltonian Replica Exchange Umbrella Sampling: HREUS). Exchanges between neighboring replicas were attempted every 500 steps, and overall at least 20000 exchanges were attempted for every replica. Using a time step of 2 fs, this results in a simulation time of at least 20 ns for every replica. The simulation time was extended up to 120 ns for regions, where convergence issues were noted. The trajectories were analyzed with VMD and CURVES+ [37, 38]. Free energy profiles were calculated with WHAM and corresponding error bars based on thermodynamic integration and block averaging [39, 40].

We followed the methodology of Strahs and Schlick to evaluate bending (curvature), which is anisotropic and can be decomposed into directional terms [41]:

$$curvature = \sqrt{\theta_T^2 + \theta_R^2}, \tag{4.2}$$

with

$$\theta_T = \sum_{j=4}^{11} \tau_j \cdot \cos\left(\sum_{i=4}^{j} t_i\right) + \rho_j \cdot \sin\left(\sum_{i=4}^{j} t_i\right), \tag{4.3}$$

and

$$\theta_R = \sum_{j=4}^{11} -\tau_j \cdot \sin\left(\sum_{i=4}^{j} t_i\right) + \rho_j \cdot \cos\left(\sum_{i=4}^{j} t_i\right). \tag{4.4}$$

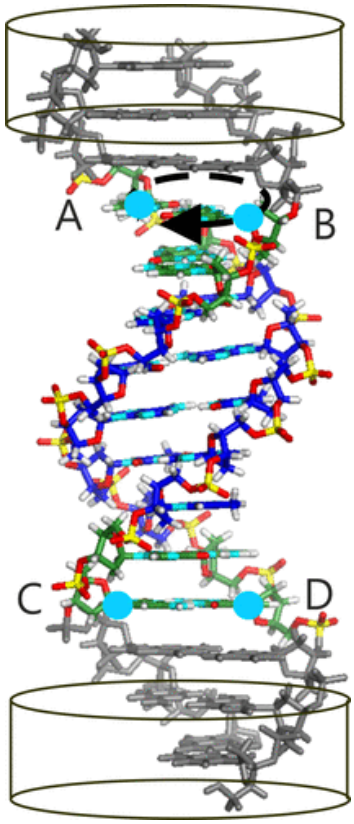


Figure 4.1: Umbrella Sampling setup. The DNA is shown as stick model, atom-color coded center indicates absence of restraints for all setups. In case of weak or strong bending suppression (WBR/SBR), cylindrical restraints with respect to the z-axis have been applied on terminal base-pairs. The torque to induce unwinding of the central segment acts on the C1' atoms of the 4th and 12th base-pair (shown as blue dots, A-D).

The variables τ_j and ρ_j denote tilt and roll of step j and t_i the twist of step i. Directional bending is indicated by global roll θ_R and global tilt θ_T: $\theta_R < 0$ means bending toward the minor, $\theta_R > 0$ toward the major groove. θ_T quantifies bending toward the backbone. For denatured structures, this methodology is error-prone due to difficulties in parameterizing the local parameters roll/tilt/twist. Therefore, we also employed the total bend angle computed with CURVES+ as an alternative. It assigns a helical axis for the DNA molecule so that total bending is mainly determined by the relative orientation of the DNA's termini.

4.3 Results and Discussion

4.3.1 Free Energy Changes related to DNA Unwinding and Comparison to Harmonic Models

We have performed free energy simulations of two 15 bp long DNA sequences, one with a central AT- (centAT) and one with a central GC-sequence (centGC). By applying a torque to the terminal parts of the DNA (4th and 12th bp), we have been able to stepwise unwind the central eight base-pair steps. Changes in the free energy along the torsional reaction coordinate ξ have been calculated with the WHAM-procedure. Notably, unwinding of DNA affects its twisting, stretching and bending behavior. However, in vivo these modes can also be restricted, e.g. changes in bending are suppressed through packing or binding by proteins. Therefore, we have investigated the impact of bending-restriction on DNA's response to unwinding by covering three scenarios of bending suppression: In a first scenario, we left bending motions unrestrained (termed UB). We have then performed simulations on a setup which includes weak bending suppression (termed WBR) and on one with strong bending suppression (termed SBR) by inclusion of cylindrical restraints on the termini. The calculated free energy profiles of the three scenarios for both sequences are shown in Fig 4.2. Close to equilibrium, the free energy changes quadratically with respect to unwinding with an in general softer response for the centAT sequence. Moreover, we observe that an increased bending suppression results in a stiffer response of the DNA to unwinding, as unwinding then implies a higher cost in the free energy. Consistently, the suppression of bending also leads to an earlier onset of local denaturation. In this section, however, we focus on the harmonic regime (denoted as regime I) in which the DNA retains its regular structure. Here, we draw comparison between the calculated free energy profiles and a harmonic approximation for DNA's deformability. In the harmonic approximation, DNA's free energy is described by

$$F(\Delta w) = \frac{1}{2}\Delta w^T \cdot K \cdot \Delta w, \tag{4.5}$$

where Δw denotes deviations of the internal coordinates from their equilibrium [28, 30,

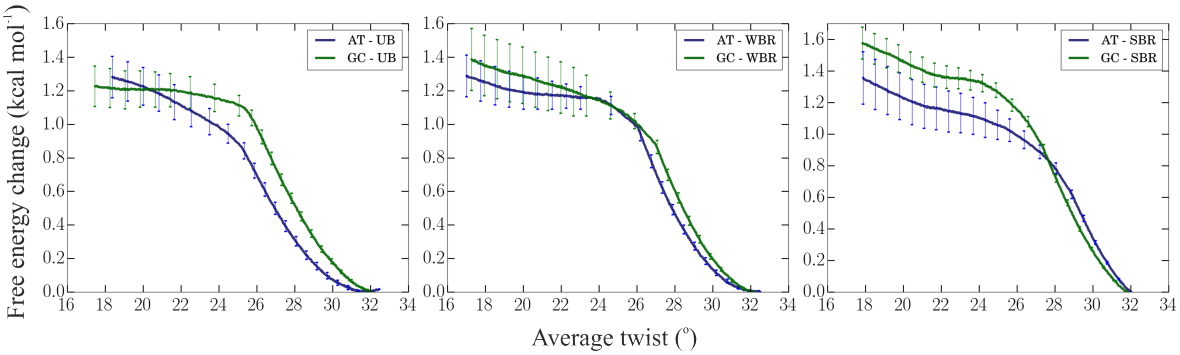


Figure 4.2: Changes in the free energy upon DNA unwinding for different levels of bending suppression.

32]. K represents the stiffness matrix, which is obtained from inversion of the covariance matrix C:

$$K = k_B T C^{-1}. \tag{4.6}$$

Note that we have calculated equilibrium parameters as well as the stiffness matrix from unrestrained 500 ns long MD simulations. Thus, the motivation for comparing free energy calculations to the harmonic model is twofold: First, how well do free energy profiles agree with the deformability description extracted from equilibrium MD simulations, and second, which modes of the DNA absorb the unwinding stress. To this purpose, we have investigated various applications of the harmonic model. In the simplest approach, we have chosen only the twist variable (between 4th and 12th bp) as the internal coordinate. The stiffness matrix then reduces to the inverse of the twist-variance. The second set of internal coordinates consists of the twist, stretch, global roll and global tilt variable, whereby the latter two describe orthogonal bending motions, hence K becomes a 4x4 matrix. Furthermore, we have also applied the harmonic model directly to the torsional reaction coordinated ξ to assess the accuracy and range of validity of the harmonic approximation.

The free energy estimates obtained from the harmonic model can be directly compared to the PMFs of the HREUS simulations. Close to equilibrium, relative free energy curves show harmonic behavior, matching the harmonic estimates from the inverse variance of the reaction coordinate ξ (compare black and pink dotted curves in Fig 4.3). Thus, we argue that our harmonic models allow for quantitative conclusions on DNA's elasticity. Application of the simplest harmonic model with the inverse of the twist variance as the only force constant already agrees well with the HREUS simulations in the harmonic regime (compare red and black dotted curve in Fig 4.3). Unwinding stress in the harmonic regime is hence absorbed by twisting motions of the base-pairs. Note, however, that this model indirectly also accounts for coupling to orthogonal degrees of freedom such as bending. It therefore relates to an effective persistence length of $\sim$ 76 and $\sim$ 83 nm for centAT and centGC, which is in good agreement with experiments [14].

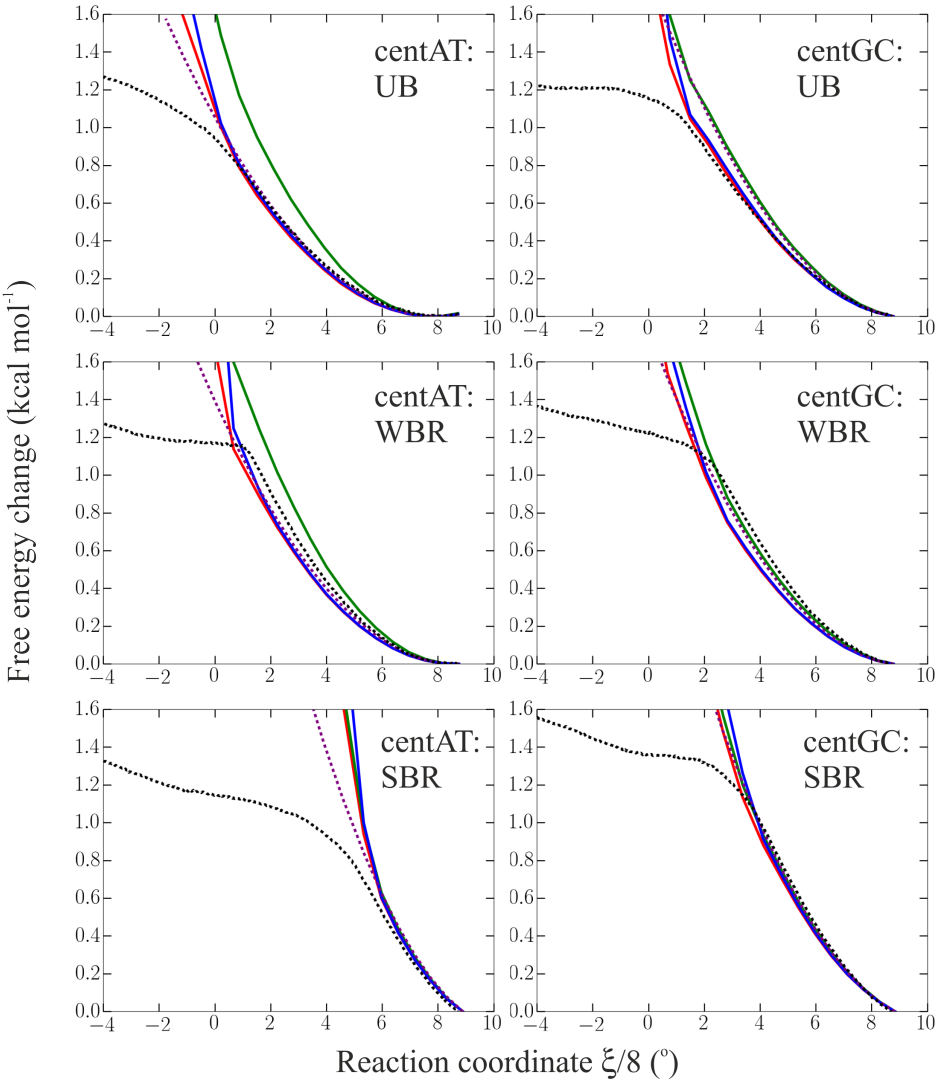


Figure 4.3: Comparison of the PMF (per base-pair step) along the reaction coordinate ξ (black dotted curves) obtained from the HREUS simulations with different harmonic models. The sequences and setups are given in the inset of each panel. The dotted pink curve is computed from the stiffness with respect to ξ from equilibrium simulations (i.e. absence of an external torque) of the respective restraining system (UB, WBR, SBR). Red curves show free energy predictions based only on the twist stiffness. In order to translate this to the reaction coordinate, the twist has been calculated as average in each HREUS window, hence obtaining free energy curves based on the quadratic model. A similar procedure was carried out to obtain the blue and green curves, but employing different modeling of the stiffness. For the blue curve, the 4x4 stiffness matrix includes twist, stretch, global roll and global tilt entries. For the green curve, only the diagonal twist stiffness of this matrix was taken into account.

The second harmonic model, the 4x4 stiffness matrix, also results in close agreement with the HREUS-PMF calculations (blue curve in Fig 4.3). This model includes twist, bending and stretching deformabilities as diagonal elements in K. The non-diagonals represent coupling terms. If we instead use just the diagonal twist stiffness element of this 4x4 matrix, we overestimate free energy changes upon unwinding (see corresponding green curves in Fig 4.3). From this, we see that twist-bend coupling plays an important role in relaxing unwinding stress. Moreover, this effect is apparently stronger for the centAT sequence, indicating that the softer response for AT-rich sequences to unwinding largely results from their more pronounced twist-bend coupling. Furthermore, we can translate the diagonal twist-stiffness element to an intrinsic twist persistence length of $\sim 102 - 120$ nm. These estimates are in close agreement with previous simulation studies [32] and experiments [14]. In the case of strong bending suppression (SBR-simulations), free energies obtained from the full 4x4 matrix and from just the diagonal twist element are indistinguishable (compare blue and green curves in Fig 4.3). This stresses the consistency of our harmonic model considerations, as in the SBR simulations the DNA can only respond through twisting motions. We therefore conclude that deformation energy caused by torsional stress is nearly fully absorbed in untwisting and coupled local bending of the DNA. Besides, we also highlight the validity range of the harmonic model: During equilibration MD simulations we sample fluctuations in mean twist around $3 - 3.5°$ (in the UB case). However, the harmonic model agrees with HREUS-PMFs in deviations up to $\sim 7°$. Eventually, we find that the breakdown of the harmonic response depends on the level of bending suppression. Analogously to the UB-case, we also observe a harmonic response for the WBR and SBR simulations, but it is limited to a smaller deformation range with a stiffer response. Note that the discussed implications of bending restriction may not only apply for straight, but also for packed or circularly closed DNA and are hence of high biological relevance [42].

4.3.2 Free Energy Change upon DNA Unwinding beyond the Harmonic Regime

As shown in Fig 4.2 and 4.3, the free energy required to overcome the harmonic regime amounts ~ 1.0 kcal/mol per bp-step for the centAT sequence, and is slightly more expensive for the centGC sequence (by ~ 0.2 kcal/mol per bp-step). Overcoming the harmonic regime implies a phase transition, resulting in local melting of the DNA. In the UB case, this phase transition occurs at an average twist of $\sim 25 - 26°$. Free energy costs for further unwinding are then significantly reduced. In the WBR-HREUS simulations, the harmonic regime ends at a higher average twist of $\sim 27°$. The relation between bending suppression and earlier phase transition holds also true for the SBR case, where the harmonic regime ends at $\sim 28°$. Note that the free energies required to induce phase transition is smaller for the centAT sequence in all setups, sequence-dependent structural changes will be discussed in the next paragraph. Intriguingly, the PMFs do not show any local minima in the melted regime. Although one can assume that this might be due to insufficient sampling in this regime, we emphasize that this indicates

	AT			GC		
regime	slope UB	slope WBR	slope SBR	slope UB	slope WBR	slope SBR
I	0.886	0.920	0.998	0.861	0.905	1.01
II	1.55	2.86	1.33	1.93	2.04	1.37
III	1.27	1.21	1.03	1.24	1.07	1.01

Table 4.1: Slopes for the average twist of the central 8 base-pair steps vs. torsion reaction coordinate ξ.

a fully reversible character of unwinding. The relaxation of underwound and disrupted structures to regular B-DNA is investigated in paragraph 4.3.5.

4.3.3 Torque Induced Helical Conformations

As shown in Fig 4.4, application of a torque results in systematic untwisting of the DNA. In the harmonic regime, marked as area I in Fig 4.4, we find a linear relationship between reaction coordinate and average twist. If the torsional stress causes only untwisting, a slope of 1 would be expected for the linear relationship. However, the fact that external stress is partially absorbed by bending explains the lower correlation of $\sim 85\%$ in the UB case (Table 4.1).

Furthermore, simulations under conditions of bending suppression show higher correlations: We obtain a correlation of $\sim 92\%$ for the WBR and almost exactly 100% for the SBR case (Table 4.1). The relationship between torque and average twist is therefore fully in line with our conclusion on the elasticity of DNA (see paragraph 4.3.1) and emphasizes the role of twist-bend coupling. For the further discussion of structural changes during phase transition, we limit on the total bend angle and relevant helical variables, as most of the other parameters do not allow conclusions due to very large fluctuations in the melted phase. The total bend angle continuously increases upon unwinding in regime I, but begins to drop upon transition to regime II (see Fig 4.5 and Fig 4.6). Moreover, we see that bending is mostly composed of strong local kinking in this phase because of structural failure. The loss in twist-bend coupling must be compensated for by increased untwisting. Thus, we obtain torque-twist slopes clearly exceeding 100% in regime II. This correlation balances again in regime III with slopes closer to 1 (Fig 4.4), although irregular bending is still observed. Structural failure occurs with the beginning of the transition phase (regime II) and its character is highly sequence-dependent (Fig 4.5, 4.7): For the centAT sequence, we observe local strand dissociations, various flipping and unstacking events. In the centGC case, however, we have mainly sampled stacking perturbations and left-handed segments. Apart from transient flipping events, the centGC sequence largely conserves base-pairing. Note that this different behavior is revealed by significant changes in the opening angle for the centAT sequence (Fig 4.6). Upon unwinding, helical rise decreases in regime I due to twist-stretch coupling [26, 43], hence shifting the DNA's topology closer to A-form. In regime II, however, this

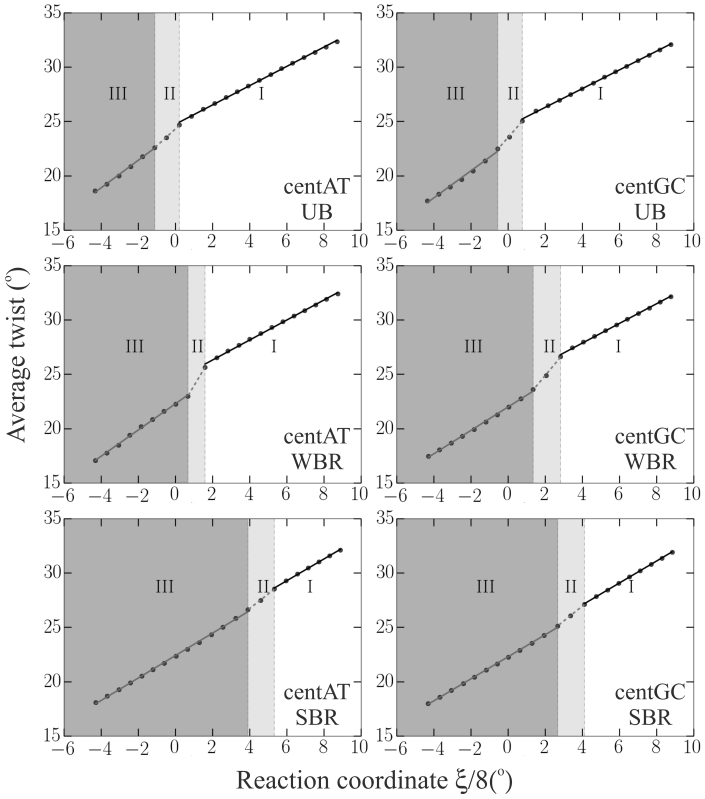


Figure 4.4: Relation between reaction coordinate $\xi/8$ and average twist over the central 8 bp steps. Only data of the last 20 ns of each replica was used, ξ was measured every ps, and snapshots were taken every 10 ps. For each replica, an average was calculated for ξ and the average twist over the central 8 bp steps. Average twist values were determined with CURVES+. Only the 4th, 8th and 12th bp were taken into account. The twist values computed for the steps $4 \mapsto 8$ and $8 \mapsto 12$ were added and divided by 8. This procedure turned out to perform robustly in regime II and III, where strong structural deformations prevent reliable parameterization of single base-pair step parameters. Linear interpolation was performed for each regime.

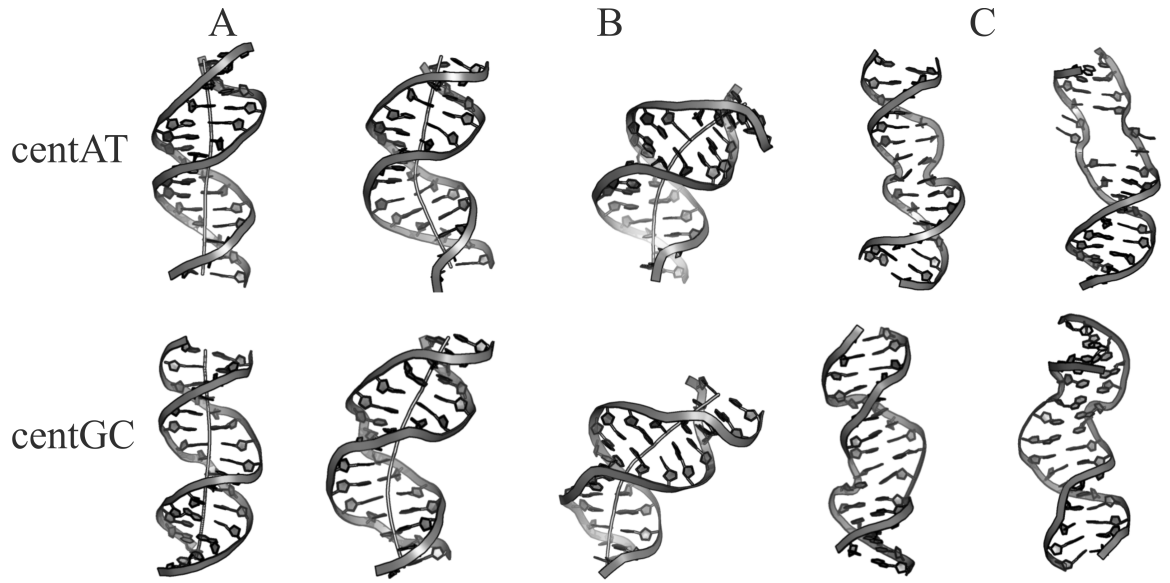


Figure 4.5: Snapshots of representative conformations. (A) Snapshots taken from equilibrium simulations. Helical axis was computed with CURVES+. (B) Snapshots taken from regime I prior to melting (C) Snapshots taken from regime II/III, exhibiting less bending but structural distortions in the central segment.

is reversed by rapid helical extension. For the centAT sequence, the helical extension relaxes back to B-form values, whereas the centGC sequence shows an excessively large extension due to the ladder-like stacking irregularities. In order to better distinguish regime II from III, we have looked at the probability distributions of the mean twist in different regimes (Fig 4.8). While regime I is characterized approximately by single Gaussian distributions, regime II shows bimodal behavior. This reflects a greater variety of accessible states in regime II, resulting from partially melted structures mixed with regular conformations. In regime III, the distribution becomes narrower and unimodal again. Thus, we argue that there is a small window for the external stress to generate a high structural polymorphism.

4.3.4 Changes in the Backbone Structure

In the previous paragraph, we have discussed that unwinding in regime I transforms DNA's shape closer to A-form like structures. This is also reflected in transitions of the backbone structure and sugar puckering (Fig 4.9): Unwinding significantly reduces the B-DNA characteristic C2'-endo substate in regime I, while the O1'-endo population (closer to A-form) increases. Upon phase transition, we find the opposite tendency, analogously to our observations on the adaption of the helical structure. Furthermore, we have quantified changes in the α/γ and BI/BII population along unwinding. In regime I, unwinding results in an increase of BI states, which is in agreement with equilibrium

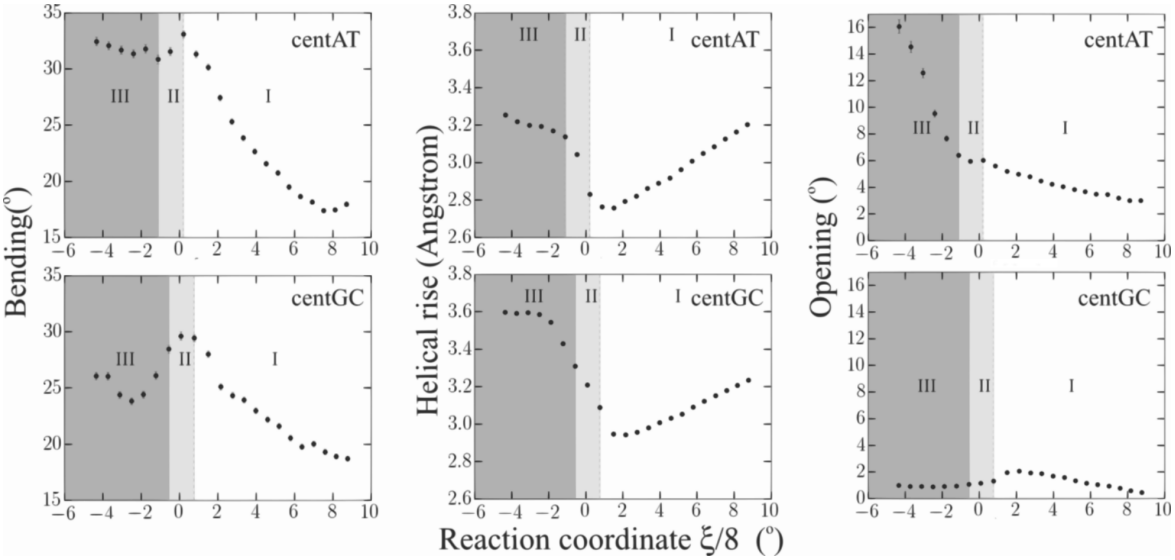


Figure 4.6: Changes in global bending (left), helical rise (middle) and base-pair opening angle (right) induced during HREUS simulations for the UB case. The plot was generated in a similar way as described in the caption of Fig 4.4.

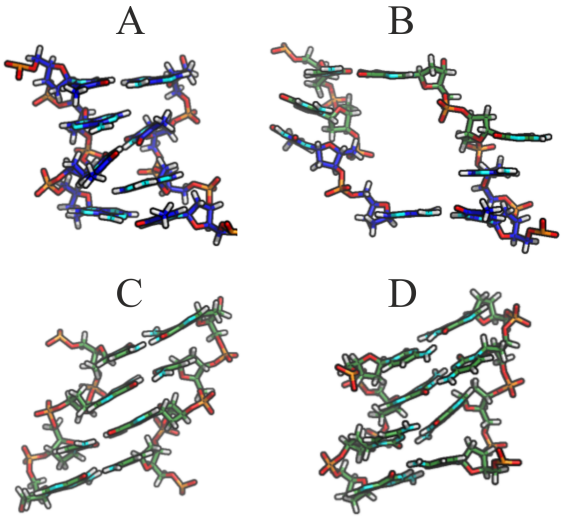


Figure 4.7: Structural snapshots (four base-pairs) taken from HREUS windows in regime III, A:T base-pairs are shown in blue (panel A,B), G:C base-pairs in green (panel C,D).

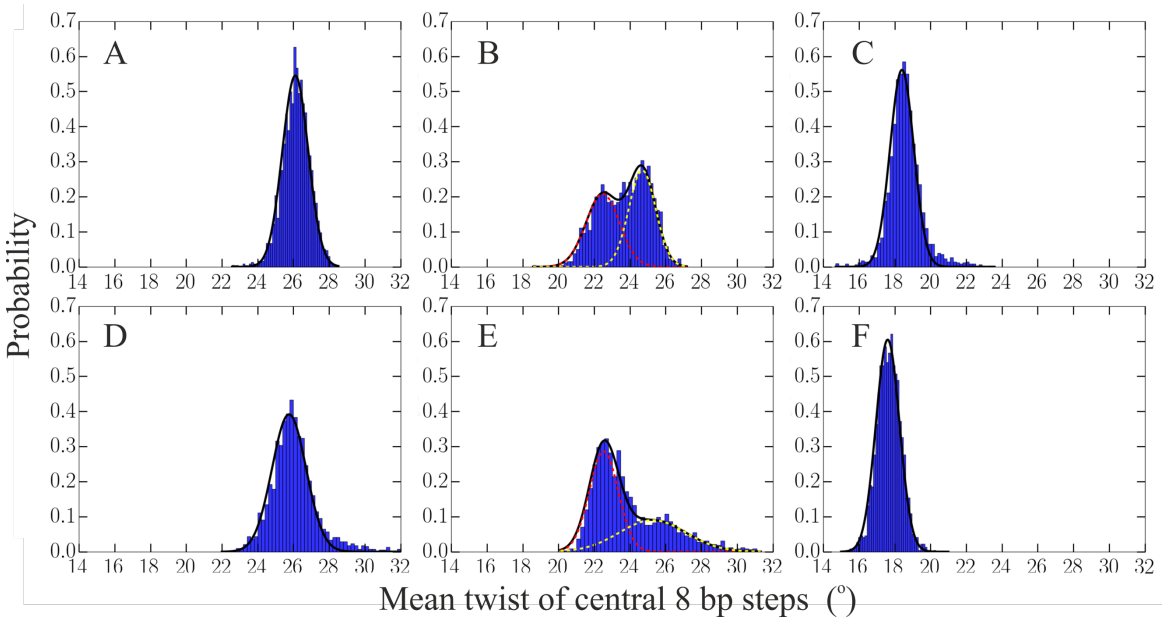


Figure 4.8: Probability distribution of the mean twist (average over the central 8 base-pair steps) in different HREUS windows. Narrow Gaussian type distributions are sampled in regime I (A:centAT, D:centGC), whereas multimodal distributions are obtained in the transition phase, regime II (B:centAT, E:centGC). The sampling in regime III is narrow Gaussian type again (C:centAT, F:centGC).

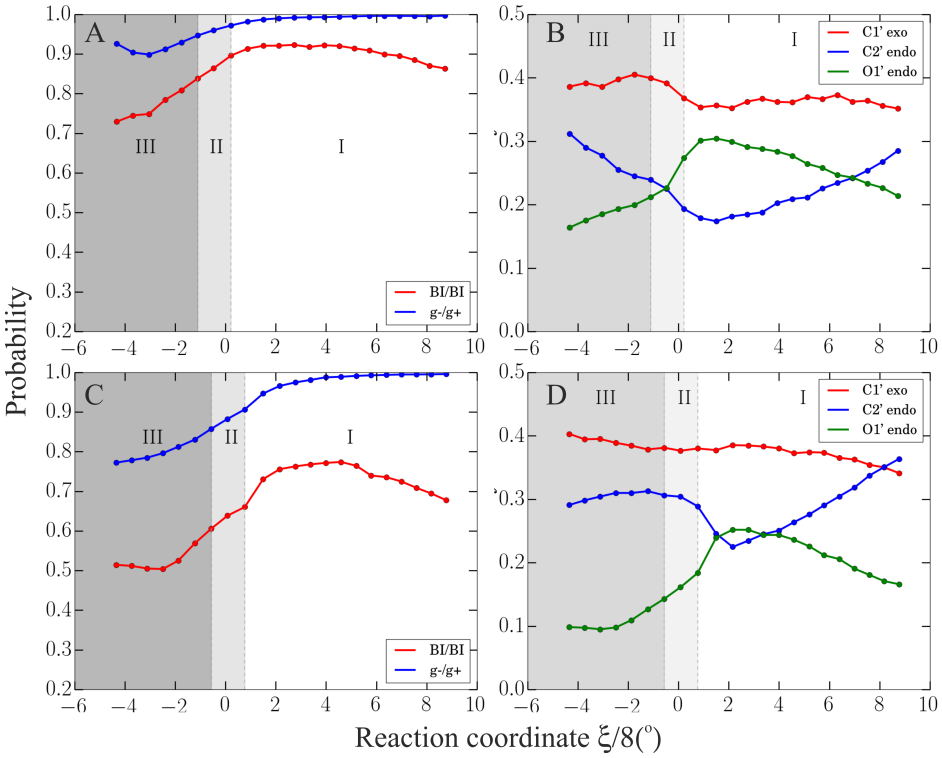


Figure 4.9: Relation between reaction coordinate ξ and sampled ϵ/ζ and α/γ backbone couplings (A:centAT, C:centGC). The population of sugar pucker states is given in (B:centAT) and (D:centGC). The data shows sampling of the central segment in the UB case.

MD simulations [19]. Entering regime II coincides with large changes in both populations, reflecting strong changes in the backbone structure. For the centGC sequence we obtain a higher affinity for α/γ flips (g-/g+ denotes the native state) compared to the centAT sequence. We therefore point out that GC-sequences absorb torsional stress predominantly through backbone-deformation, because α/γ flips are associated with large structural consequences.

4.3.5 Relaxation of Denatured DNA to Intact B-DNA

The calculated free energy curves indicate full reversibility with respect to unwinding, i.e. unwound DNA structures are not kinetically trapped. We have challenged this view by starting MD simulations from denatured DNA conformations (sampled in regime III) in the absence of torsional restraints (Fig 4.10). For the locally melted centAT sequence, the DNA recovers to fully regular B-DNA after only $\sim 50\,\text{ns}$. Moreover, the average twist relaxes to its equilibrium value after only $\sim 10\,\text{ns}$, thus showing that global helical structures relax on a shorter timescale than local distortions. The helical relaxation proceeds through sharp transient bending (up to $\sim 60°$) and a drop of the

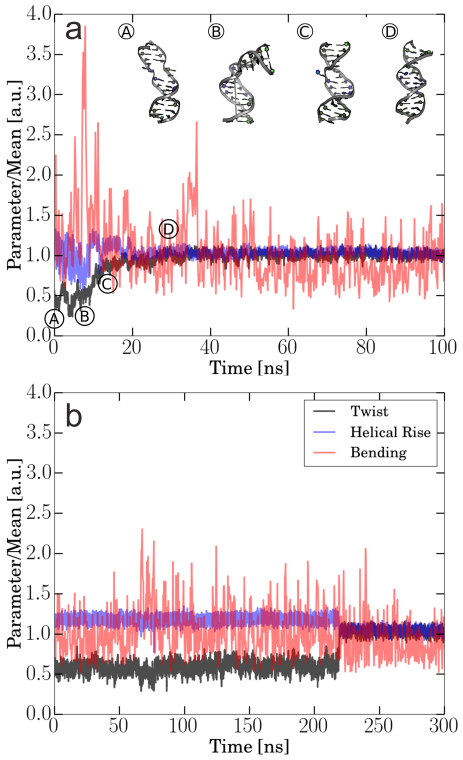


Figure 4.10: Relaxation of unwound DNA toward regular B-DNA in unrestrained MD simulations. (a) Time evolution of global parameters twist, helical rise and bending during 100 ns unrestrained MD simulation starting from a locally melted centAT structure. The parameters are given relative to their mean values. Characteristic snapshots are shown in the inset of the panel and corresponding simulation times are labeled accordingly (snapshots A-D). (b) Time evolution of global parameters by starting an unrestrained 300 ns long MD simulation from a strongly unwound centGC conformation.

helical rise. Intriguingly, this reflects the transition from regime I to II as sampled in our HREUS simulations. Once the helical structure has relaxed, it remains intact, hence enabling also local deformations to decay. Instead, for the centGC case, denatured structures mostly remained base-paired with a higher amount of backbone irregularities. Relaxation from these structures to B-DNA occurred in abrupt transitions on longer timescales (> 200 ns).

4.4 Conclusion

DNA unwinding is a central mechanism in many biological processes. It is implicated in supercoiling, packing, damage repair and even proposed to be a key determinant in transcription and replication as it promotes access to DNA's single strands [1–5]. In this

study, we have investigated DNA unwinding by means of Hamiltonian Replica Exchange Umbrella Sampling simulations on two 15 bp long sequences varying in the AT-content of the central segment. We have found that DNA unwinding can be distinguished in three regimes: Close to equilibrium, regime I, unwinding results in quadratic changes in the free energy. Thus, DNA's deformability can be described through a harmonic model [28, 30, 32]. By employing different kinds of harmonic models, we have figured out that unwinding stress is absorbed by untwisting. Here, however, the coupling to orthogonal degrees of freedom, mostly bending, plays an important role. In line with this, we find that simulations conducted under bending restriction also show a stiffer response of the DNA to unwinding. Furthermore, we see that a collapse in the harmonic response coincides with a phase transition to regime II. In this phase, the DNA begins to denature locally, coupling between twist and bending breaks down, which leads to a sudden helical extension, stochastic kinking and an overshoot in untwisting. Upon further unwinding, the DNA equilibrates to a continuous phase, regime III. Calculated PMFs indicate full reversibility and show that further unwinding requires only minimal free energy. Importantly, torsional stress leads to local strand dissociation in an AT-rich sequence. For GC bps, however, stress is mainly absorbed in deformation of the backbone and stacking perturbation. The reversible nature of DNA unwinding has been confirmed by starting unrestrained simulations from denatured structures, where highly damaged structures relax back to B-DNA on a timescale of $50 - 250\,\mathrm{ns}$. In this study, we have characterized the unwinding process of DNA, discussed induced sequence-dependent structural defects and revealed the impact of bending flexibility. So far, however, we have focused on aspects of DNA's deformability and local effects. In a follow-up project (chapter 6), we have studied longer DNA duplexes, hence also clarifying the impact of local melting on DNA's global behavior and coupling between distant sites.

4.5 Bibliography

[1] Christophe Lavelle. DNA Torsional Stress Propagates Through Chromatin Fiber and Participates in Transcriptional Regulation. *Nat. Struct. Mol. Biol.*, 15(2):123–125, 2008.

[2] Tae-Kyung Kim, Richard H Ebright, and Danny Reinberg. Mechanism of ATP-Dependent Promoter Melting by Transcription Factor IIH. *Science*, 288(5470):1418–1421, 2000.

[3] Yogambigai Velmurugu, Xuejing Chen, Phillip Slogoff Sevilla, Jung-Hyun Min, and Anjum Ansari. Twist-Open Mechanism of DNA Damage Recognition by the Rad4/XPC Nucleotide Excision Repair Complex. *Proc. Natl. Acad. Sci. USA*, 113(16):2296–2305, 2016.

[4] Terence R. Strick, Vincent Croquette, and David Bensimon. Single-Molecule Analysis of DNA Uncoiling by a Type II Topoisomerase. *Nature*, 404(6780):901–904, 2000.

[5] Christopher Collier, Cristina Machn, Geoff S Briggs, Wiep Klaas Smits, and Panos Soultanas. Untwisting of the DNA Helix Stimulates the Endonuclease Activity of Bacillus Subtilis Nth at AP Sites. *Nucleic Acids Res.*, 40(2):739–750, 2011.

[6] Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, and Richard S Mann. Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.*, 79:233–269, 2010.

[7] TR Strick, JF Allemand, D Bensimon, A Bensimon, and V Croquette. The Elasticity of a Single Supercoiled DNA Molecule. *Science*, 271(5257):1835–1837, 1996.

[8] T R Strick, J F Allemand, D Bensimon, and V Croquette. Behavior of Supercoiled DNA. *Biophys. J.*, 74(4):2016–2028, 1998.

[9] Zev Bryant, Michael D Stone, Jeff Gore, Steven B Smith, Nicholas R Cozzarelli, and Carlos Bustamante. Structural Transitions and Elasticity from Torque Measurements on DNA. *Nature*, 424(6946):338–341, 2003.

[10] Bryant S Fujimoto and J Michael Schurr. Dependence of the Torsional Rigidity of DNA on Base Composition. *Nature*, 344(6262):175–178, 1990.

[11] DP Millar, RJ Robbins, and AH Zewail. Torsion and Bending of Nucleic Acids Studied by Subnanosecond Time-Resolved Fluorescence Depolarization of Intercalated Dyes. *J. Chem. Phys.*, 76(4):2080–2094, 1982.

[12] TR Strick, J-F Allemand, D Bensimon, and V Croquette. Stress-Induced Structural Transitions in DNA and Proteins. *Annu. Rev. Biophys.*, 29(1):523–543, 2000.

[13] J David Moroz and Philip Nelson. Torsional Directed Walks, Entropic Elasticity, and DNA Twist Stiffness. *Proc. Natl. Acad. Sci. USA*, 94(26):14418–14422, 1997.

[14] Stefanos K Nomidis, Franziska Kriegel, Willem Vanderlinden, Jan Lipfert, and Enrico Carlon. Twist-Bend Coupling and the Torsional Response of Double-Stranded DNA. *Phys. Rev. Lett.*, 118(21):217801, 2017.

[15] Irene Collins, Achim Weber, and David Levens. Transcriptional Consequences of Topoisomerase Inhibition. *Mol. Cell Biol.*, 21(24):8437–8451, 2001.

[16] Franziska Kriegel, Niklas Ermann, and Jan Lipfert. Probing the Mechanical Properties, Conformational Changes, and Interactions of Nucleic Acids with Magnetic Tweezers. *J. Struct. Biol.*, 197(1):26 – 36, 2017.

[17] Jan Lipfert, Jacob WJ Kerssemakers, Tessa Jager, and Nynke H Dekker. Magnetic Torque Tweezers: Measuring Torsional Stiffness in DNA and RecA-DNA Filaments. *Nat. Methods*, 7(12):977–980, 2010.

[18] John F Marko and Eric D Siggia. Fluctuations and Supercoiling of DNA. *Science*, 265(5171):506–508, 1994.

[19] Marco Pasi, John H. Maddocks, David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham, III, Pablo D. Dans, B. Jayaram, Filip Lankaš, Charles Laughton, and et al. μABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. *Nucleic Acids Res.*, 42(19):12272–12283, 2014.

[20] David L Beveridge, Thomas E Cheatham, and Mihaly Mezei. The ABCs of Molecular Dynamics Simulations on B-DNA. *J. Biosci.*, 37(3):379–397, 2012.

[21] Pablo D Dans, Alberto Pérez, Ignacio Faustino, Richard Lavery, and Modesto Orozco. Exploring Polymorphisms in B-DNA Helical Conformations. *Nucleic Acids Res.*, 40(21):10668–10678, 2012.

[22] Pablo Daniel Dans, Ignacio Faustino, Federica Battistini, Krystyna Zakrzewska, Richard Lavery, and Modesto Orozco. Unraveling the Sequence-Dependent Polymorphic Behavior of d (CpG) Steps in B-DNA. *Nucleic Acids Res.*, 42(18):11304–11320, 2014.

[23] Pablo D Dans, Linda Danilane, Ivan Ivani, Tomáš Dršata, Filip Lankas, Jürgen Walther, Ricard Illa Pujagut, Federica Battistini, Josep Lluis Gelpí, and Richard Lavery. Long-Timescale Dynamics of the Drew–Dickerson Dodecamer. *Nucleic Acids Res.*, 44(9):4052–4066, 2016.

[24] Alexandra Balaceanu, Marco Pasi, Pablo D Dans, Adam Hospital, Richard Lavery, and Modesto Orozco. The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, 8(1):21–28, 2016.

[25] Srinivasaraghavan Kannan, Kai Kohlhoff, and Martin Zacharias. B-DNA Under Stress: Over- and Untwisting of DNA during Molecular Dynamics Simulations. *Biophys. J.*, 91(8):2956 – 2965, 2006.

[26] Korbinian Liebl, Tomáš Dršata, Filip Lankaš, Jan Lipfert, and Martin Zacharias. Explaining the Striking Difference in Twist-Stretch Coupling between DNA and RNA: A Comparative Molecular Dynamics Analysis. *Nucleic Acids Res.*, 43(21):10143–10156, 2015.

[27] Filip Lankaš, Jiří Šponer, Jörg Langowski, and Thomas E Cheatham. DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.*, 85(5):2872–2883, 2003.

[28] Tomáš Dršata, Alberto Pérez, Modesto Orozco, Alexandre V Morozov, Jiří Šponer, and Filip Lankaš. Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, 9(1):707–721, 2013.

[29] F Lankaš, Oscar Gonzalez, LM Heffler, G Stoll, M Moakher, and John H Maddocks. On the Parameterization of Rigid Base and Basepair Models of DNA from Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.*, 11(45):10565–10588, 2009.

[30] Tomáš Dršata, Nad'a Špačkova, Petr Jurečka, Marie Zgarbová, Jiří Šponer, and Filip Lankaš. Mechanical Properties of Symmetric and Asymmetric DNA A-tracts: Implications for Looping and Nucleosome Positioning. *Nucleic Acids Res.*, 42(11):7383–7394, 2014.

[31] Emmanuelle Bignon, Tomáš Dršata, Christophe Morell, Filip Lankaš, and Elise Dumont. Interstrand Cross-Linking Implies Contrasting Structural Consequences for DNA: Insights from Molecular Dynamics. *Nucleic Acids Res.*, 45(4):2188–2195, 2017.

[32] Filip Lankaš, Jiří Šponer, Pavel Hobza, and Jörg Langowski. Sequence-Dependent Elastic Properties of DNA. *J. Mol. Biol.*, 299(3):695–709, 2000.

[33] DA Case, V Babin, JT Berryman, RM Betz, Q Cai, DS Cerutti, TE Cheatham III, TA Darden, RE Duke, and H Gohlke. *Amber14*. University of California, 2014.

[34] In Suk Joung and Thomas E Cheatham III. Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters. *J. Phys. Chem. B*, 113(40):13279–13290, 2009.

[35] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[36] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiří Šponer, Thomas E Cheatham, Charles A Laughton, and Modesto Orozco. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.*, 92(11):3817–3829, 2007.

[37] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. *J. Molec. Graphics*, 14(1):33–38, 1996.

[38] R Lavery, M Moakher, JH Maddocks, D Petkeviciute, and K Zakrzewska. Conformational Analysis of Nucleic Acids Revisited: Curves+. *Nucleic Acids Res.*, 37(17):5917–5929, 2009.

[39] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[40] Fangqiang Zhu and Gerhard Hummer. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.*, 33(4):453–465, 2012.

[41] Daniel Strahs and Tamar Schlick. A-Tract Bending: Insights into Experimental Structures by Computational Models. *J. Mol. Biol.*, 301(3):643–663, 2000.

[42] J Michael Schurr. Possible Origin of the Increased Torsion Elastic Constant of Small Circular DNAs: Bending-Induced Axial Tension. *J. Phys. Chem. B*, 121(23):5709–5717, 2017.

[43] Jan Lipfert, Gary M Skinner, Johannes M Keegstra, Toivo Hensgens, Tessa Jager, David Dulin, Mariana Köber, Zhongbo Yu, Serge P Donkers, and Fang-Chieh Chou. Double-Stranded RNA under Force and Torque: Similarities to and Striking Differences from Double-Stranded DNA. *Proc. Natl. Acad. Sci. USA*, 111(43):15408–15413, 2014.

5 How Methyl-Sugar Interactions Determine DNA Structure and Flexibility [1]

5.1 Introduction

The backbone of DNA has a large impact on DNA's structure and flexibility and plays an important role in the recognition by proteins [1–3]. Thereby, the ϵ and ζ dihedral angles define the most prominent conformational polymorphism: The BI and BII substates. The population of these substates not only determines the bimodal distribution of base-pair step parameters or is altered through the binding of proteins, but it is also strongly dependent on the base-pair sequence [4–19]. By means of MD simulations, it has recently been shown that the formation of unconventional hydrogen bonds between base and backbone atoms (C8-H8...O3' between purine base (R) and backbone for RpR and YpR steps, C6-H6...O3' between pyrimidine base (Y) and backbone for RpY and YpY steps) correlates excellently with population of the BII state [15, 19]. Thus, the differential formation of these hydrogen bonds serves as an explanation for the sequence-dependent population of BI/BII states. However, in principle all DNA bases could build such contacts and the hierarchy of the hydrogen bonds itself is unclear, hence a sterical explanation has still remained elusive. A major unexplained phenomenon in this respect is that methyl groups (in thymine or C5-methylated cytosine) stabilize the BI opposed to BII substates [20–22]. Three major mechanisms are hypothetically causal to this interdependency: First, the methyl groups are hydrophobic and directed towards the major groove. Therefore, it has been argued that an altered hydration pattern in the major groove is a central effect of DNA methylation [23, 24]. Furthermore, BI states may be stabilized by a water molecule bridging between methyl-carbon and a phosphate bound oxygen [22]. This hypothesis, however, has been revised by Wibowo et al in a follow-up study, suggesting that methyl group induced BI stabilization results from an increased mean water residence time around base atoms [25].
Second, steric repulsion between the methyl group and sugar ring of the 5'-neighboring base may be causal. The existence of this effect has been hypothesized by Hunter [26] and more recently in a MD based study by Peguero-Tejada and coworkers [21].

[1] This chapter has been previously published in similar form in Nucleic Acids Res., 47(3):1132-1140, 2018.

Third, BI states are stabilized through methyl-π stacking. Interactions between the thymine methyl group and a 5'-neighboring base are attractive, which has been argued to be strongly implicated in DNA's deformability [27–29].
We have investigated the molecular mechanisms of a methyl group in thymine and in C5-methylated cytosine through comparative MD simulations, during which we have included or explicitly omitted nonbonded interactions between the methyl group and solvent or other parts of the DNA. The simulations reveal that neither methyl-π stacking nor methyl-solvent interactions have considerable impact on DNA's backbone structure. Switching off interactions between methyl group and the sugar C2' atom of the 5'-neighboring nucleotide and its hydrogen atoms, however, significantly increases the population of BII states. We further show that under these conditions the degree of unconventional hydrogen bonding between nucleobase and backbone is also strongly increased. The drastic effects on the BI/BII ratio imply a strong impact on DNA's structure and flexibility: Methyl-sugar clashes increase DNA's intrinsic bending but considerably reduce its stiffness. Bending and stretching stiffness are reduced by up to $\sim 40\%$ when methyl-sugar clashes are allowed, and for GpT and CpT steps the twist stiffness can lower even about $\sim 60\%$. Given the substantial impact of methyl-sugar clashes on DNA's local structure and flexibility, we suppose that this atomistic effect is causal to many genetic functions, e.g. C5-methylation of cytosine is a central mechanism in regulating gene expression. Moreover, thymine's methyl group and the chemical environment of the C2' atom are the only difference between RNA and DNA. In this study, we propose a fundamental interaction between both moieties.

5.2 Materials and Methods

5.2.1 Force Field Modifications

All starting structures were generated with the nab module of the Amber16 package [30]. In total, seven different 15bp long DNA duplexes have been studied. The DNA structures were neutralized with potassium ions and solvated to explicit solvent (TIP3P water model [31]) with the xleap module. The minimum distance between box boundary and DNA was thereby set to $10\mathring{A}$, and parmbsc1 was selected as force field [32]. For the base atoms of C5-methylated cytosine, the parameters by Rauch et al, which served as extension to the parm99 force field, were employed [22]. Updates to the parm99 force field concerned only torsional backbone angles and not nonbonded parameters. Thus, the parameters by Rauch et al for C5-methylated cytosine are fully compatible with the bsc1 parameters. Classical force fields have the form:

$$U(\mathbf{r}_1, ..., \mathbf{r}_N) = \sum_{Bonds} k_d \cdot (d - d_0)^2 + \sum_{Angles} k_\theta \cdot (\theta - \theta_0)^2 + \\ + \sum_{Dihedrals} \frac{V_n}{2} \cdot [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon \cdot r_{ij}} \right]. \tag{5.1}$$

The last sum represents Coulomb and van-der-Waals interactions, which describe all nonbonded interactions. Besides regular parameter topologies, also topology files with modified force field descriptions were generated. Thereby, nonbonded interactions between methyl groups and specific partner groups were switched off using the parmed module of Amber16 [30]. This required the removal of partial charges on the methyl group (C7, H71,H72,H73) and redistribution of the base atoms' charges according to the demethylated analogs. Furthermore, the pairwise van-der-Waals parameters between methyl and defined partner group were set to zero. Thus, three additional topologies were generated for each sequence, neglecting either:

- interactions between methyl groups and all water molecules
- interactions between methyl groups and the C2' and its hydrogen atoms of the 5' neighboring sugar
- or interaction between methyl groups and their 5' neighboring base atoms.

5.2.2 Simulation Setup and Equilibration

Initially, the systems were energy minimized (steepest descent method) in 2500 steps with the sander module of the Amber16 package [30]. All subsequent MD simulations were carried out with the pmemd.cuda module. First, systems were heated up to 300 K in three stages. Each stage induced a temperature increase of 100 K and was simulated for 100 ps including positional restraints (with respect to the starting structure) on all of DNA's non-hydrogen atoms. Afterwards, the restraints were gradually reduced from 25 $\frac{\text{kcal}}{\text{mol}\text{Å}^2}$ to 0.5 $\frac{\text{kcal}}{\text{mol}\text{Å}^2}$ in five consecutive simulations at 300 K and at a constant pressure of 1 bar (weak coupling with a time constant of 5 ps). The equilibration was completed with a 2 ns simulation in which only the first two base-pairs were positionally restrained with a small force constant of 0.1 $\frac{\text{kcal}}{\text{mol}\text{Å}^2}$. The output structures served as input for the production runs, where we maintained the soft restraints on the first two base-pairs. Data gathering simulations were performed for 900-4000 ns, and coordinates were written out every 5000 steps. Activation of hydrogen-mass-repartitioning allowed a time step of 4 fs.

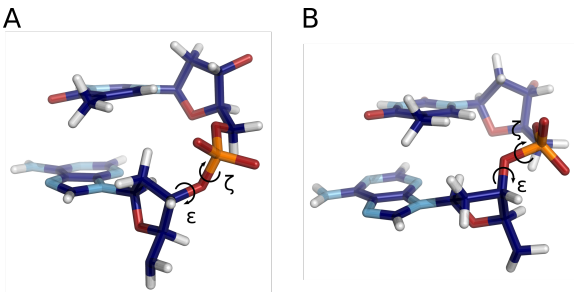


Figure 5.1: ApT base-pair step in BI (A) and in BII conformation (B), respectively.

DNA-sequence	abbrev.	P(BI) [%]
$5' - CGCGC\mathbf{ATATA}CGCGC - 3'$	AT	83.6
$5' - CGCGC\mathbf{AUAUA}CGCGC - 3'$	AU	75.5
$5' - CGCGC\mathbf{GCGCG}CGCGC - 3'$	CG	73.4
$5' - CGCGC\mathbf{GC^*GC^*G}CGCGC - 3'$	C*G	78.8
$5' - CGCGC\mathbf{AAAAA}CGCGC - 3'$	AA	88.3
$5' - CGCGC\mathbf{TCTCT}CGCGC - 3'$	CT	77.2
$5' - CGCGC\mathbf{TGTGT}CGCGC - 3'$	GT	76.4

Table 5.1: Sequences of the studied DNA duplexes and population of BI states. C* denotes methylated cytosine

5.3 Results and Discussion

5.3.1 How the Methyl Group Affects DNA's Backbone

In a dinucleotide step, a DNA backbone strand can either adopt BI or BII conformations which are determined by the ϵ and ζ dihedral angles (Fig 5.1):

$$\epsilon - \zeta < 0 : \text{BI}, \quad \epsilon - \zeta > 0 : \text{BII}. \tag{5.2}$$

We have performed MD simulations on seven different DNA duplexes to capture the sequence dependence of these backbone states. The computed probabilities for the BI state as a function of the sequence are summarized in Table 5.1.

We notice two major findings: First, increasing the A:T content results in a higher population of the BI substate (by $\sim 10 - 15\%$). Second, replacing thymine or C5-methylated cytosine by their demethylated analogs destabilizes the BI state ($\sim 5 - 8\%$). Thus, our MD simulations point out that the methyl group in thymine and C5-methylated cytosine stabilizes the BI conformation. Similar tendencies have also been reported in other studies [21]. In order to reveal the molecular mechanism behind this stabilization,

we have performed MD simulations of the AT-sequence (central ATATA segment, Table 5.1) during which we have excluded physical interactions between the methyl group and specific atom groups in the system. We have either turned off all interactions of the methyl group with solvent molecules, with the atoms of the 5'-neighboring base or with the C2' and its hydrogen atoms of the 5' sugar. Corresponding free energy profiles along the $\epsilon - \zeta$ coordinate have been computed by Boltzmann inversion of the sampled probability distributions which represent averages over the central DNA segment (steps 6 to 9). Intriguingly, the exclusion of interactions between methyl groups and all water molecules (Fig 5.2, purple curve) or the atoms of the 5' neighboring bases (Fig 5.2, yellow curve) results in negligible changes with respect to standard conditions (Fig 5.2, blue curve). This means that neither the hydrophobic nor the methyl-π stacking effect impacts DNA's backbone structure. However, turning off interactions between methyl groups and the C2' atom and its hydrogens causes a large drop of the free energy in the BII region (Fig 5.2, red curve). Thus, we conclude that the steric methyl sugar clashes trigger BI promiscuity of C5-methylated sequences.

5.3.2 Why Methyl-Sugar Clashes Trigger BI Promiscuity

Methyl-sugar clashes can only occur in base-pair steps if the thymine (or C5-methylated cytosine) base is at the 3' position (e.g. in ApT but not in TpA steps). In the ApT steps, the BII states are even more stabilized (by $\sim 2\,\mathrm{kcal/mol}$) through the exclusion of methyl-sugar interactions than expected from the free energy profiles obtained as averages over the central segments (compare Fig 5.2 with Fig 5.3). On the contrary, for adjacent TpA steps the exclusion results in a remarkable destabilization of BII states ($\sim 0.5\,\mathrm{kcal/mol}$). Note that this behavior reflects the anticorrelation of DNA base-pair steps and has been reported in previous studies [18, 33]. Here, we provide a qualitative sterical explanation for this nearest-neighbor anticorrelation, see Fig 5.3. When an ApT step is in BI configuration, the 3' sugar (T nucleotide) preferentially adopts a lower pucker phase than the 5' sugar (A nucleotide). On the other hand, if a base-pair step is in BII configuration, the 3' sugar preferentially adopts a higher pucker phase than the 5' sugar. Note that the 5' sugar adopts the role of a 3' sugar in the consecutive step. Having shifted the sugar to a lower phase due to the BII conformation therefore suppresses another BII conformation in the neighboring step.

Based on our MD simulations, we can further illustrate the methyl-group induced BII destabilization (Fig 5.4). In order to avoid clashing with the 5' sugar, the bases are locked to a conformational subspace where the backbone preferentially adopts the BI configuration. Switching off this sterical hindrance allows both components to come closer together and hence also the BII subspace becomes accessible. These states are then stabilized by unconventional hydrogen bonds between thymine's H6-atom and the O3'-atom of the backbone.

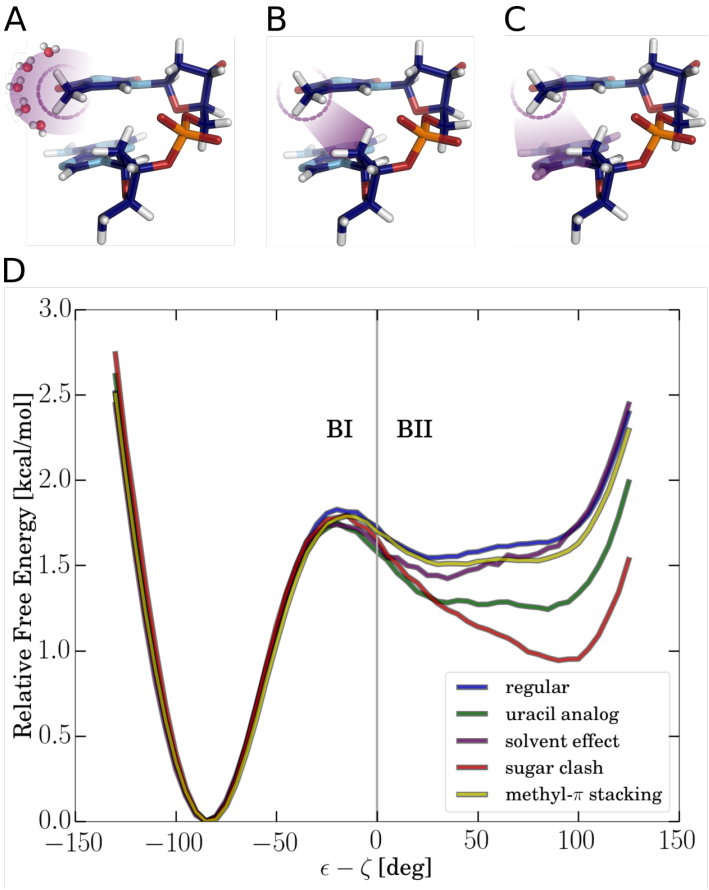


Figure 5.2: Sterical effects which may hypothetically trigger BI promiscuity: Hydrophobic interactions, methyl-sugar clashes and methyl$-\pi$ stacking (A-C). Illustration for excluding interactions of the methyl group with all water molecules (A), with the 5'-neighboring C2' atom and its hydrogens (B) and with the atoms of the 5'-neighboring base (C). (D) Free energy profiles along the $\epsilon - \zeta$ coordinate for the AT-sequence including all interactions or switching off specific interactions (indicated by different colors). The free energy curve for the uracil containing sequence (AU) is shown in green. The free energies along the $\epsilon - \zeta$ coordinate have been calculated by Boltzmann inversion, $F = -k_B T \cdot \ln(p)$, and as average over all dinucleotide steps in the central segments.

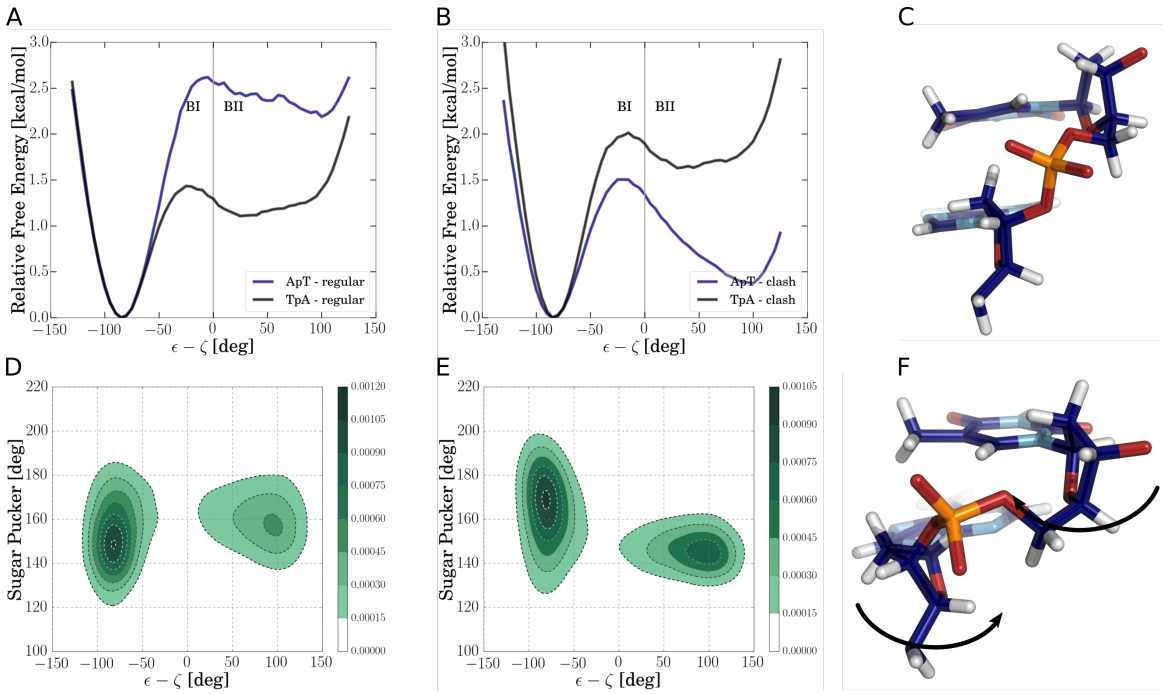


Figure 5.3: Calculated free energies along $\epsilon - \zeta$ for individual base-pair steps in the AT-sequence. (A) Free energy profiles obtained under regular conditions. (B) Free energy profiles obtained from simulations with exclusion of methyl-sugar interactions. (C) Snapshot of a BI conformation. (D) Normalized density plot of 3'-sugar's pucker phase versus $\epsilon - \zeta$. Here, BI states prefer a lower phase compared to BII states. (E) Normalized density plot of 5'-sugar's pucker phase versus $\epsilon - \zeta$. In this case, BI states prefer a higher phase compared to BII states. (F) Snapshot of a BII conformation with arrows indicating the shift in the puckering phases.

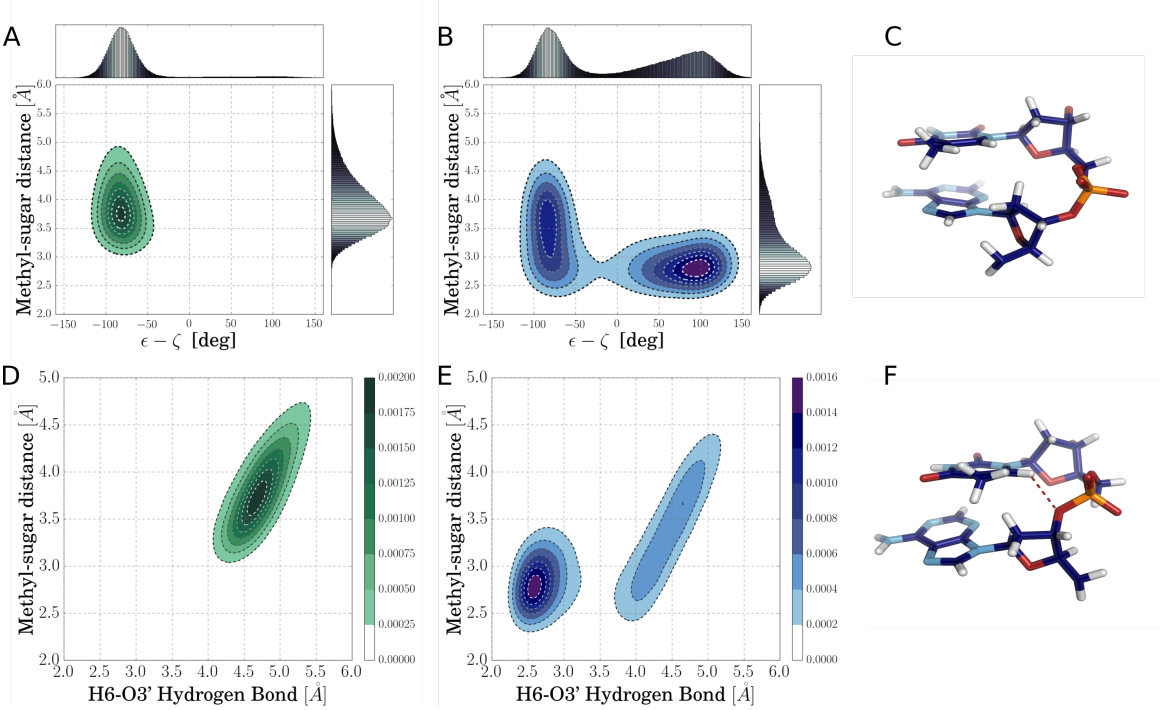


Figure 5.4: Methyl-sugar clashes destabilize the BII subspace in ApT steps. (A) Center of mass distance between methyl group and C2' atom of 5'-neighboring base as a function of $\epsilon - \zeta$, sampled in the regular MD simulation. BII states are populated very rarely. (B) Same as in (A), but methyl-sugar interactions have been excluded, hence both groups can approach each other making the BII subspace accessible. (C) BI conformation, taken from regular MD simulation. (D) Center of mass distance between methyl group and C2' atom of 5'-neighboring base versus H6-O3' distance. The normalized density has been computed from regular MD simulations and shows no unconventional hydrogen bonding. (E) Same as in D, but for simulations during which methyl-sugar interactions have been excluded. Methyl and 5'-sugar group can come close to each other, thereby also giving rise to unconventional H6-O3' hydrogen bonding. (F) Snapshot of a BII conformation sampled during MD simulation with excluded methyl-sugar interactions. The dashed red line illustrates the unconventional H6-O3 hydrogen bond, which stabilizes the BII subspace.

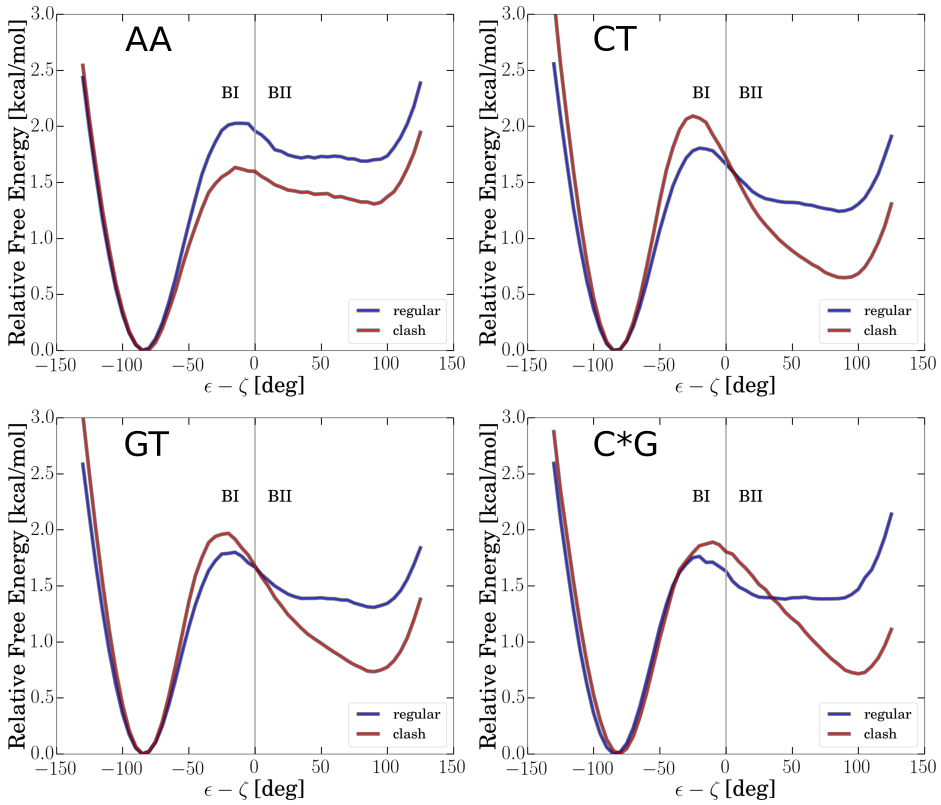


Figure 5.5: Free energy along $\epsilon - \zeta$ for the AA, CT, GT and C*G sequences. Blue lines correspond to simulations without the exclusion of specific nonbonded interactions, and the red lines to simulations with excluded methyl-sugar interactions.

In addition, we have studied the effect of methyl-sugar clashes for other sequences (Fig 5.5). Overall, we find that methyl-sugar clashes destabilize BII states for each of the investigated sequences. For the CT, GT and C*G sequences, this effect turned out to be even stronger than for the AT-sequence.

The sequence dependence of DNA backbone substates has also been studied with NMR experiments. Thereby, it has been reported that out of the ten dinucleotide steps the four which contain a thymine on the 3'-position (ApT, GpT, TpT and CpT) clearly exhibit the lowest BII population [2]. This confirms our results remarkably well and emphasizes that the effect of methyl-sugar clashes is indeed realistic.

5.3.3 Methyl-Sugar Clashes Influence DNA's Global Structure and Flexibility

The conformation of DNA's backbone is strongly coupled to base-pair stacking [1, 2, 34, 35]. In the following, we quantify the impact of methyl-sugar clashes on DNA's structure and flexibility by considering mean twist, stretch and bending of the central DNA segments as relevant parameters (Fig 5.6). Bending has been computed as curvature as introduced in chapter 4. While we find only negligible changes in mean twist and stretching, methyl-sugar clashes cause a pronounced increase in the intrinsic bending of the DNA double helix (up to $\sim 18\%$ for methylated cytosine tracts). We have calculated the stiffnesses based on the harmonic approximation, $K = k_B T C^{-1}$ [36–40]. Here, we have computed the stiffnesses K directly by inversion of the parameters variance (denoted by C). In general, methyl-sugar repulsion has a strong softening effect: Excluding methyl-sugar interactions decreases the bending stiffness by $\sim 40\%$ for C*G and by $\sim 20-30\%$ for the GT and CT sequence. The stretching stiffness is also reduced (by up to $\sim 35\%$), while the twist stiffness decreases drastically for the CT and GT sequence ($\sim 50-60\%$). Nevertheless, we emphasize that these findings do not necessarily indicate that methylated sequences are more flexible than their demethylated analogs, as we have studied only the direct impact of methyl-sugar interactions. Van-der-Waals interactions between methyl groups and other chemical groups are still included.
The impact of methylation on stiffness has been investigated in previous studies with different conclusions having been made [40–42]. Based on our MD simulations, we have made the following observations: GpC* steps are stiffer than GpC, whereas C*pG are generally more flexible than CpG. Globally, methylated cytosine sequences exhibit a lower stretching and bending stiffness ($\sim 5-7\%$), but they are much stiffer in response to twisting ($\sim 20\%$).
Note that the increase in flexibility induced by methyl-sugar clashes is counterintuitive. However, we argue that this is a direct influence of the backbone: Within the BI subspace, the twist distribution is broader than in the BII subspace (Fig 5.7). An increased population of BII states may therefore result in a narrower overall twist distribution, and this lower variance directly relates to a higher stiffness.
Furthermore, one might also expect from the $\epsilon - \zeta$ free energy profiles that the twist distribution of BI is narrower. However, the backbone states do not map linearly to stacking geometry, as for instance similar $\epsilon - \zeta$ conformations can be coupled to entirely different twisting (illustrated in Fig 5.7).

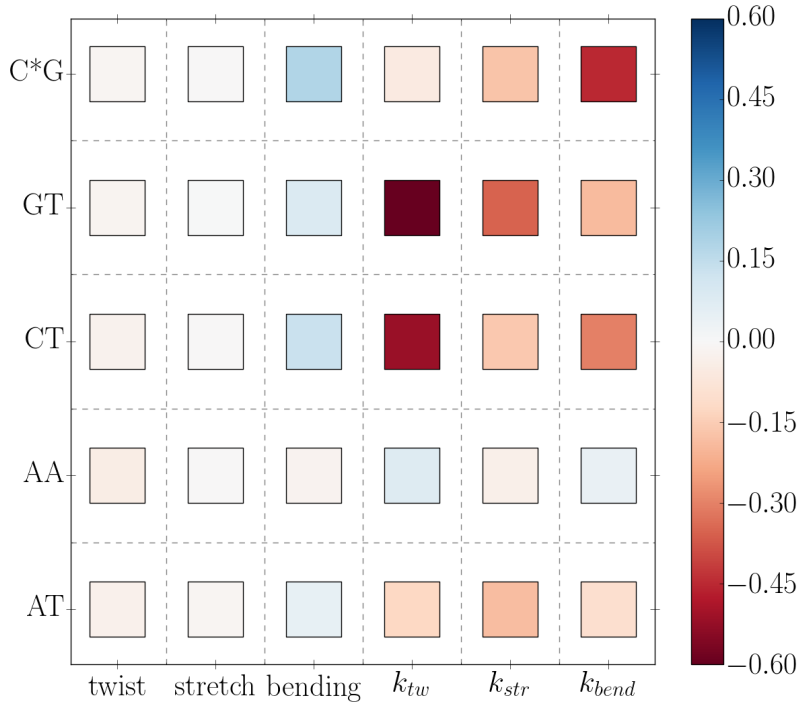


Figure 5.6: Relative changes in structure and flexibility due to methyl-sugar clashes. The first three columns show changes in equilibrium twist, stretch and bending of the central segment, whereas the last three columns represent changes in the stiffness. For the latter cases, red entries indicate that methyl-sugar clashes have a decreasing/softening effect, and blue entries reflect increased stiffness. All changes are given as relative to the reference case of reassignment of thymines' and methylated cytosines' charges, i.e. the entries depict the effect of van-der-Waals interactions between methyl and sugar group.

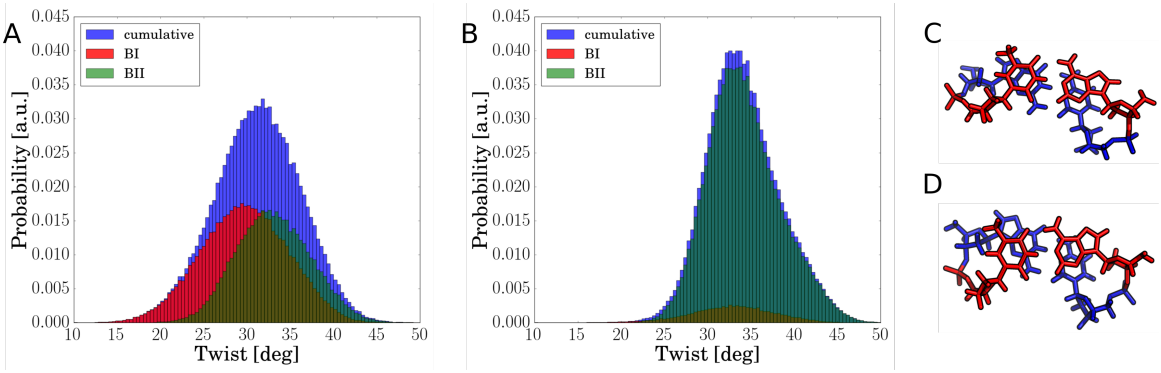


Figure 5.7: Methyl-sugar clashes increase DNA's flexibility. (A) Calculated twist distribution of a GpT step in case of simulations including methyl-sugar interactions. Distribution of the BI and BII states as well as the cumulative distribution are shown separately. The distribution of BI states is shifted and broader (standard deviation $\sigma = 7.2°$) than that for BII states ($\sigma = 6.2°$). The standard deviation of the cumulative distribution amounts to $\sigma = 7.1°$. (B) Similar as (A), but with excluded methyl-sugar interactions. As a consequence of the higher BII population, the variance of the cumulative distribution becomes lower ($\sigma = 5.7°$). (C) Snapshot of an undertwisted GpT step ($\sim 20°$) at $\epsilon - \zeta \cong -80°$ in Watson and Crick strand. (D) Snapshot of an overtwisted GpT step ($\sim 38.5°$) at $\epsilon - \zeta \cong -80°$ in both strands.

5.4 Conclusion

DNA methylation is a central epigenetic mechanism to regulate the expression of genes [43–45]. These biological functions are likely to be caused by the altered deformability of DNA [40, 41, 46]. In this study, we have shown that the methyl group in thymine and C5-methylated cytosine impacts the population of BI/BII backbone substates and consequently DNA's deformability. Performing comparative MD simulations, we have been able to identify methyl-sugar clashes as the sterical origin of the BI promiscuity induced by the methyl groups. Neither the exclusion of nonbonded interactions between methyl group and water molecules nor between methyl group and 5'-neighboring base atoms changes the population of BI states. However, switching off interactions between methyl group and the C2' and its hydrogen atoms stabilizes the BII subspace remarkably and hence destabilizes the BI subspace. Notably, thymine's methyl group and the absence of a hydroxyl group on the C2' atom are the only chemical difference between DNA and RNA. In the presented study we have pointed out that both components are strongly coupled, which regulates DNA's fine structure. This interaction has been found for each of the investigated sequences. Thereby, the strongest changes have been observed for sequences including CpT and GpT steps or methylated cytosine.
Previous studies have put forward the role of unconventional hydrogen bonds (H6-O3' or H8-O3') for stabilization of BII conformations [15, 19]. Within our study, we largely

agree with this view. Nevertheless, we emphasize that this is not sufficient to explain the sequence dependence of BI/BII population. This becomes clear from our simulations during which we have switched off methyl-sugar interactions: Allowing the clashes also suddenly leads to formation of these unconventional hydrogen bonds.
Besides, we have been able to give a sterical explanation for the anticorrelation of DNA's base-pair steps: The backbone states are tightly coupled to the sugar puckering. In adjacent base-pair steps, a given sugar ring adopts the opposite roles (5' or 3'), thus promoting opposite backbone behavior in both steps.
As a consequence of the large impact of methyl-sugar clashes on the BI/BII population, also DNA's global structure and flexibility is strongly affected. Methyl-sugar interactions increase DNA's intrinsic bending, but significantly decrease DNA's stiffness. For instance, the torsional rigidity of CpT and GpT rich sequences is decreased by $50-60\%$. This finding can be explained by the higher variance of the BI subspace, so that its stabilization results in an overall higher flexibility.
Finally, we emphasize that our simulation methodology of excluding specific nonbonded interactions may be applied in future studies to also elucidate other phenomena, e.g. the role of hydrogen bonding in structure formation.

5.5 Bibliography

[1] Brigitte Hartmann, Daniel Piazzola, and Richard Lavery. BI-BII transitions in B-DNA. *Nucleic Acids Res.*, 21(3):561–568, 1993.

[2] Brahim Heddi, Christophe Oguey, Christophe Lavelle, Nicolas Foloppe, and Brigitte Hartmann. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.*, 38(3):1034–1047, 2009.

[3] Martin J. Packer and Christopher A. Hunter. Sequence-dependent DNA structure: the role of the sugar-phosphate backbone. *J. Mol. Biol.*, 280(3):407 – 420, 1998.

[4] Tomáš Geršata, Alberto Pérez, Modesto Orozco, Alexandre V Morozov, Jiří Šponer, and Filip Lankaš. Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, 9(1):707–721, 2013.

[5] Tomáš Geršata and Filip Lankaš. Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3(4):355–363, 2013.

[6] Korbinian Liebl and Martin Zacharias. Unwinding Induced Melting of Double-Stranded DNA Studied by Free Energy Simulations. *J. Phys. Chem. B*, 121(49):11019–11030, 2017.

[7] Christophe Oguey, Nicolas Foloppe, and Brigitte Hartmann. Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions. *PLOS ONE*, 5(12):1–8, 2011.

[8] Fajar R. Wibowo, Christine Rauch, Michael Trieb, and Klaus R. Liedl. M.TaqI facilitates the base flipping via an unusual DNA backbone conformation. *Biopolymers*, 79(3):128–138, 2005.

[9] James C. Robertson and Thomas E Cheatham. DNA Backbone BI/BII Distribution and Dynamics in E2 Protein-Bound Environment Determined by Molecular Dynamics Simulations. *J. Phys. Chem. B*, 119(44):14111–14119, 2015.

[10] Rashmi S. Hegde, Steven R. Grossman, Laimonis A. Laimins, and Paul B. Sigler. Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DMA-binding domain bound to its DNA target. *Nature*, 359(6395):505–512, 1992.

[11] D. Djuranovic and B. Hartmann. DNA fine structure and dynamics in crystals and in solution: The impact of BI/BII backbone conformations. *Biopolymers*, 73(3):356–368, 2004.

[12] D. Djuranovic and B. Hartmann. Molecular Dynamics Studies on Free and Bound Targets of the Bovine Papillomavirus Type I E2 Protein: The Protein Binding Effect on DNA and the Recognition Mechanism. *Biophys. J.*, 89(4):2542 – 2551, 2005.

[13] Brahim Heddi, Nicolas Foloppe, Nadia Bouchemal, Edith Hantz, and Brigitte Hartmann. Quantification of DNA BI/BII Backbone States in Solution. Implications for

DNA Overall Structure and Recognition. *J. Am. Chem. Soc.*, 128(28):9170–9177, 2006.

[14] A. Madhumalar and Manju Bansal. Sequence Preference for BI/BII Conformations in DNA: MD and Crystal Structure Data Analysis. *J. Biomol. Struct. Dyn.*, 23(1):13–27, 2005.

[15] Marco Pasi, John H. Maddocks, David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham, III, Pablo D. Dans, B. Jayaram, Filip Lankaš, Charles Laughton, and et al. μABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. *Nucleic Acids Res.*, 42(19):12272–12283, 2014.

[16] Akli Ben Imeddourene, Ahmad Elbahnsi, Marc Guroult, Christophe Oguey, Nicolas Foloppe, and Brigitte Hartmann. Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLOS Comput. Biol.*, 11(12):1–27, 2015.

[17] Daniel Svozil, Jan Kalina, Marek Omelka, and Bohdan Schneider. DNA conformations and their sequence preferences. *Nucleic Acids Res.*, 36(11):3690–3706, 2008.

[18] Marie Zgarbová, Petr Jurečka, Filip Lankaš, Thomas E. Cheatham, Jiří Šponer, and Michal Otyepka. Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Inf. Model.*, 57(2):275–287, 2017.

[19] Alexandra Balaceanu, Marco Pasi, Pablo D. Dans, Adam Hospital, Richard Lavery, and Modesto Orozco. The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, 8(1):21–28, 2017.

[20] Nuri A. Temiz, Duncan E. Donohue, Albino Bacolla, Brian T. Luke, and Jack R. Collins. The Role of Methylation in the Intrinsic Dynamics of B- and Z-DNA. *PLOS ONE*, 7(4):1–9, 2012.

[21] Alfredo Peguero-Tejada and Arjan van der Vaart. Biasing Simulations of DNA Base Pair Parameters with Application to Propellor Twisting in AT/AT, AA/TT, and AC/GT Steps and Their Uracil Analogs. *J. Chem. Inf. Model.*, 57(1):85–92, 2017.

[22] Christine Rauch, Michael Trieb, Bernd Wellenzohn, Markus Loferer, Andreas Voegele, Fajar R. Wibowo, and Klaus R. Liedl. C5-Methylation of Cytosine in B-DNA Thermodynamically and Kinetically Stabilizes BI. *J. Am. Chem. Soc.*, 125(49):14990–14991, 2003.

[23] Sebastian Wärmländer, Judit E. Šponer, Jiří Šponer, and Mikael Leijon. The Influence of the Thymine C5 Methyl Group on Spontaneous Base Pair Breathing in DNA. *J. Biol. Chem.*, 277(32):28491–28497, 2002.

[24] Xiaojing Teng and Wonmuk Hwang. Effect of Methylation on Local Mechanics and Hydration Structure of DNA. *Biophys. J.*, 114(8):1791–1803, 2018.

[25] Fajar R. Wibowo, Michael Trieb, Christine Rauch, Bernd Wellenzohn, and Klaus R. Liedl. The N6-Methyl Group of Adenine Further Increases the BI Stability of DNA Compared to C5-Methyl Groups. *J. Phys. Chem. B*, 109(1):557–564, 2005.

[26] Christopher A. Hunter. Sequence-dependent DNA Structure: The Role of Base Stacking Interactions. *J. Mol. Biol.*, 230(3):1025 – 1054, 1993.

[27] Trent M. Parker, Edward G. Hohenstein, Robert M. Parrish, Nicholas V. Hud, and C. David Sherrill. Quantum-Mechanical Analysis of the Energetic Contributions to π Stacking in Nucleic Acids versus Rise, Twist, and Slide. *J. Am. Chem. Soc.*, 135(4):1306–1316, 2013.

[28] Daniel Svozil, Pavel Hobza, and Jiří Šponer. Comparison of Intrinsic Stacking Energies of Ten Unique Dinucleotide Steps in A-RNA and B-DNA Duplexes. Can We Determine Correct Order of Stability by Quantum-Chemical Calculations? *J. Phys. Chem. B*, 114(2):1191–1203, 2010.

[29] Yoji Umezawa and Motohiro Nishio. Thymine-methyl/π interaction implicated in the sequence-dependent deformability of DNA. *Nucleic Acids Res.*, 30(10):2183–2192, 2002.

[30] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[31] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[32] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep Lluis Gelpí, Carlos González, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. PARMBSC1: A REFINED FORCE-FIELD FOR DNA SIMULATIONS. *Nature methods*, 13(1):55–58, 2015.

[33] Filip Lankaš, Jiří Šponer, Jörg Langowski, and Thomas E. Cheatham. DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.*, 85(5):2872–2883, 2003.

[34] Joe W. Keepers, Peter A. Kollman, Paul K. Weiner, and Thomas L. James. Molecular mechanical studies of DNA flexibility: Coupled backbone torsion angles and base-pair openings. *Proc. Natl. Acad. Sci. USA*, 79(18):5537–5541, 1982.

[35] Tamara M. Okonogi, Stephen C. Alley, Eric A. Harwood, Paul B. Hopkins, and Bruce H. Robinson. Phosphate backbone neutralization increases duplex DNA flexibility: A model for protein binding. *Proc. Natl. Acad. Sci. USA*, 99(7):4156–4160, 2002.

[36] Wilma K. Olson, Andrey A. Gorin, Xiang-Jun Lu, Lynette M. Hock, and Victor B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, 95(19):11163–11168, 1998.

[37] Tomáš Geršata, Mahmut Kara, Martin Zacharias, and Filip Lankaš. Effect of 8-Oxoguanine on DNA Structure and Deformability. *J. Phys. Chem. B.*, 117(39):11617–11622, 2013.

[38] Tomáš Geršata, Nad'a Špačkova, Petr Jurečka, Marie Zgarbová, Jiří Šponer, and Filip Lankaš. Mechanical Properties of Symmetric and Asymmetric DNA A-tracts: Implications for Looping and Nucleosome Posi tioning. *Nucleic Acids Res.*, 42(11):7383–7394, 2014.

[39] Korbinian Liebl, Tomáš Geršata, Filip Lankaš, Jan Lipfert, and Martin Zacharias. Explaining the Striking Difference in Twist-Stretch Coupling between DNA and RNA: A Comparative Molecular Dynamics Analysis. *Nucleic Acids Res.*, 43(21):10143–10156, 2015.

[40] Alberto Pérez, ChiaraLara Castellazzi, Federica Battistini, Kathryn Collinet, Oscar Flores, Ozgen Deniz, MariaLuz Ruiz, David Torrents, Ramon Eritja, Montserrat Soler-López, and Modesto Orozco. Impact of Methylation on the Physical Properties of DNA. *Biophys. J.*, 102(9):2140–2148, 2012.

[41] Thuy TM Ngo, Jejoong Yoo, Qing Dai, Qiucen Zhang, Chuan He, Aleksei Aksimentiev, and Taekjip Ha. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.*, 7(1):1–9, 2016.

[42] T. Zaichuk and J. F. Marko. Single-Molecule Micromanipulation Studies of Methylated DNA. *bioRxiv*, 2020.

[43] Sari Pennings, James Allan, and Colin S. Davey. DNA methylation, nucleosome formation and positioning. *Brief. Funct. Genomics*, 3(4):351–361, 2005.

[44] Peter A. Jones and Daiya Takai. The Role of DNA Methylation in Mammalian Epigenetics. *Science*, 293(5532):1068–1070, 2001.

[45] Peter W. Laird and Rudolf Jaenisch. The Role Of DNA Methylation In Cancer Genetics And Epigenetics. *Annu. Rev. Genet.*, 30(1):441–464, 1996.

[46] Philip M D Severin, Xueqing Zou, Hermann E Gaub, and Klaus Schulten. Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Res.*, 39(20):8740–8751, 2011.

6 How Global DNA Unwinding Causes Non-uniform Stress Distribution and Melting of DNA [1]

6.1 Introduction

DNA unwinding is an important process which is involved in protein-binding, gene-expression and melting of double-stranded DNA [1–9]. Intriguingly, theoretical studies by Benham suggest that regulatory regions are intrinsically destabilized due to super-helical stress in bacteria [10, 11]. Experimentally, the effect of torsional stress on DNA has been studied in single-molecule torque measurements, whereby not only the twist flexibility but also twist-stretch coupling and DNA denaturation have been characterized [12–15]. In chapter 4, we have elucidated DNA's elastic response to unwinding, sequence-dependent structural changes, the role of global restraints and local denaturation. Furthermore, we have also pointed out that this mechanism is highly reversible. However, this study was limited to relatively short DNA duplexes (15 base-pairs) [16]. This leaves many questions open: How do distant sites interact in larger DNA sequences upon unwinding, how is this altered through local melting, and how does local melting depend on the global sequence? In this study, we have performed explicit-solvent all-atom MD simulations on two 50 base-pair long DNA duplexes, which both contain a transcription start like sequence (TATA-box). Unwinding of the duplexes has been implemented using a torque restraining potential. From these simulations, we infer that global unwinding induces coupling between distant sites of the DNA, a phenomenon which is absent in the relaxed case. By employing a harmonic stiffness model [16–22] we find that this also causes a non-uniform stress distribution along the DNA: Cytosine/Guanine-rich segments absorb a higher amount of elastic energy due to enhanced transitions from the BII to the BI state. The TATA-box, instead, remains rather relaxed. This character changes abruptly upon phase transition, where we observe local melting within the TATA-box for both sequences. Here, the global deformation energy is absorbed by the denaturation bubble in the TATA-box, whereas other sequences relax back to near B-form. From our MD simulations, we have also calculated PMFs for the unwinding process. We have related the free energy curves to an Ising model, which we have established from experimental parameters [23]. The direct comparison with the

[1] This chapter has been previously published in similar form in Plos One, 15(5):e0232976, 2020.

DNA-sequence	Label
$5' - cgcgcatgaactgcag\mathbf{ttatat}ggacctcgatgcggcgtacagtacgcgc - 3'$	AT
$5' - cgcgcgcgcgcggccg\mathbf{ttatat}gggccgcgcggcggcgcgccgcgcgcgc - 3'$	GC

Table 6.1: Simulated sequences. TATA-boxes are highlighted in bold.

MD simulations shows a remarkably good agreement for the torsional stress required to initiate the melting transition but also for the position of the melted segment. Moreover, the Ising model gives excellent agreement with the PMFs in the harmonic regime, whereas a slight overstabilization of the DNA force field (bsc1) is indicated. Overall, this study gives new insights into the molecular mechanism of torsional stress induced DNA melting and reveals a high degree of reliability of MD simulations to also capture large-scale transitions in DNA realistically.

6.2 Materials and Methods

6.2.1 Simulation Setup

Two 50 base-pair long dsDNA sequences varying in their C/G content were studied, see Table 6.1.
The sequences were processed with the xleap-tool of the Amber16 package [24]. Thereby, each system was neutralized by sodium ions and solvated with the TIP3P water model [25] using a minimum distance of 12 Å between DNA and rectangular box boundary. The topology of the DNA molecules was described through the bsc1 force field [26]. The DNA structures were initially aligned along the z-axis of the box, and afterwards energy-minimized in 5000 steps (steepest-descent) with the sander module of the Amber16 package. All following simulations were performed with the pmemd.cuda module. We equilibrated the systems to 300 K in three steps while including positional restraints on DNA's non-hydrogen atoms. In five subsequent simulations (T=300 K, p=1 bar), the restraints were gradually removed from 25 $\frac{\text{kcal}}{\text{mol}\text{Å}^2}$ to 0.5 $\frac{\text{kcal}}{\text{mol}\text{Å}^2}$. Resulting structures served as the starting structures for production runs of at least 500 ns, where we applied cylindrical restraints (introduced in chapter 4) with a small force constant of 0.1 $\frac{\text{kcal}}{\text{mol}\text{Å}^2}$ on the two terminal base-pairs of both ends. From these simulations, we inferred the stiffness matrices for the two DNA sequences (see below). The output structures served as input for the Umbrella Sampling (US) simulations. In the US simulations, we used a dihedral angle as reaction coordinate, $\xi = \angle(: 4@C1' : 97@C1' : 53@C1' : 47@C1')$. This dihedral angle represents a rotation of the 47th base-pair with respect to the 4th base-pair. The reaction coordinate was biased by a harmonic potential $V = k \cdot (\xi - \xi_0)^2$, with $k = 0.0122\,\text{kcal}/(\text{mol} \cdot \text{deg}^2) = 40.0\,\text{kcal}/(\text{mol} \cdot \text{rad}^2)$. The reference value ξ_0 was changed in steps of 10° per US interval (window). Simulations of every window were performed for at least 100 ns and the first 50 ns have been skipped in the analysis.

6.2.2 Calculation of Stiffness and Covariance Matrices

Stiffness matrices for both sequences were obtained by inversion of the covariance matrix: $K = k_B T \cdot C^{-1}$, whereby we used the twist, roll and tilt angles of all 43 central base-pair steps. These parameters were calculated with the Curves+ program [27]. Furthermore, we calculated the twist persistence length for both sequences by $P = \frac{Contour \cdot k_{tw}}{k_B T}$, where k_{tw} denotes the twist stiffness as the inverse of the twist variance between 4th and 47th base-pair, and Contour the corresponding contour length measured by the sum of the rise-parameters.

6.2.3 Computation of Geometric Parameters through Rigid-Body Transformation

Besides parameterization with Curves+, we have also developed a new protocol to calculate base-pair and base-pair stacking parameters. The major differences to other protocols (such as Curves+) are:

- All rotational parameters are Euler angles in our method (including also opening and buckling).
- We can easily obtain a reference-axis system for every base and base-pair. This enables us to measure bending as the angle of the axis-vectors, which are orthogonal to the respective base-pairs.
- No specifications are required, except topology- and trajectory files as input.

The protocol works as follows: First, we span a local reference system on every base. The local **x**-vector is the glycosidic bond. We then define a preliminary vector y_p as the vector-difference between the two base-carbon atoms next to the glycosidic nitrogen (e.g. C2, C6 in case of thymine). In the Crick-strand, we define this vector in opposite direction. The local **z**-vector is obtained by $\mathbf{z} = \mathbf{x} \times y_p$. We then calculate the actual **y**-vector through $\mathbf{y} = \mathbf{x} \times \mathbf{z}$. We place this axis-system (denoted $\mathbf{A}$) on the corresponding C1' atom, which serves as anchor-point. Second, we calculate six intra base-pair parameters. The three translational parameters are given by the vectors connecting the anchor-points of paired bases. Afterwards, we calculate the rotation matrix $\mathbf{R}$, which transforms the Watson-axis system into the Crick-axis system: $\mathbf{R} = \mathbf{A}_{Crick}\mathbf{A}^T_{Watson}$. The three rotational intra base-pair parameters are obtained from this matrix as Euler-angles. Third, we calculate the six base-pair step parameters. For this, we first determine a mid-axis system for every base-pair by performing half-translation and half-rotation between the axis-systems of the paired bases. Analog to the calculation of the intra base-pair parameters, the base-pair step parameters are computed by means of rigid-body transformations between successive mid-axis systems.
Moreover, bending angles are calculated as the angle between the **z**-vectors of the mid-axis systems of the chosen base-pairs.

6.2.4 Specification of Base-Pair Fraying

In order to estimate the fraying/breathing tendency of adenine vs thymine base-pairs in the melting region, we computed the center of the two sugar rings of a base-pair, which yields a center point. The distance B of the center point to a reference point in the base (N1-atom for adenine, N3-atom for thymine) was then compared to the distance S between the sugar and the center point. If distance S < distance B, then the base is counted as frayed, otherwise not. The rationale behind this procedure is that in the frayed state, the base rotates away from the center point of the base-pair and hence the sugar is then closer to the center of the base-pair.

6.3 Results and Discussion

6.3.1 DNA as a Set of Harmonic Oscillators

Close to the equilibrium structure, dsDNA's energy landscape can be well described as a harmonic system in terms of intra and inter base-pair parameters [17–20]. For an unrestrained system, one expects on average equipartitioning of the deformation energy E. On a simple mechanical system we first demonstrate how equipartitioning of deformation energy changes upon introduction of a global restraint (such as an external restraint on the twist of DNA). We consider two harmonic oscillators, for simplicity we neglect coupling terms as it does not affect the general mechanism outlined here:

$$E = k_1 \cdot (q_1 - q_1^0)^2 + k_2 \cdot (q_2 - q_2^0)^2, \tag{6.1}$$

whereby the individual deformation energies are given by $E_1 = k_1 \cdot (q_1 - q_1^0)^2$ and $E_2 = k_2 \cdot (q_2 - q_2^0)^2$. The expected energy is determined by the virial theorem:

$$\left\langle q_i \frac{\partial E}{\partial q_j} \right\rangle = \delta_{ij} k_B T. \tag{6.2}$$

When exposed to a thermal heat bath, both oscillators absorb on average the same amount of deformation energy, $\left\langle q_i \frac{\partial E}{\partial q_j} \right\rangle = \langle 2k_1 q_1 \cdot (q_1 - q_1^0) \rangle = \langle 2k_1 (q_1 - q_1^0)^2 \rangle = \langle 2E_1 \rangle = \langle 2E_2 \rangle = k_B T$.

However, if we impose a global restraint on the system, $E^{res} = k_R \cdot (L - L_0)^2$, all parts of the system are coupled through the geometric condition $q_1 + q_2 = L$. As a consequence of the virial theorem, thermal energy is then no longer equipartitioned between the individual oscillators but between coupled modes:

$$k_B T = \left\langle 2k_i q_i \cdot (q_i - q_i^0)^2 + 2k_R q_i \cdot (q_i + q_j - L_0) \right\rangle. \tag{6.3}$$

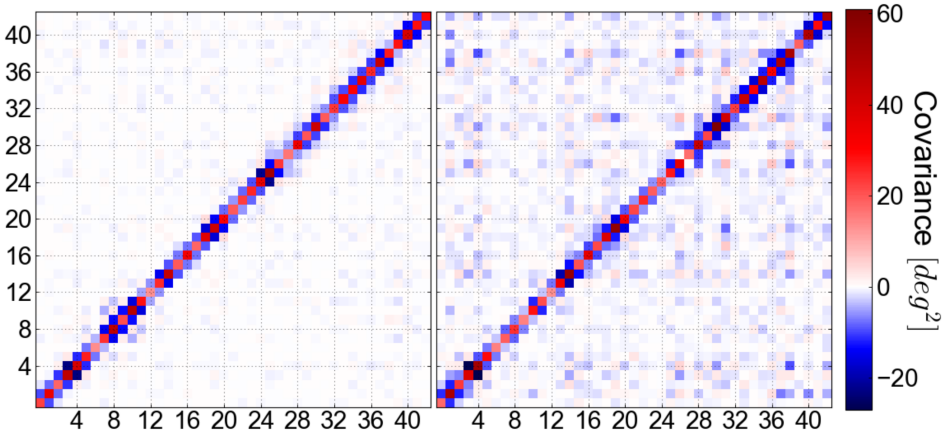


Figure 6.1: Covariance matrix of the twist fluctuations obtained for the central 43 base-pair steps during MD-simulations (AT-sequence). In the unrestrained case (left panel), relevant coupling occurs only between nearest-neighbor steps. Upon global unwinding (right panel, $\sigma = 0.067$) also distant sites become coupled to a significant extent.

Thus, equipartition of deformation energy, $\langle E_1 \rangle = \langle E_2 \rangle$, is no longer obligatory, global restraints rather induce correlations between distant sites. Intriguingly, this is also revealed by our MD simulations of DNA subjected to global unwinding. In the unrestrained case, the changes in twist at a given base-pair step exhibit relevant correlation only with the nearest-neighbors, whilst non-diagonal elements in the covariance matrix are close to zero. However, significant non-nearest-neighbor correlations arise upon addition of the global restraints (Fig 6.1). Qualitatively, this can be understood by changes at one side being able to neutralize opposite changes at a distant side so that the entire system is set in line with the global restraints. Indeed, such long-range correlations have been reported for minicircles in a MD study by Sutthibutpong and coworkers [28].

6.3.2 Local Bimodality in the Twist Distribution of dsDNA

As discussed in chapter 5, the ϵ and ζ dihedral angles give rise to conformational polymorphism in DNA by adopting either the canonical BI or BII configuration [17, 18, 22, 29–31]. These backbone states affect the base-pair stacking geometry and can trigger deviations from the harmonic behavior (Fig 6.2). In general, BII states are associated with a larger equilibrium twist than BI states [22, 30]. Thus, upon unwinding, torsional stress may partially be relaxed by transitions from BII to BI states. Since the population of the BI/BII states is sequence-dependent, Reymer et al have proposed that specific sequences act as 'twist capacitors'. In their study, the authors demonstrate a sequence-dependence in relative changes in twist, but have not related this to the local stiffness [31]. It has therefore remained unclear how much of the deformation energy is absorbed due to backbone polymorphism.

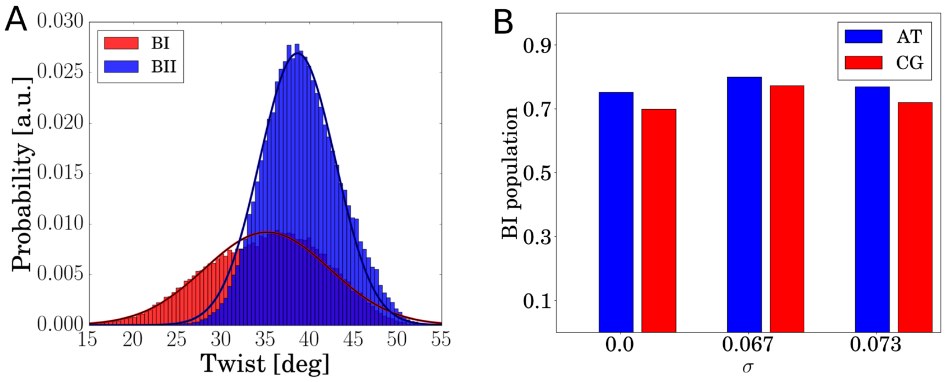


Figure 6.2: (A) Twist distribution sampled for a CpG step during unrestrained MD simulation. The probability distribution is bimodal due to the BI/BII-backbone polymorphism. BI states generally have a lower twist and a higher variance than BII states. (B) Changes in the overall BI population at different levels of global unwinding (for AT and CG steps). At $\sigma = 0.067$ an increase in the BI population is observed which relaxes for higher σ as a consequence of TATA-box melting and relaxation to standard B-DNA of all other segments.

6.3.3 Structural Changes in dsDNA upon Global Unwinding

We have performed MD simulations including a torque on the dsDNA termini to continuously unwind two different, 50 base-pair long, dsDNA molecules. Note that our sequences are apparently longer than those of previous MD studies (usually 10-20 base-pairs) [16, 32] and contain an AT-rich segment ('TATA-box'). Otherwise, our first sequence is heterogeneous (labeled 'AT'), whereas the second sequence contains in addition to the TATA-box only G-C base-pairs. Employing the Umbrella Sampling (US) technique with a torsional reaction coordinate, it is feasible to unwind the central 43 base-pair steps down to local melting. The setup is shown in Fig 6.3. In the following, we indicate the degree of unwinding by just the magnitude of supercoiling density σ (positive number). As pointed out previously, a global unwinding restraint induces long-range couplings in the DNA polymer. At a global unwinding of $\sigma = 0.067$, most base-pair steps are significantly undertwisted relative to regular B-DNA (Fig 6.4). However, further unwinding ($\sigma = 0.073$) causes melting of the TATA-box (and no other region) for both sequences. This transition allows the rest of the DNA to relax back to B-DNA, i.e. almost the entire global stress is then absorbed in the TATA-box. This not only implies untwisting but also bending/stochastic kinking of the TATA-box (Table 6.2). In accordance with our observations from MD studies on shorter sequences (chapter 4), we characterize global unwinding by three phases. Plotting changes in twist of just the TATA-box segment vs global unwinding (Fig 6.5) reveals small gradual unwinding of the TATA-box at modest global stress (regime I). Melting of the TATA-box occurs then to absorb large proportions of global unwinding (regime II), and further unwinding entails steeper changes in twist compared to regime I. In regime III, global unwinding is hence accommodated in the denaturation bubble of the TATA-box (Fig 6.4). This also

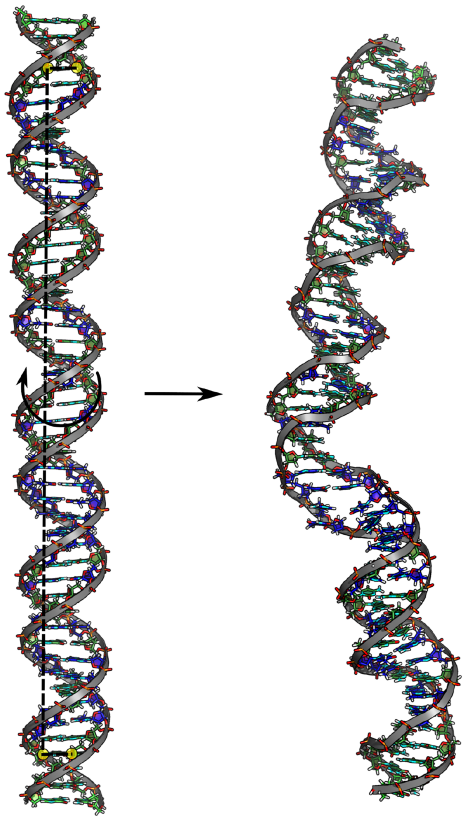


Figure 6.3: Setup of the simulation systems (as stick models with the backbone shown as cartoon). Yellow circles highlight the position of the C1' atoms of the 4th and 47th base-pair, and constitute the torsional reaction coordinate (illustrated as dashed line, curved arrow indicates unwinding). Stepwise unwinding eventually causes melting of the TATA-box segment. C/G bases are shown in green, A/T bases in blue.

outlines the scenario of forming left-handed DNA structures in the melted region for longer sequences. Indeed, experimental studies have reported that left-handed segments can occur in strongly supercoiled DNA [12, 33–35]. Consistent with previous studies on base-pair formation and melting [36–38], we have found an asymmetric fraying/breathing tendency of adenine vs thymine bases in the melted region. Overall, we obtain on average an increased breathing of thymine bases in the denatured TATA-box by a factor of 1.2 (AT-sequence) and 1.3 (GC-sequence), respectively. Details on the computation of the breathing tendency are given in paragraph 6.2.4. Due to long-range couplings induced by the global unwinding stress, structural changes are also representable for TATA-box distant sites: Small global unwinding stress results in untwisting and helical-shortening of those sites, but increasing torsional stress does not induce structural transitions. On the contrary, the twist and other helical parameters relax to equilibrium values at these sites when the TATA-box begins to denature. This mechanism is illustrated by snapshots in Fig 6.6.

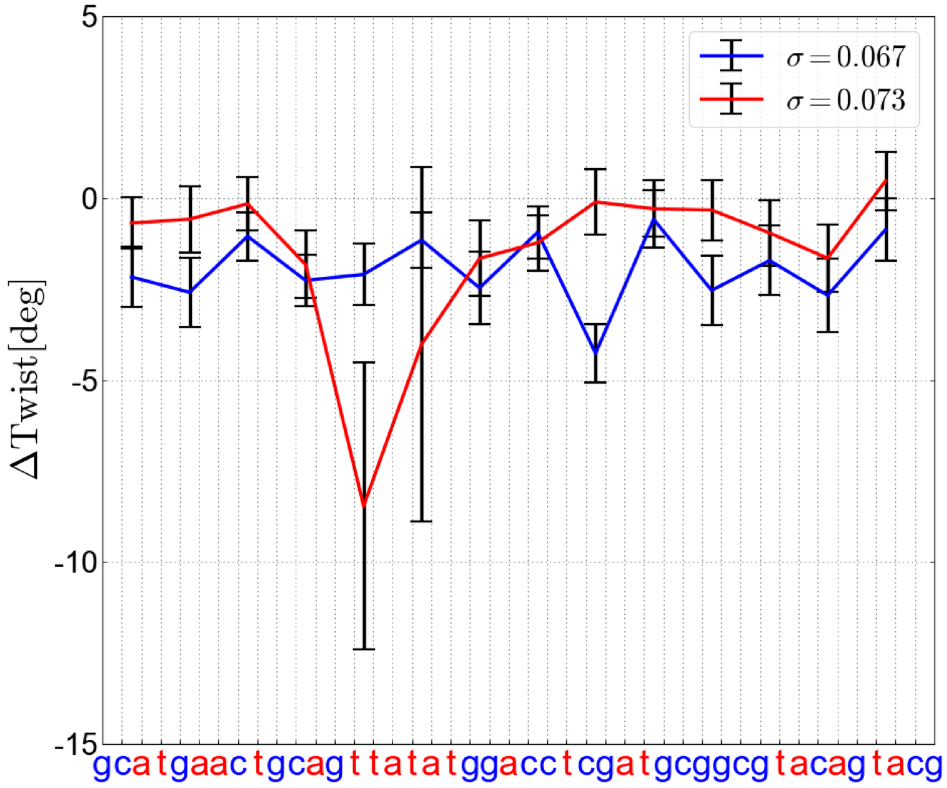


Figure 6.4: Average changes in twist at each base-pair step (relative to unrestrained DNA, AT-sequence) for two different levels of global unwinding: Slightly below TATA-box melting (blue), and for melted TATA-box (red). The average deviations and standard deviations were calculated as averages of three adjacent base-pair steps.

	Twist/bp [deg]	Bending/bp [deg]
$\sigma = 0.0$	34.37 ± 1.21	2.32 ± 1.24
$\sigma = 0.067$	31.88 ± 1.39	3.13 ± 1.46
$\sigma = 0.073$	21.59 ± 2.79	4.45 ± 1.83

Table 6.2: Deformation of the TATA-box under different levels of global unwinding, averaged over both sequences. Parameters have been obtained by rigid-body transformation of the first C-G base-pair prior to the TATA-box to the first C-G base-pair after it. Our parameterization methodology is described in paragraph 6.2.3

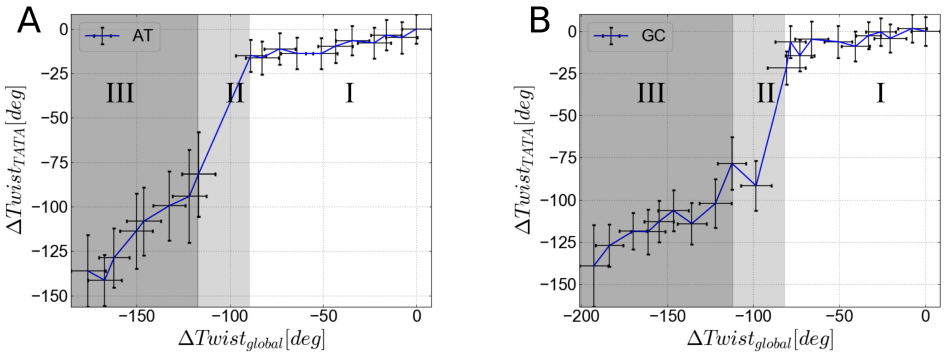


Figure 6.5: Change in total twist of the TATA-box segment upon global unwinding for the AT-sequence (A) and GC-sequence (B). Close to equilibrium (regime I), slight continuous unwinding of the TATA-box is observed, which changes abruptly during the transition phase (regime II). In this phase, the TATA-box begins to denature and thus largely stores global unwinding. Further unwinding causes continuous unwinding of the TATA-box (regime III) with a steeper slope than in regime I. Due to its smaller persistence length, the denaturation bubble absorbs most of the global stress. Global twisting and twisting of the TATA-box segment have been computed based on our protocol presented in paragraph 6.2.3.

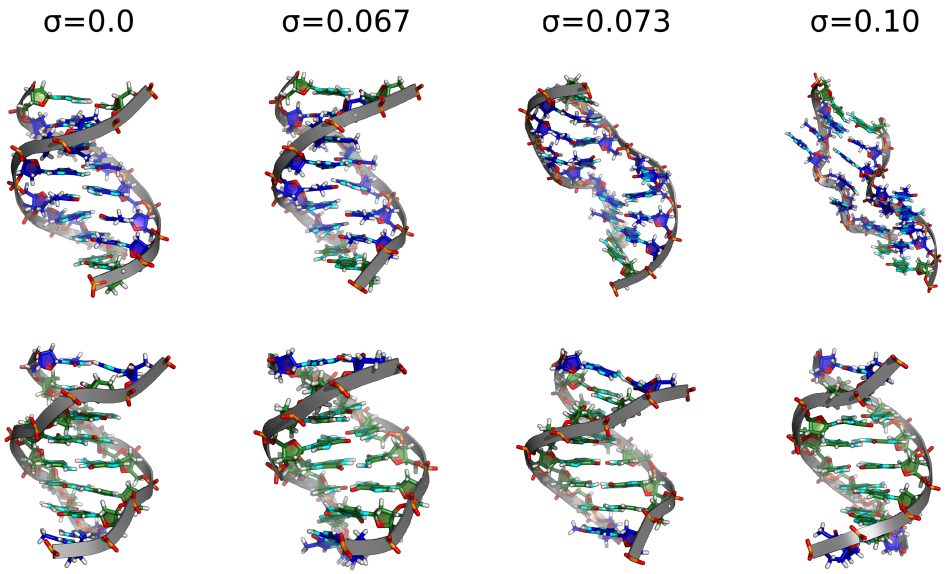


Figure 6.6: Snapshots of two selected 8 base-pair segments of the DNA at different levels of global unwinding. Structures of the TATA-box containing segment are shown in the upper series. Structures of a distant GC-rich segment are illustrated in the lower series (same color-coding as in Fig 6.3.)

6.3.4 Redistribution of Elastic Energy upon Global Unwinding

From our US simulations, we have calculated mean deformation energies for each base-pair step as a function of global unwinding employing the harmonic model [17–20]:

$$E(\Delta w) = \frac{1}{2}\Delta w^T K \Delta w, \tag{6.4}$$

Δw denotes deviations of the internal coordinates form their equilibrium values and K the stiffness matrix, obtained as inverse of the covariance matrix of the internal coordinates: $K = k_B T \cdot C^{-1}$. Note that the stiffness matrix and equilibrium values have been determined from extensive unrestrained MD simulations. Since changes in global unwinding may be coupled to bending of the DNA, we have selected twist, roll and tilt of each of the 43 central base-pair steps as internal coordinates. Thereby, we assume that other internal parameters relax according to the corresponding covariations. However, our description fully specifies all twist and bending modes, hence K becomes a 129x129 matrix. The elastic energy for each base-pair step includes diagonal and half of the secondary diagonal terms. In the absence of external stress, we obtain $\sim \frac{k_B T}{2}$ per degree of freedom, which is expected due to the equipartition theorem. In the presence of torsional stress, the local elastic behavior of the DNA changes completely: Already at modest global unwinding, the deformation energy profiles along the DNA sequence show a non-uniform distribution (upper panels in Fig 6.7). Further increase of global torsional stress results in strong fluctuations of absorbed deformation energy. On average, the stress levels of A/T segments are lower compared to C/G segments and the TATA-boxes are rather relaxed (middle panels in Fig 6.7). Segments storing higher amounts of elastic energy exhibit significant transitions in the backbone structure. Since C/G segments have a higher propensity for adopting a BII state in unrestrained DNA [22, 39–41], these segments also have a higher capacity towards transitions from BII to BI states, which relax the unwinding stress. Especially for strong unwinding (prior to melting) we observe a significantly increased BI population mostly for C/G segments. This finding indicates a protection mechanism of the DNA, as in this way unwinding of the DNA does not lead to higher flipping rates of A/T base-pairs. In chapter 5, we have pointed out that clashing of thymine's methyl group with the 5' sugar promotes population of BI states. We now argue that this steric effect has a strong impact on how the DNA responds to global stress. Unwinding to $\sigma \sim 0.07$ induces a phase transition characterized by local denaturation. In this phase, most of the deformation energy is absorbed in the denaturation bubble, whereas all other segments of the DNA relax to equilibrium conformation (bottom panels in Fig 6.7). This is also in line with the adaption of the twist parameters, discussed in paragraph 6.3.3. In the melted region, the computed elastic energies of the TATA-box only have a qualitative meaning, as the harmonic model is not valid for denatured conformations. The impact of global unwinding on the sequence-dependent deformation energy distribution is presented schematically in Fig 6.8.

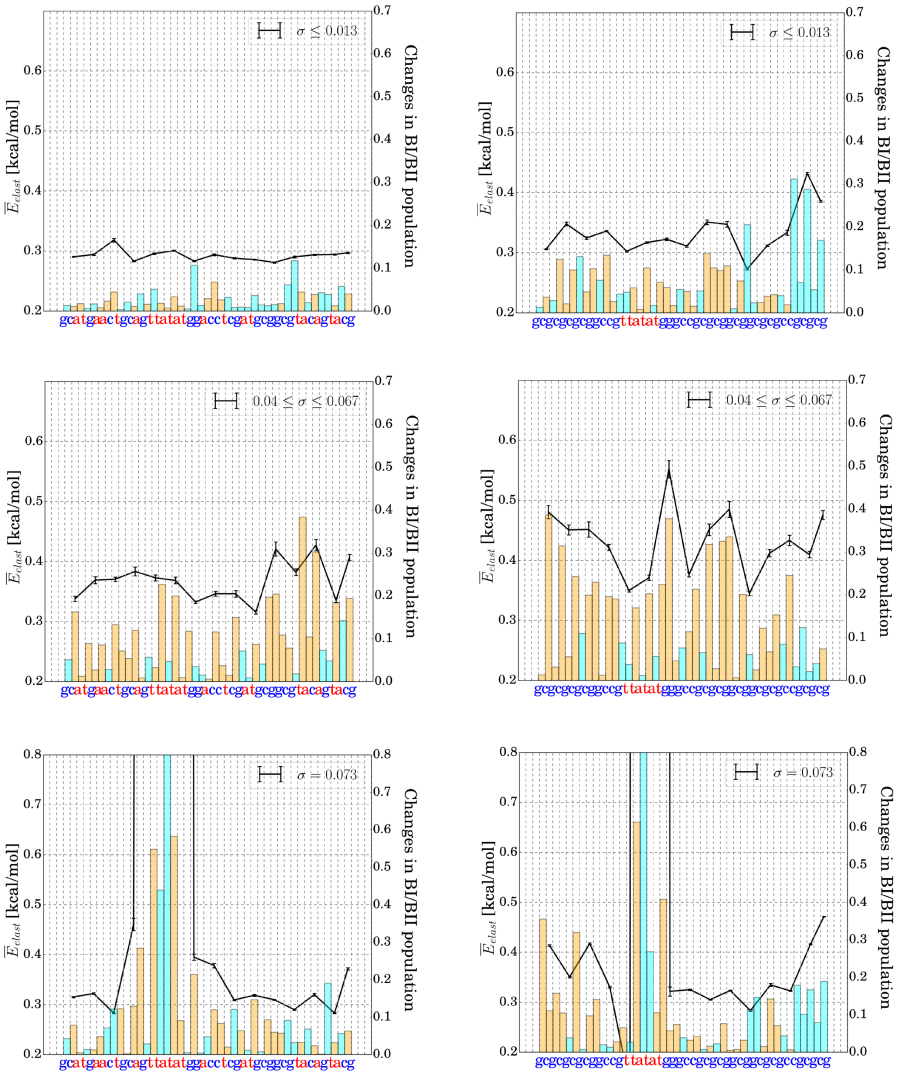


Figure 6.7: Average elastic energy per degree of freedom and changes in the BI/BII population with respect to unrestrained simulations for each of the central 43 bp steps of the AT sequence (left column) and the GC sequence (right column). Vertical dashed lines in the panels correspond to a base-pair (sequence indicated on the x-axis). The black lines indicate the average deformation energy at each base-pair step. Cyan bars indicate an increase in BII, orange bars an increase in BI states relative to the unrestrained simulation. The profiles have been generated for different levels of global DNA unwinding. Upper panels show profiles of Umbrella windows with low unwinding. Middle panels illustrate moderate unwinding conditions, where the DNA structure is still intact. Bottom panels show simulations at strong unwinding, which results in TATA-box melting for both sequences. In order to calculate E_{elast}, we determined deformation energies for every frame and then calculated averages (over three adjacent base-pair steps). Note that a harmonic description for the deformation energy is insufficient for denatured structures, corresponding values therefore only have qualitative meaning. Error bars have been calculated as standard errors by splitting the simulation into bins of 25 ns.

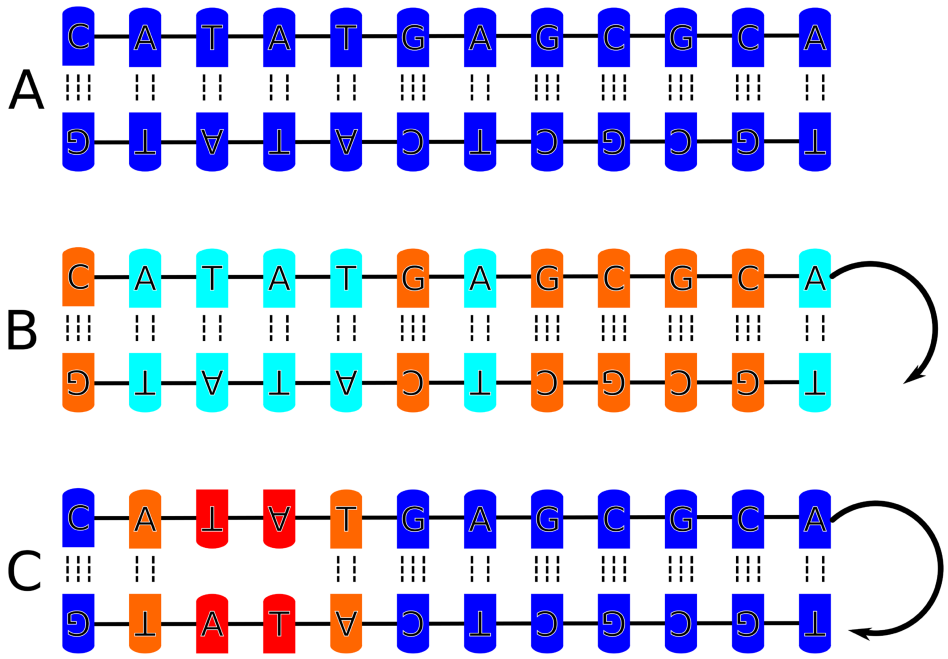


Figure 6.8: Scheme for sequence-dependent response to global unwinding. In the absence of global stress, all sequences undergo equal levels of deformation ($\sim \frac{k_B T}{2}$, A, highlighted in blue). Global unwinding then causes G/C sequences to absorb most of the stress (orange color in B), whereas A/T rich segments are less stressed (light blue color in B). At strong levels of unwinding, global stress is absorbed through melting of the TATA-Box (red in C), whereby distant sequences are relaxed again (blue in C).

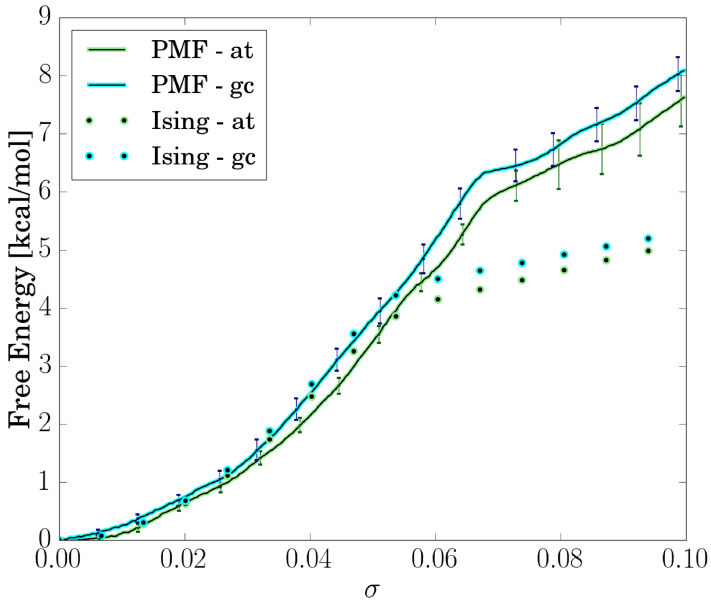


Figure 6.9: Free energy profiles for global unwinding of the DNA molecules obtained from US simulations (PMFs, solid lines) and calculated by Ising models (dotted). Error bars of the PMFs have been calculated by thermodynamic integration and block-averaging [43].

6.3.5 Calculation of Free Energy Changes

We have calculated relative free energy profiles (PMFs) as a function of global unwinding σ with the Weighted Histogram Analysis Method (WHAM) [42], see Fig 6.9. In agreement with chapter 4, we find a quadratic form of the PMFs up to relatively large σ [16]. Upon local melting ($\sigma \sim 0.07$), the free energy curves change drastically to a flat shape with a small slope. This behavior reflects the description of structural changes during unwinding (paragraph 6.3.3): After local melting, global unwinding is absorbed in the denaturation bubble (TATA-box), which has a lower stiffness. Thus, costs in the free energy for further unwinding are significantly reduced. Consistently, TATA-box melting occurs in both sequences at similar values of σ and requires $\sim 6\,$kcal/mol.

6.3.6 Modeling Melting Behavior upon Unwinding by an Ising Model

In a next step, we have compared the thermodynamic behavior extracted from our all-atom US simulations with predictions from an Ising model. In an Ising model, the stability of dsDNA is determined by nearest-neighbor interactions (base-pairing and stacking interactions), and mechanical destabilization is described by the global twist stiffness (persistence length) and a smaller stiffness for denatured segments [6–8]. Note that the Ising model allows us to calculate the partition function and associated ther-

	ag	ac	at	aa	ta
E_{stack}	-1.44	-2.19	-1.72	-1.49	-0.57
	tc	tg	gg	gc	cg
E_{stack}	-1.81	-0.93	-1.82	-2.55	-1.29

Table 6.3: Stacking energies of the ten different base-pair steps used in the Ising model. For base-pairing, we used $E_{bp} = 0.64$ and $E_{bp} = 0.12$ for AT and GC pairs, respectively. All values have been adopted from [44] and are in units of kcal/mol.

modynamic quantities as a function of global unwinding σ.
Our Ising model is based on experimentally determined base-stacking and base-pairing parameters [44]. Furthermore, we use an empirical bubble initiation parameter $\epsilon = 4.1\,\mathrm{kcal/mol}$, which has been shown by Krueger et al [23] to reconcile the base-stacking and pairing energies with measured probabilities for base-pair opening. We set the stiffness constant for denaturation bubbles to $C = 0.79\,\mathrm{cal/(mol \cdot deg^2)}$, and write the elastic energy as $E = C \cdot \frac{\Delta\tau - 34.5\cdot(n+1))^2}{2\cdot n}$. This approximation for the elasticity of denaturation bubbles has been introduced by Benham [6]. Importantly, the effective stiffness depends on the number of molten base-pairs n. For regular dsDNA, we use a twist persistence length $P = \frac{Contour\cdot k_{tw}}{k_B T}$ extracted from our unrestrained MD simulations by measuring the mean contour-length and the stiffness k_{tw} as the inverse of the twist variance. At this point, we have used parameters extracted from our simulations in order to account for the sequence dependence. Moreover, the extracted twist persistence lengths (110.8 nm for the AT and 119.9 for the GC-sequence) are close to experimental estimates (around 100 nm) [13]. In the Ising model, we discretize base-pairing by the eigenstates $\mid i >, \mid j > := \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ for an intact base-pair and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ for a denatured base-pair. The resulting Hamiltonian is then given in 6.5. The elements of the interaction matrix H_{ij} are related to the state of the base-pairs $< i \mid$ and $\mid j >$, e.g. the entry on the upper left of H_{ij} describes the state where both base-pairs are intact. This state is therefore not penalized by any potential.

$$H = \begin{cases} \Sigma_{i=0,j=i+1}^{N} < i| \begin{pmatrix} 0 & \epsilon - E_{bp}^{j} - E_{stack}^{ij} \\ -E_{bp}^{i} - E_{stack}^{ij} & -E_{bp}^{j} - E_{bp}^{i} - E_{stack}^{ij} \end{pmatrix} \mid j > + E_{bubble} & , \quad n > 0 \\ \frac{k_B T P}{2L} \Delta\tau^2 & , \quad n = 0 \end{cases} \tag{6.5}$$

The partition sum can then be obtained through a transfer-matrix calculation that sums over all possible states and the corresponding Boltzmann weight:

$$Z = \exp\left[-\beta E_{bubble}\right] \cdot < 0 \mid \Pi_{i=0,j=i+1}^{N} \exp\left[-\beta H_{ij}\right] \mid N{+}1 > -\exp\left[-\beta E_{bubble}\right] + \exp\left[-\beta H_{n=0}\right]. \tag{6.6}$$

We set $< 0 \mid \quad = \quad \mid N+1 > = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, as these two overhang base-pairs are not affected by torsional stress. Note that this treatment reflects the setup of our MD simulations. However, we must ensure that the number of denatured eigenstates equals n (in the elastic bubble energy), hence we only keep according terms in our implementation of the Ising model. For optimization purposes, we neglect macrostates containing more than one denaturation bubble. This approximation is justified as our DNA sequences are relatively short compared to the size of plasmids and such states are penalized by another 4.1 kcal/mol. Importantly, the partition sum is a function of global unwinding $\Delta\tau$. Therefore, we can also calculate relative free energies as a function of σ by $F = -k_B T \cdot \ln(Z)$. These free energy curves can be directly compared to the PMFs obtained from our US simulations. As shown in Fig 6.9, the agreement between both free energy curves is remarkably good: Close to equilibrium, the Ising model calculations coincide very well the MD-based PMFs. Interestingly, the Ising model also shows a flattening in the free energy profile upon transition to the melted phase, which is a consequence of the low elastic constant of the denaturation bubble and represents an enthalpic character of DNA melting. Compared to our all-atom simulations, the Ising model predicts an earlier onset of the phase transition and hence lower stability of the intact phase (by ~ 2 kcal/mol). We identify two likely reasons therefore: First, in our simulations we only find melting of the TATA-box region, as bubble closing and reopening at different positions occurs on a too large timescale to be sampled in current all-atom MD simulations. Thus, our simulations entropically underestimate the denaturation phase. By means of the Ising model, we estimate this effect to correspond to only $\sim 0.09/0.40$ kcal/mol (GC/AT). Second, we suppose base-stacking to be overstabilizing in current DNA force fields. This argument has also been made in previous studies [45–47].

Besides, the Ising model allows us to determine local melting probabilities (Fig 6.10). Here, we obtain a dominant contribution from the TATA-box, and see that increased global C/G content promotes earlier melting of this region. This phenomenon has already been found experimentally by Vlijm et al [15]. We explain this finding as a melting competition between different sites of the DNA, as melting of one segment has a relaxation contribution on all other segments and hence decreases their melting probability. Thus, we think of it as an entropic rather than an enthalpic phenomenon, as the persistence lengths of our sequences are also quite similar. Furthermore, we predict the length of the denaturation bubble to $\sim 1-3$ bps for our sequences. This also nicely resembles the observation from our simulations.

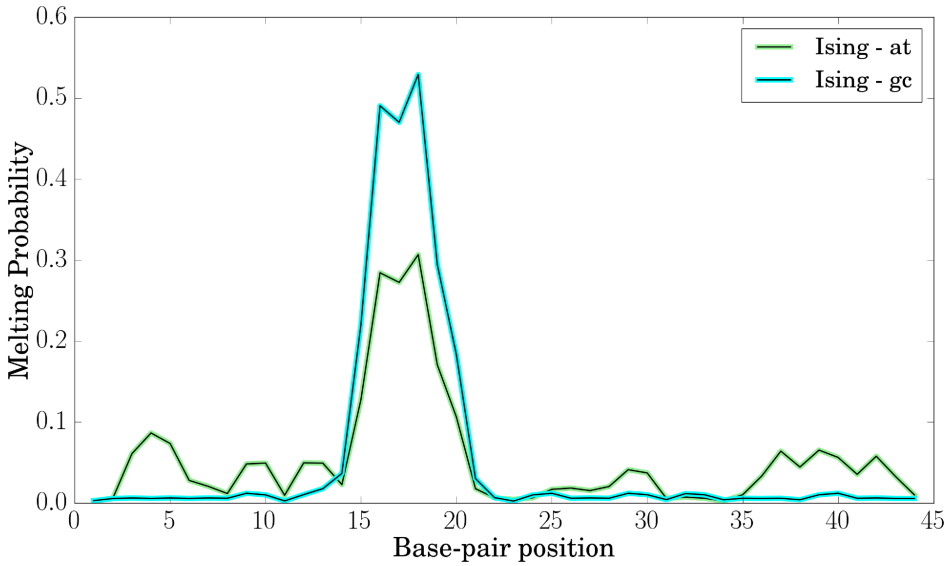


Figure 6.10: Calculated melting probabilities along the sequence using the Ising model at a global supercoiling stress of $\sigma = 0.067$.

6.4 Conclusion

In this study, we have achieved atomistic simulation of localized melting of a promoter-like TATA-box sequence for the first time. Thereby, we have not only calculated the required free energy for unwinding induced DNA melting but also complemented our MD simulations with an Ising model. This physical model is independent from our unwinding simulations and based on empirical parameters. Though the Ising calculations indicate an overstabilization of the MD force field, we overall obtain good agreement between both theories.

In addition, we also give new insights into structural changes and the absorbance of local deformation energies as a function of global unwinding. Whereas elastic energy is evenly distributed in the absence of global restraints, it is absorbed highly non-uniform along the sequence upon global unwinding. Here, we find that C/G sites take up most of the global stress, hence relaxing the TATA-box. The reason therefore is that in equilibrium C/G segments exhibit a higher population of BII states [22, 31, 40, 41]. This increases their capacity to flip into the BI subspace which absorbs global unwinding stress. One can interpret this as a protection mechanism, as it implies that modest unwinding does not promote base flipping in A/T sites. Note that we have given a sterical explanation which underlies this mechanism in chapter 5: In equilibrium, A/T sites have a higher population of BI (and hence less BII states) than C/G sites due to thymine's methyl group clashing with the 5' neighboring sugar [22]. It is therefore possible that overwinding causes the opposite behavior and preferentially stresses A/T sites by switching from BI to BII states.

However, unwinding to $\sigma \sim 0.07$ causes a phase transition including melting in the

TATA-box and requires $\sim 6\,\mathrm{kcal/mol}$ (as obtained from our US-simulations). Simultaneously, all other sites of the DNA relax to equilibrium conformation. Further unwinding then requires only little energy, because the denaturation bubble in the TATA-box absorbs nearly the entire global stress. This principle, that strong global stress localizes the least stable part in the DNA, may also arise in other biological processes: Inhibition of the DNA gyrase enzymes, for instance, decreases the repair fidelity of UV damages in bacteria drastically [2], and Dittmore et al [48] have shown that supercoiling locates mismatches in DNA.

6.5 Bibliography

[1] C J Benham. Torsional stress and local denaturation in supercoiled DNA. *Proc. Natl. Acad. Sci. USA*, 76(8):3870–3874, 1979.

[2] J B Hays and S Boehmer. Antagonists of DNA gyrase inhibit repair and recombination of UV-irradiated phage lambda. *Proc. Natl. Acad. Sci. USA*, 75(9):4125–4129, 1978.

[3] Alan R. Bishop, Anny Usheva, Boian S. Alexandrov, Kim . Rasmussen, Ludmil B. Alexandrov, Sang Wook Yoo, Vladimir Gelev, and Yayoi Fukuyo. DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation. *Nucleic Acids Res.*, 38(6):1790–1795, 2009.

[4] Sung Hyun Kim, Mahipal Ganji, Eugene Kim, Jaco van der Torre, Elio Abbondanzieri, and Cees Dekker. DNA sequence encodes the position of DNA supercoils. *eLife*, 7:e36557, 2018.

[5] D Kowalski, D A Natale, and M J Eddy. Stable DNA unwinding, not "breathing "accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences. *Proc. Natl. Acad. Sci. USA*, 85(24):9464–9468, 1988.

[6] Craig J. Benham. Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.*, 225(3):835–847, 1992.

[7] Daniel Jost and Ralf Everaers. Genome wide application of DNA melting analysis. *J. Phys: Condensed Matter*, 21(3):034108, 2008.

[8] Daniel Jost, Asif Zubair, and Ralf Everaers. Bubble statistics and positioning in superhelically stressed DNA. *Phys. Rev. E*, 84(3):031912, 2011.

[9] Rossitza N. Irobalieva, Jonathan M. Fogg, Daniel J. Catanese Jr, Thana Sutthibutpong, Muyuan Chen, Anna K. Barker, Steven J. Ludtke, Sarah A. Harris, Michael F. Schmid, Wah Chiu, and Lynn Zechiedrich. Structural diversity of supercoiled DNA. *Nat. Communications*, 6(8440):8440, 2015.

[10] Craig J. Benham. Duplex Destabilization in Superhelical DNA is Predicted to Occur at Specific Transcriptional Regulatory Regions. *J. Mol. Biol.*, 255(3):425–434, 1996.

[11] Huiquan Wang, Michiel Noordewier, and Craig J Benham. Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. *Genome Res.*, 14(8):1575–1584, 2004.

[12] Carlos Bustamante, Zev Bryant, and Steven B Smith. Ten years of tension: single-molecule DNA mechanics. *Nature*, 421(6921):423–427, 2003.

[13] Jan Lipfert, Gary M. Skinner, Johannes M. Keegstra, Toivo Hensgens, Tessa Jager, David Dulin, Mariana Kber, Zhongbo Yu, Serge P. Donkers, Fang-Chieh Chou, Rhiju Das, and Nynke H. Dekker. Double-stranded RNA under force and torque:

Similarities to and striking differences from double-stranded DNA. *Proc. Natl. Acad. Sci. USA*, 111(43):15408–15413, 2014.

[14] Domenico Salerno, Alessia Tempestini, I Mai, Doriano Brogioli, Roberto Ziano, Valeria Cassina, and Francesco Mantegazza. Single-molecule study of the DNA denaturation phase transition in the force-torsion space. *Phys. Rev. Lett.*, 109(11):118303, 2012.

[15] Rifka Vlijm, Jaco V D Torre, and Cees Dekker. Counterintuitive DNA Sequence Dependence in Supercoiling-Induced DNA Melting. *PLOS ONE*, 10(10):e0141576, 2015.

[16] Korbinian Liebl and Martin Zacharias. Unwinding Induced Melting of Double-Stranded DNA Studied by Free Energy Simulations. *J. Phys. Chem. B*, 121(49):11019–11030, 2017.

[17] Tomáš Geršata, Alberto Pérez, Modesto Orozco, Alexandre V Morozov, Jiří Šponer, and Filip Lankaš. Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, 9(1):707–721, 2013.

[18] Tomáš Geršata and Filip Lankaš. Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3(4):355–363, 2013.

[19] O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks. A sequence-dependent rigid-base model of DNA. *J Chem Phys*, 138(5):055102, 2013.

[20] F. Lankaš, O. Gonzalez, L. M. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks. On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, 11(45):10565–10588, 2009.

[21] Korbinian Liebl, Tomáš Geršata, Filip Lankaš, Jan Lipfert, and Martin Zacharias. Explaining the Striking Difference in Twist-Stretch Coupling between DNA and RNA: A Comparative Molecular Dynamics Analysis. *Nucleic Acids Res.*, 43(21):10143–10156, 2015.

[22] Korbinian Liebl and Martin Zacharias. How methyl-sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.*, 47(3):1132–1140, 2018.

[23] Andrew Krueger, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Sequence-dependent base pair opening in DNA double helix. *Biophys. J.*, 90(9):3091–3099, 2006.

[24] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[25] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[26] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep Lluis Gelpí, Carlos González, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. PARMBSC1: A REFINED FORCE-FIELD FOR DNA SIMULATIONS. *Nature methods*, 13(1):55–58, 2015.

[27] R Lavery, M Moakher, JH Maddocks, D Petkevičiutė, and K Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37(17):5917–5929, 2009.

[28] Thana Sutthibutpong, Christian Matek, Craig Benham, Gabriel G. Slade, Agnes Noy, Charles Laughton, Jonathan P. K. Doye, Ard A. Louis, and Sarah A. Harris. Long-range correlations in the mechanics of small DNA circles under topological stress revealed by multi-scale simulation. *Nucleic Acids Res.*, 44(19):9121–9130, 2016.

[29] D. Djuranovic and B. Hartmann. DNA fine structure and dynamics in crystals and in solution: The impact of BI/BII backbone conformations. *Biopolymers*, 73(3):356–368, 2004.

[30] Marie Zgarbová, Petr Jurečka, Filip Lankaš, Thomas E. Cheatham, Jiří Šponer, and Michal Otyepka. Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Info. Mod.*, 57(2):275–287, 2017.

[31] Anna Reymer, Krystyna Zakrzewska, and Richard Lavery. Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. *Nucleic Acids Res.*, 46(4):1684–1694, 2017.

[32] Srinivasaraghavan Kannan, Kai Kohlhoff, and Martin Zacharias. B-DNA Under Stress: Over- and Untwisting of DNA during Molecular Dynamics Simulations. *Biophys. J.*, 91(8):2956–2965, 2006.

[33] Florian C Oberstrass, Louis E Fernandes, Paul Lebel, and Zev Bryant. Torque spectroscopy of DNA: base-pair stability, boundary effects, backbending, and breathing dynamics. *Phys. Rev. Lett.*, 110(17):178103, 2013.

[34] Zev Bryant, Michael D. Stone, Jeff Gore, Steven B. Smith, Nicholas R. Cozzarelli, and Carlos Bustamante. Structural transitions and elasticity from torque measurements on DNA. *Nature*, 424(6946):338–341, 2003.

[35] E Di Capua, A Stasiak, T Koller, S Brahms, R Thomae, and F M Pohl. Torsional stress induces left-handed helical stretches in DNA of natural base sequence: circular dichroism and antibody binding. *EMBO J.*, 2(9):1531–1535, 1983.

[36] Francesco Colizzi, Cibran Perez-Gonzalez, Remi Fritzen, Yaakov Levy, Malcolm F White, J Carlos Penedo, and Giovanni Bussi. Asymmetric base-pair opening drives helicase unwinding dynamics. *Proc. Natl. Acad. Sci. USA*, 116(45):22471–22477, 2019.

[37] C Izanloo, GA Parsafar, H Abroshan, and H Akbarzadeh. Denaturation of Drew-Dickerson DNA in a high salt concentration medium: Molecular dynamics simulations. *Journal of computational chemistry*, 32(16):3354–3361, 2011.

[38] Xiaojun Xu, Tao Yu, and Shi-Jie Chen. Understanding the kinetic mechanism of RNA single base pair formation. *Proc Natl. Acad. Sci. USA*, 113(1):116–121, 2016.

[39] Alexandra Balaceanu, Marco Pasi, Pablo D. Dans, Adam Hospital, Richard Lavery, and Modesto Orozco. The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, 8(1):21–28, 2017.

[40] Pablo D. Dans, Linda Danilāne, Ivan Ivani, Tomáš Dršata, Filip Lankaš, Adam Hospital, Jürgen Walther, Ricard Illa Pujagut, Federica Battistini, Josep Lluis Gelpí, Richard Lavery, and Modesto Orozco. Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res.*, 44(9):4052–4066, 2016.

[41] Pablo Daniel Dans, Ignacio Faustino, Federica Battistini, Krystyna Zakrzewska, Richard Lavery, and Modesto Orozco. Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, 42(18):11304–11320, 2014.

[42] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[43] Fangqiang Zhu and Gerhard Hummer. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.*, 33(4):453–465, 2012.

[44] Ekaterina Protozanova, Peter Yakovchuk, and Maxim D. Frank-Kamenetskii. Stacked-Unstacked Equilibrium at the Nick Site of DNA. *J. Mol. Biol.*, 342(3):775–785, 2004.

[45] Florian Häse and Martin Zacharias. Free energy analysis and mechanism of base pair stacking in nicked DNA. *Nucleic Acids Res.*, 44(15):7100–7108, 2016.

[46] Dazhi Tan, Stefano Piana, Robert M Dirks, and David E Shaw. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. USA*, 115(7):E1346–E1355, 2018.

[47] Jejoong Yoo and Aleksei Aksimentiev. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Phys. Chem. Chem. Phys.*, 20(13):8432–8449, 2018.

[48] Andrew Dittmore, Sumitabha Brahmachari, Yasuharu Takagi, John F. Marko, and Keir C. Neuman. Supercoiling DNA Locates Mismatches. *Phys. Rev. Lett.*, 119(14):147801, 2017.

7 Coarse-graining DNA: A Multivariate Ising Model

7.1 Introduction

DNA's sequence encodes not only the sequence of proteins but also its own local deformabilities [1–5]. This relationship plays an important role in protein-DNA binding, processes of outmost biological relevance [4, 6–10]. Chromatin folding, for instance, depends on DNA's sequence-dependent deformability, as nucleosome formation strongly bends the DNA [8, 11–15]. Another interesting example is the binding specificity of papillomavirus proteins. These proteins locally deform DNA upon binding and show binding affinities which strongly depend on DNA's sequence [16–19]. Understanding of these processes, how proteins recognize and bind to specific DNA sequences is very limited so far. As a possible explanation, the concept of indirect readout mechanism, which states that proteins find their target sites by scanning DNA's local deformability, has been proposed [6, 20]. In this view, the energetic cost to deform a contacted DNA segment serves as a selection criterion for binding. However, this concept of indirect readout strategies has remained vague and hypothetical. A major reason therefore is the difficulty in describing DNA's deformability.
It has been well established that the deformability of a dsDNA polymer can be assembled from the deformabilities of tetranucleotide sequences [21–24]. Nevertheless, modeling DNA's deformability is complicated by two factors: First, nearest-neighbor base-pair stacking is correlated [22, 25, 26]. Second, DNA exhibits a high level of bimodality, which is caused by conformational substates in the backbone [26–29]. In general, coarse-grained models have approached DNA's deformability with a simple, additive and harmonic Hamiltonian [10, 14, 23, 24, 30]. However, such a Hamiltonian fails to simultaneously capture both, nearest-neighbor correlation and multimodality. The ThreaDNA- as well as the cgDNA-model account for nearest-neighbor correlation but are unimodal [24, 30]. On the other hand, the coarse-grained method by Walther et al treats multimodality, but their Hamiltonian neglects nearest-neighbor coupling. Furthermore, it also does not explicitly account for intra base-pair parameters [23]. In this study, we propose a novel approach to DNA's deformability: By combining the multivariate harmonic approximation with the Ising model, we account for nearest-neighbor correlation and multimodality. Our model is fully parameterized from $\sim 1\,\mu s$ MD simulations of the ABC-sequences [21], which include all 136 tetranucleotides, and accounts for all base-pairing and base-pair stacking degrees of freedom. We show that our Ising model is more

accurate by $\sim 0.03\,\mathrm{kcal/mol}$ per degree of freedom than the standard model in the evaluation of MD simulations on regular B-DNA oligomers. We have also applied the Ising model to simulations on papillomavirus-protein/DNA complexes. Here, the free energy estimations from the Ising model correlate very well with experimental binding affinities, whereas the harmonic model is too imprecise to resolve sequence-dependent effects accurately. In a next step, we have studied sequence-dependent deformation energies in nucleosome structures. Thereby, we have applied our Ising model to comparatively quantify energies along DNA's sequence for the human nucleosome core particle with its original and with an A-tract replaced sequence. Our Ising model based calculations clearly reveal a strong increase in deformation energies upon substitution by A-tracts. This explains why A-tracts deplete the formation of nucleosomes in vivo, an effect of high importance for chromatin folding [31].
Thus, our multivariate Ising model provides a new and reliable approach to quantify sequence-dependent deformation contributions that arise in protein-DNA binding. Our routine is fast and hence enables evaluation of crystal structures and μs-long MD-simulations within a few seconds.

7.2 Materials and Methods

7.2.1 MD Simulations of the ABC and Test Sequences

For the parameterization of the coarse-grained models (pure harmonic as well as Ising model), we simulated all 39 sequences of the ABC consortium [21]. Together, these sequences (each consists of 18 base-pairs) contain all possible 136 tetranucleotides. Initial structures were generated with the nab module of the Amber16 package [32]. The structures were then neutralized with sodium ions and solvated to explicit solvent (TIP3P water model [33], 10 Å minimum distance between box boundary and DNA) with xleap. The OL15 force field was used [34]. Afterwards, the systems were energy minimized in 2500 steps with the sander module of the Amber16 package. Subsequent simulations were then performed with the pmemd.cuda module. Initially, the systems were heated up to 300 K in three consecutive simulations. Each of these simulations lasted 100 ps and induced a temperature increase of 100 K. Thereby, positional restraints were applied on all of DNA's non-hydrogen atoms. The restraints were reduced from 25 $\frac{\mathrm{kcal}}{\mathrm{mol\AA^2}}$ to 0.5 $\frac{\mathrm{kcal}}{\mathrm{mol\AA^2}}$ in five consecutive simulations at 300 K and a constant pressure of 1 bar (with a coupling constant of 5 ps). The equilibration was completed with a 2 ns simulation in which only the first two base-pairs were restrained with a small force constant of 0.1 $\frac{\mathrm{kcal}}{\mathrm{mol\AA^2}}$. The output structures were then used as input for the production runs, which lasted $\sim 1\,\mu$s and include the soft restraints on the terminal base-pairs. We used the same protocol for the equilibration of two 16 base-pair long B-DNA molecules (Table 7.1), for which we tested the harmonic and the Ising model. The production runs were performed for $2\,\mu$s.

sequence 1	5'-gcgcaatggagtacgc-3'
sequence 2	5'-cgtcaaagtacccagc-3'

Table 7.1: Sequence of the Watson-strand for the two simulated double-stranded DNA molecules. The Crick-strand is complementary. Based on the simulation of these sequences, we evaluated the accuracy of the Ising and standard harmonic model.

7.2.2 Parameterization and Implementation of the Ising Model

The standard harmonic and the Ising model were exclusively parameterized based on the simulations of the ABC-set [21]. Base-pair and base-pair step parameters for all simulations were calculated with Curves+ [35]. For each tetranucleotide, we calculated the equilibrium values for these parameters. In case of the Ising model, equilibrium values were calculated for the (backbone) substates. The backbone substates were specified by $\epsilon - \zeta$, these dihedral angles were also computed with Curves+. The stiffness parameters were obtained by inversion of the covariance matrices, $K = k_B T \cdot C^{-1}$. The covariance matrices were calculated starting from two base-pair steps prior up to two base-pair steps behind the corresponding tetranucleotide. From the resulting stiffness matrices, we extracted only the central part which represents the tetranucleotide segment. The reason for calculating covariance matrices for longer segments and cutting its inversed matrix afterwards is that in this way the stiffness matrix represents the case of being explicitly coupled to neighboring steps. This is desired for building a description for DNA's elasticity composed of tetranucleotide fragments. For the Ising model, this procedure was performed for every substate. Since our interaction matrix in the Ising model is spanned over two base-pair steps (less than a tetranucleotide segment), we have split tetranucleotide parameters in two fragments: The first fragment contains equilibrium and stiffness parameters for the first two base-pairs and base-pair steps, the second fragment for the second and third base-pairs and the last two base-pair steps. The application of our Ising model requires only the sequence, the number of the central base-pair step of the first tetranucleotide in this sequence, and prior processing of the MD simulation/crystal structure with Curves+. In our implementation of the Ising model, all parameters are then read in automatically, interaction matrices are calculated according to equation 7.2. The partition sum is then computed according to equation 7.9. In this way, the free energy is calculated for every frame in the trajectory.

7.2.3 MD Simulations of the Papillomavirus E2 Systems and Nucleosome Complexes

The papillomavirus E2 systems were built based on the 1jj4 pdb-structure [18]. In order to generate complexes with the five different central spacer sequences, all base-atoms of the spacer sequence except N1/N9, C2/C4 and C6/C8 were cut. The residue-names of

Position	original	A-tract
24-33	ACCAAAAGTG	AAAAAAAAAA
49-58	CCATCAAAAG	TTTTTAAAAA
89-100	CTTTTGATGGAG	TTTTTTAAAAAA
121-130	GGTAGAATCT	AAAAAAAAAA

Table 7.2: Substituted DNA sequences of the Watson-strand. The Crick-strand is always complementary.

these three atoms, which largely define the geometry of the base, and of the corresponding backbone atoms were reassigned according to the intended substitution. The missing base-atoms are then automatically added in xleap. The rest of the preparation of the systems, equilibration and MD simulations largely followed the same methodology as for the ABC-simulations. The production runs lasted 20 ns, the protein was described with the ff14SB force field [36] and no restraints were applied. The analog procedure was carried out for the five DNA duplexes alone, where the production runs lasted for 10 ns.
Nucleosome systems were constructed from the 2CV5 pdb-structure [37]. Here, two nucleosome systems differing in the DNA sequence were studied. In the first system (termed 'original'), the DNA sequence and structure of the pdb-file was left unmodified. In the second, we substituted several sites of the DNA by A-tract sequences. The sites of the substitutions are given in table 7.2. Note that substitution and all of the subsequent procedure was carried out in the same way as for the papillomavirus systems. Production runs for the nucleosome systems were performed for 20 ns. For the evaluation of all systems, we used only the second half of the production runs.

7.3 Results

7.3.1 Splitting DNA's Mechanical Information into a Subset of Tetranucleotides

The structure of double-stranded DNA can be described through three translational and three rotational degrees of freedom for each base-pairing as well as base-pair stacking [35, 38, 39]. Thus, it has become standard to also described DNA's deformability in terms of these parameters (e.g. twist, rise, slide, hereafter denoted by $\mathbf{q}$). As a starting point, we outline a procedure to model DNA's deformability based on the harmonic approximation from predetermined parameters for all 136 tetranucleotides. This description serves as a basis for the development of our Ising model. The parameters of the models are derived from $\sim 1\,\mu s$ long all-atom MD simulations of all 39 sequences studied in the ABC consortium. This set of DNA sequences includes all 136 tetranucleotides [21]. From these simulations we have extracted the stiffness matrices (with respect to the

internal base-pair (stacking) parameters) by inversion of the covariance matrices: $K = k_B T \cdot C^{-1}$. Furthermore, we also measure the mean values for all parameters, $\mathbf{q}_0$. The deformation energy for a tetranucleotide structure can then be approximated in harmonic form [14, 28, 40–42]:

$$E = \frac{1}{2} \cdot (\mathbf{q} - \mathbf{q}_0)^T \cdot K \cdot (\mathbf{q} - \mathbf{q}_0). \tag{7.1}$$

In order to describe now the deformation energy of any dsDNA-structure of arbitrary sequence, we first decompose the DNA's sequence into pieces of tetranucleotides. We then merge the associated stiffness matrices in the following way: For every tetranucleotide, we use the full 'self-terms' of its central step (indicated by cross-hatched area in the tetranucleotides of Fig 7.1). The 'self-terms' of the flanking sites are set to zero, while we scale coupling terms between central and flanking sites by a factor of 1/2 (single-hatched area in the tetranucleotide, Fig 7.1). Finally, we successively sum over all tetranucleotides, hence we obtain a mean deformability for nearest-neighbor coupling terms (shown as cross-hatched with green stroke in the full matrix of Fig 7.1). It has been shown that nearest-neighbor coupling terms play an important role in DNA deformability [22, 26]. This is already established in the aforementioned method. Note that it excludes non-nearest-neighbor interaction (indicated by 0 in the full matrix of Fig 7.1), which is justified as distant sites do not physically interact [43]. However, so far our method relies on a pure harmonic description of DNA's deformability, but it has already been addressed by several studies that some sequences show a high level of multimodality [21, 26, 28, 44]. Multimodality means the existence of several states in one base-pair step. While a multimodal treatment of one base-pair step itself is fairly easy, the inclusion of nearest-neighbor coupling becomes challenging. This requires a theory, in which the states of all base-pairs are coupled to all nearest-neighbor states. Such a theory cannot be realized by expressing the Hamiltonian as a simple pure additive function, which has been state of the art so far. The derivation of a Hamiltonian which fulfills both, nearest-neighbor coupling and multimodality is the objective of this study. Therefore, we combine the harmonic approximation with an Ising model.

7.3.2 Inclusion of Multimodality

A harmonic model as introduced in the previous section fails to reproduce multimodal behavior, hence it is insufficient to accurately describe DNA's free energy landscape. The major source for DNA's deviation from unimodality is the BI and BII backbone substates. These substates are defined by the ϵ and ζ dihedrals (BI: $\epsilon - \zeta < 0$, BII: $\epsilon - \zeta > 0$), and their population is strongly sequence-dependent [21, 26, 28, 44, 45]. Generally, a backbone strand preferentially adopts the BI substate. In the following, we denote a base-pair step with BI, when Watson- and Crick-strand are in BI configuration and otherwise with BII. Dissecting now DNA's local conformational space into BI and BII harmonically-approximated subspaces provides a very sound basis for capturing

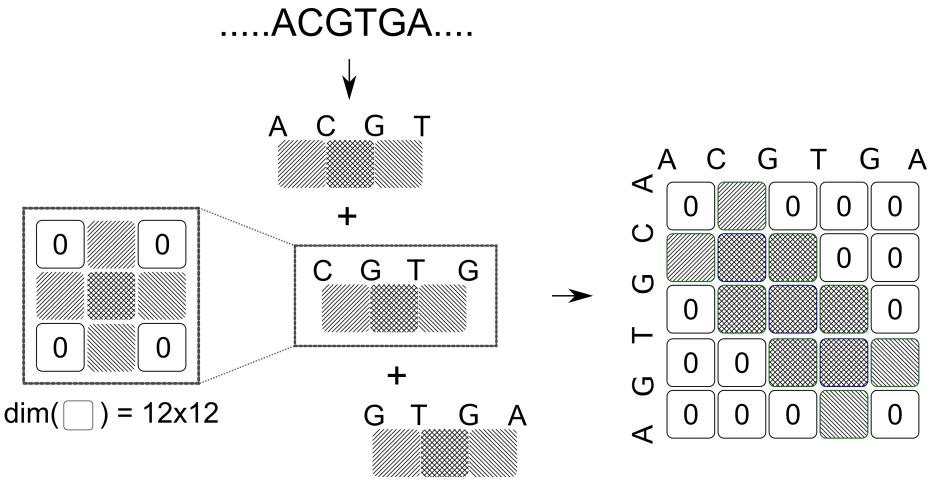


Figure 7.1: Modeling the deformability of a DNA polymer: DNA's sequence is split up into tetranucleotides. For every tetranucleotide, the deformability of the central step is fully retained (cross-hatched), whereas nearest-neighbor coupling is scaled by 1/2 (single-hatched). The 'self-terms' of the flanking sites as well as non-nearest-neighbor couplings are set to zero (indicated by 0-panels). The deformability for the entire polymer is then obtained by connecting the tetranucleotide-descriptions additively.

DNA's multimodal behavior (Fig 7.2) [26, 28]. We then link the splitting of one base-pair (step) into subspaces to its nearest-neighbors' subspaces by means of an Ising model. In this way, our description of local subspaces also includes coupling terms to nearest-neighbor subspaces. As shown in Fig 7.2, our Ising model reproduces the multimodal probability distribution significantly better than the harmonic approximation.

7.3.3 Ising Model

We model DNA's multimodality by discretizing DNA's conformational space with respect to the backbone substates: $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ denotes the BI, and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ the BII substate. In the following, the numbers of the base-pair (steps) are indexed by i and j, the corresponding sequences by k and l, and the (backbone) substates by m and n (can be either 0 or 1). The interaction between adjacent base-pair (steps) i and j is then determined by $E =< i \mid H_{ij} \mid j >=< i \mid H^{kl}_{mn} \mid j >$, with :

$$H^{kl}_{mn} = \begin{pmatrix} \frac{1}{2}\Delta q^{kl\,T}_{00} K^{kl}_{00} \Delta q^{kl}_{00} + E^{kl}_{00} & \frac{1}{2}\Delta q^{kl\,T}_{01} K^{kl}_{01} \Delta q^{kl}_{01} + E^{kl}_{01} \\ \frac{1}{2}\Delta q^{kl\,T}_{10} K^{kl}_{10} \Delta q^{kl}_{10} + E^{kl}_{10} & \frac{1}{2}\Delta q^{kl\,T}_{11} K^{kl}_{11} \Delta q^{kl}_{11} + E^{kl}_{11} \end{pmatrix}. \tag{7.2}$$

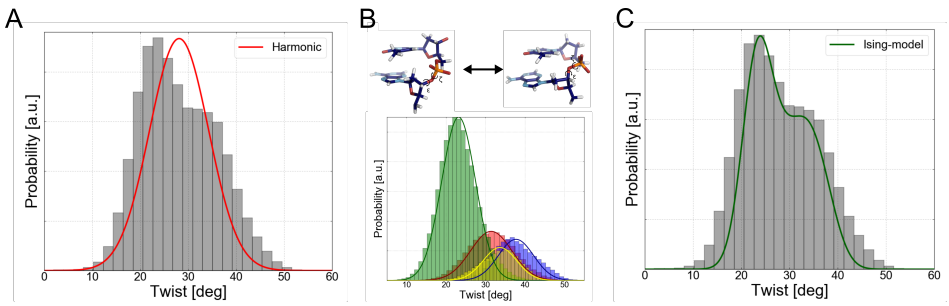


Figure 7.2: Bimodal distribution of a base-pair step parameter (twist). The standard harmonic model fails to accurately reproduce the probability distribution (A). The distribution can be decomposed into Gaussian distributions, which represent BI/BII states. The different colors represent different conformational spaces for the backbones (B). Building an Ising model based on discretization towards backbone states yields a significantly better reproduction of the probability distribution (C).

For each of the four states, we make a harmonic ansatz for the deformation energy. The stiffness matrices are computed from inversion of the covariance matrices, which have been calculated by accounting only for frames with the corresponding backbone configuration, and we tailor the stiffness matrices analogously to the full harmonic model. Additionally, we measure the set of equilibrium coordinates for every substate. In the following, we directly account for the deviation from the equilibrium parameters, denoted by Δq_{mn}^{kl}. All parameters are calculated from the simulations of the ABC-set containing the 136 tetranucleotides [21]. Note that the ground state energies E_{mn}^{kl} are so far undetermined. However, a harmonic approximation for the energy implies a Gaussian probability distribution:

$$\rho_{mn}^{kl} = \exp\left[-\beta \cdot \left(\frac{1}{2}\Delta q_{mn}^{kl\,T} K_{mn}^{kl} \Delta q_{mn}^{kl} + E_{mn}^{kl}\right)\right]. \tag{7.3}$$

The overall probability for state mn of sequence kl is then given by:

$$p_{mn}^{kl} = \int_{-\infty}^{+\infty} \exp\left[-\beta \cdot \left(\frac{1}{2}\Delta q_{mn}^{kl\,T} K_{mn}^{kl} \Delta q_{mn}^{kl} + E_{mn}^{kl}\right)\right] \mathrm{d}^N q = \tag{7.4}$$

$$= \sqrt{\frac{(2\pi)^N}{det(\beta \cdot K_{mn}^{kl})}} \exp(-\beta E_{mn}^{kl}). \tag{7.5}$$

On the other hand, p_{mn}^{kl} is a quantity we can directly measure from our simulations, and hence we are also able to calculate the ground state energies:

$$E_{mn}^{kl} = -\frac{1}{\beta}\ln\left(p_{mn}^{kl}\sqrt{\frac{det(\beta\cdot K_{mn}^{kl})}{(2\pi)^N}}\right). \tag{7.6}$$

In our description, we have scaled the E_{mn}^{kl} values by a factor of 1/2, due to the overlap of juxtaposed matrices. In this way, we obtain mean ground state energies. As shown in Fig 7.3, our approach provides a sound basis for capturing multimodal probability distributions. Thus, we can evaluate the full Hamiltonian for a DNA duplex with known configuration of the N base-pair (steps):

$$H =< 1 \mid H_{12} \mid 2 > + < 2 \mid H_{23} \mid 3 > +....+ < N-1 \mid H_{N-1,N} \mid N > . \tag{7.7}$$

The partition sum results from summation over all (backbone) substates:

$$Z = \sum_{|1>=\binom{1}{0},\binom{0}{1}} \sum_{|2>=\binom{1}{0},\binom{0}{1}} ... \sum_{|N>=\binom{1}{0},\binom{0}{1}} \exp[-\beta H]. \tag{7.8}$$

Applying the transfer-matrix method, we obtain a simple matrix product:

$$Z = \sum_{|1>=\binom{1}{0},\binom{0}{1}} \sum_{|N>=\binom{1}{0},\binom{0}{1}} < 1 \mid \Pi_{i=1}^{N-1}\exp[-\beta H_{i,i+1}] \mid N > . \tag{7.9}$$

Finally, we calculate the deformation free energy through $F = -k_B T \cdot \ln(Z)$.

7.3.4 Free Energy Profiles

In order to check the quality of our Ising model, we have performed MD simulations on two DNA test sequences (Table 7.1). From these simulations, we have computed free energy profiles along various modes through Boltzmann inversion (red dots, Fig 7.4 and Fig 7.5). We have then calculated the deformation energy for every snapshot of the trajectories by means of the pure harmonic model (blue dots) and our Ising model (green dots, Fig 7.4, 7.5). Obviously, the Ising model substantially decreases overestimations in the free energy and is hence overall in better agreement with the actual free energy profiles. Note that we achieve significant improvements for base-pair parameters (e.g. buckle and propeller), which have been omitted in the recent multimodality approach by Walther et al [23]. Besides, we have also compared the entire deformation energies between the two theories (Fig 7.6). For the harmonic model, we find an average deformation energy of $\sim$ 0.36 kcal/mol per degree of freedom for both sequences. The Ising model, instead, yields $\sim$ 0.33 kcal/mol in both cases and is thus in better agreement with the equipartition theorem ($\sim$ 0.30 kcal/mol). This

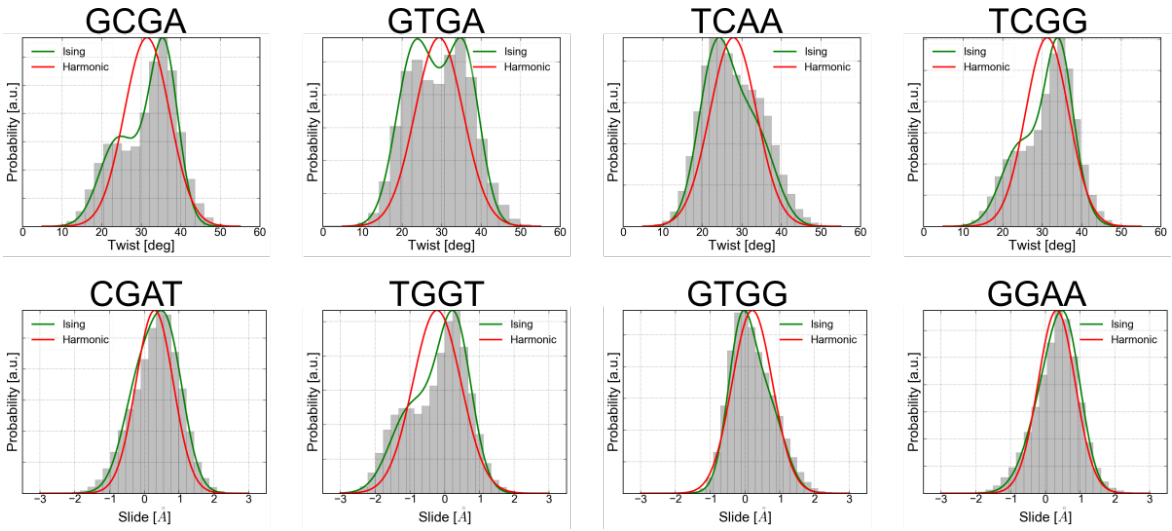


Figure 7.3: Probability distributions of the twist and slide parameter for the central step of various tetranucleotide sequences. The sampled distribution is shown as histogram plot in grey bars, predictions by the harmonic model in red and by the Ising model in green. The curves for the harmonic and the Ising model have been calculated from the parameter's variance, which has been obtained by inversion of the stiffness matrix of the corresponding model.

means that energy overestimation is on average reduced by $\sim$ 4.3 kcal/mol for 11 bp long unrestrained B-DNA structures. Due to the improvements in the free energy profiles and the better reproduction of the equipartition theorem, we therefore conclude that the Ising model offers a clearly better description of DNA's deformability than the standard harmonic model.

7.3.5 Application to DNA-Papillomavirus Complexes

We have applied the Ising model to the DNA/papillomavirus-E2 complex (see Fig 7.7) [18]. Papillomavirus E2 proteins (short E2 proteins) bind to ACGGNNNNCGGT sequences. Thereby, the binding affinity of the E2 protein strongly depends on the central spacer-sequence -NNNN-, although the protein hardly interacts with this sequence and instead contacts the adjacent major grooves [18]. However, interaction with the major grooves results in pronounced bending of the spacer sequence (Fig 7.7). Thus, the E2 system represents a model system for the indirect readout hypothesis: Proteins bind to specific sites by probing the local deformability. Experimental studies have shown that the ranking of the binding affinities for the spacer sequence is $AATT > AAAT > AAAA > TTAA > ACGT$ [18]. Moreover, it has also been pointed out that the BI/BII polymorphism plays an important role in this complex [46]. We have subjected the indirect readout hypothesis to the harmonic model and our Ising model. To this purpose, we have performed MD simulations for the protein-DNA complexes (20

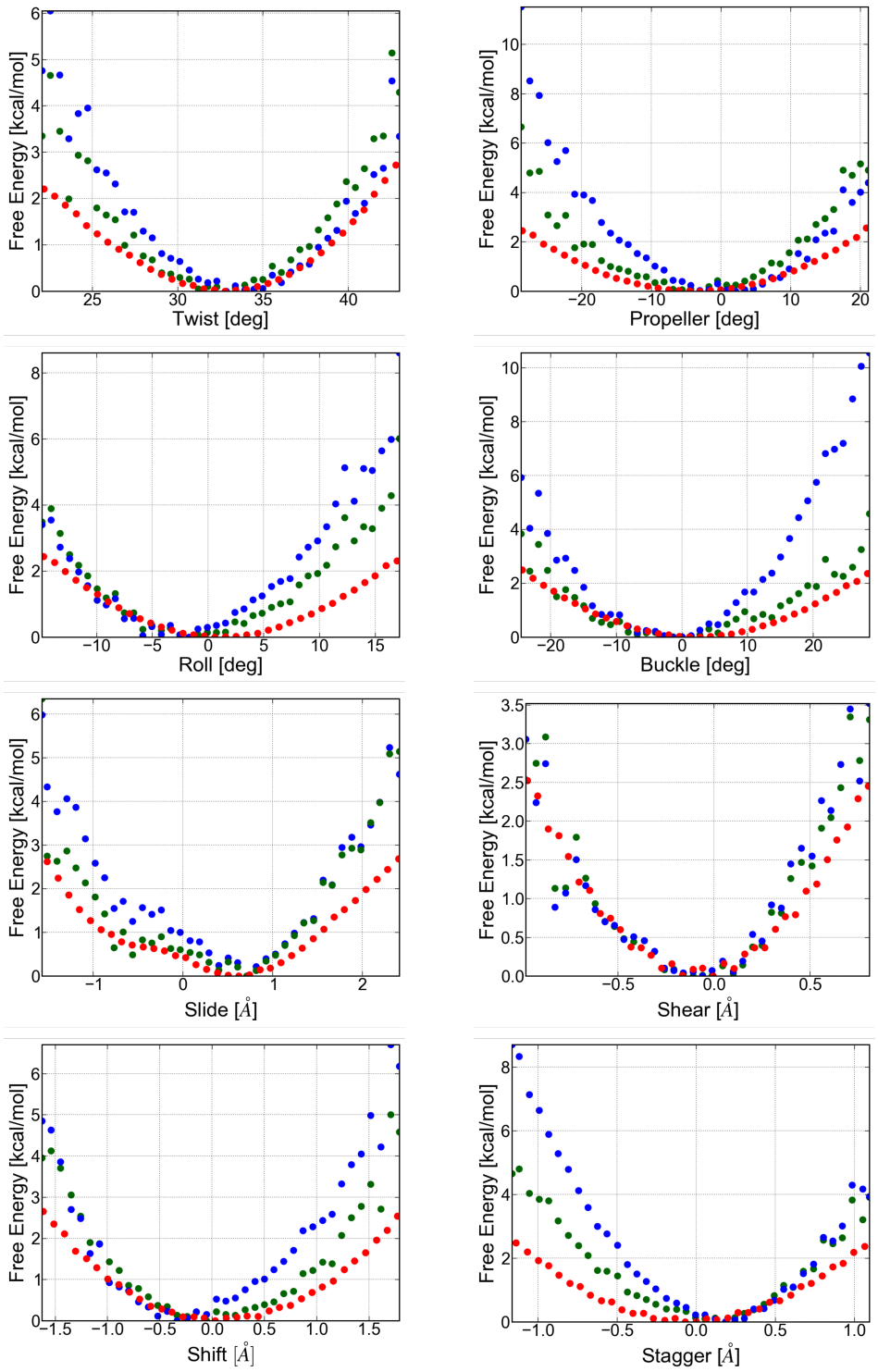


Figure 7.4: Relative free energy profiles obtained for different parameters from simulation of sequence 1. Red dots show the exact free energy curves, calculated by Boltzmann inversion. Free energy estimates from our Ising model are depicted in green and from the harmonic model in blue dots.

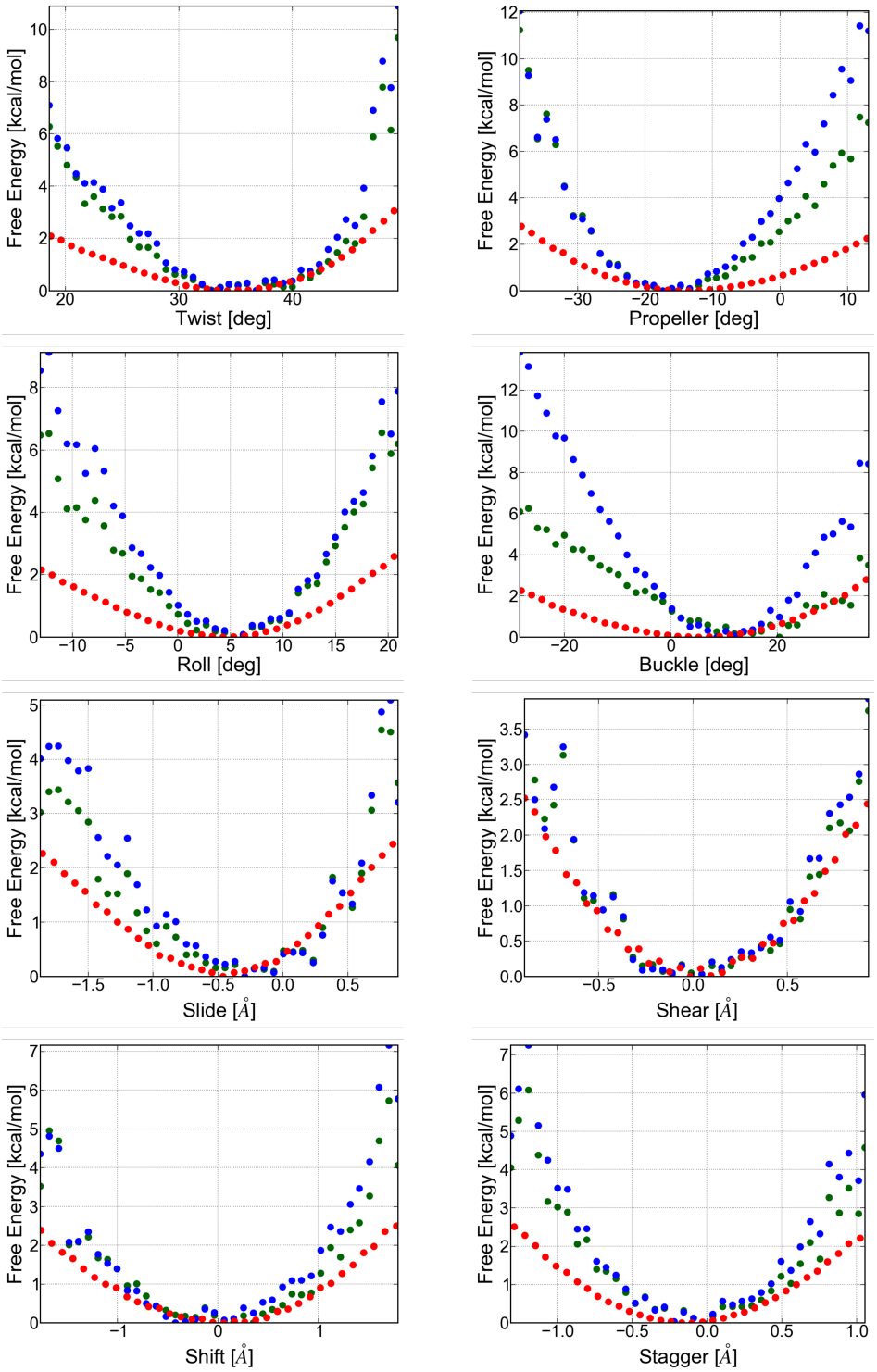


Figure 7.5: Relative free energy profiles obtained for different parameters from simulation of sequence 2, following the same methodology as in Fig 7.4.

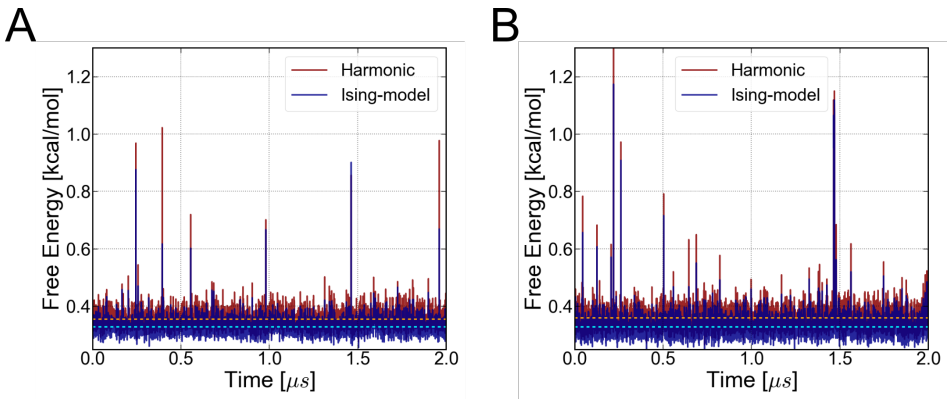


Figure 7.6: Free energy computations on MD simulations for sequence 1 (A), and sequence 2 (B). Blue curves represent calculations based on our Ising model, red curves based on the standard harmonic model. Free energies have been averaged in bins of 10 ps. Average free energies are shown by the dashed cyan and dashed orange lines for the Ising model and the harmonic description.

	ΔG^{exp}	HM single-molecule	Ising single-molecule	HM complex	Ising complex
AATT	-12.0	37.4 ± 0.3	33.9 ± 0.3	48.3 ± 0.4	38.2 ± 0.3
AAAT	-10.8	37.0 ± 0.3	33.9 ± 0.3	46.0 ± 0.3	39.2 ± 0.3
AAAA	-10.6	39.2 ± 0.3	36.3 ± 0.3	48.4 ± 0.4	41.7 ± 0.3
TTAA	-10.2	40.8 ± 0.3	36.9 ± 0.3	44.5 ± 0.4	41.6 ± 0.3
ACGT	-9.2	38.0 ± 0.3	35.1 ± 0.3	47.3 ± 0.3	43.2 ± 0.3

Table 7.3: Experimental binding affinities [18] and deformation energies calculated with the harmonic and the Ising model for DNA configurations simulated as single-molecule systems and in complex with the E2 protein. Values are in kcal/mol.

ns) and the DNA duplexes alone (10 ns) for each of the five sequences. We have then calculated the deformation energy as the difference in the average free energy between single-molecule and complex configurations for both models. For the Ising model, we find an excellent correlation between deformation energy and experimentally obtained binding affinity. In contrast, the harmonic model fails to predict correct ranking of the sequences (Fig 7.7 A). As shown in Table 7.3, the harmonic model estimates higher free energies than the Ising model for every simulation. The higher fluctuations in free energies eventually prevent the resolution of sequence-dependent effects based on a pure harmonic treatment. Thus, the Ising model performs significantly better in the computation of deformation free energies of DNA complexed to proteins than the harmonic model. Apparently, the Ising model is accurate enough to reproduce good agreement with experimental binding affinities, and is hence also supporting the indirect readout theory.

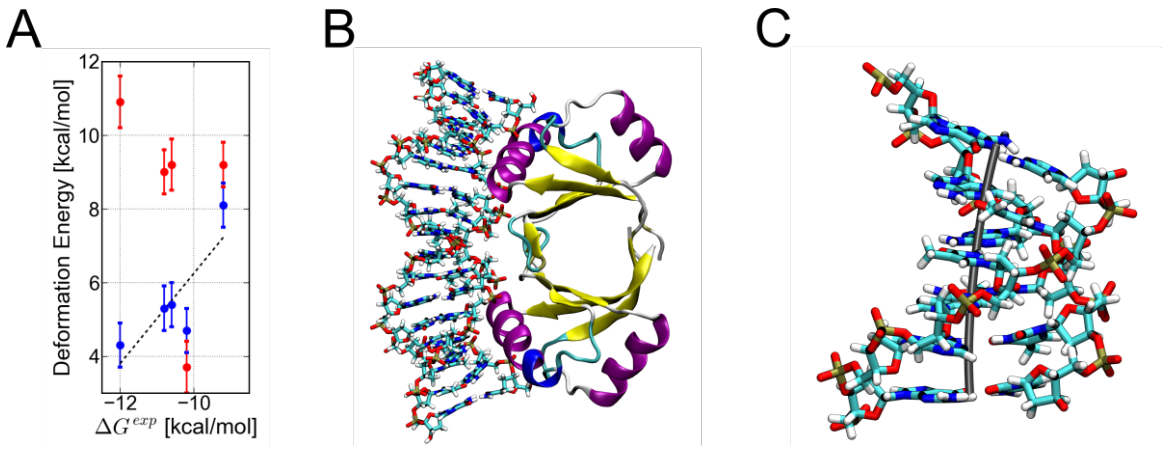


Figure 7.7: Free energy calculations on the papillomavirus E2 system: Correlation of computed deformation energies (blue: Ising model, red: harmonic model) with experimental binding affinities for the spacer-sequences AATT, AAAT, AAAA, TTAA and ACGT [18] (A). Snapshot of the E2 protein in complex with the DNA (B) and bending of the central spacer region (C).

7.3.6 Deformation Energies in Nucleosome Complexes - A-tract Depletion

In a next step, we have applied our Ising model to quantify deformation energies for DNA in the nucleosome complex. To this purpose, we have performed 20 ns simulations for the human nucleosome core particle (pdb 2CV5 [37], denoted as 'original' sequence) and for the same complex where we have substituted specific sites by A-tract sequences (Table 7.2). From these simulations, we have computed deformation energies along the DNA sequences. Deformation energies for A-tracts in the nucleosome complex are of particular interest because of their overabundance in eukaryotic genomes and their resistance to actually undergo nucleosome complex formation. This phenomenon is known as A-tract depletion in nucleosomes and plays a major role in chromatin folding [14, 31]. Though the intrinsic resistance of A-tracts towards nucleosome formation is uncontested and of high biological relevance, the mechanism behind it has not been clarified [31]. Dršata et al have studied the positioning of A-tracts in nucleosomes based on a harmonic model. Here, the authors find that A-tracts are associated with larger deformation energies than CG-sequences. However, this study is not only based on a rather simple description for DNA's deformability but also only resolves a very small subset of base-pair (step) parameters without allowing full structural adaption of the A-tracts to nucleosome binding [14]. Our methodology, in contrast, allows full structural adaption of the A-tracts, accounts for all base-pair (step) parameters and includes multimodal behavior. Application of the Ising model to both trajectories (Fig 7.8) yields the following results: First, the deformation energy is distributed rather uniformly for the heterogeneous, original sequence, except at the 50-60 base-pair position. At this site, kinking/partial unstacking was observed in the trajectory which may be a deficiency

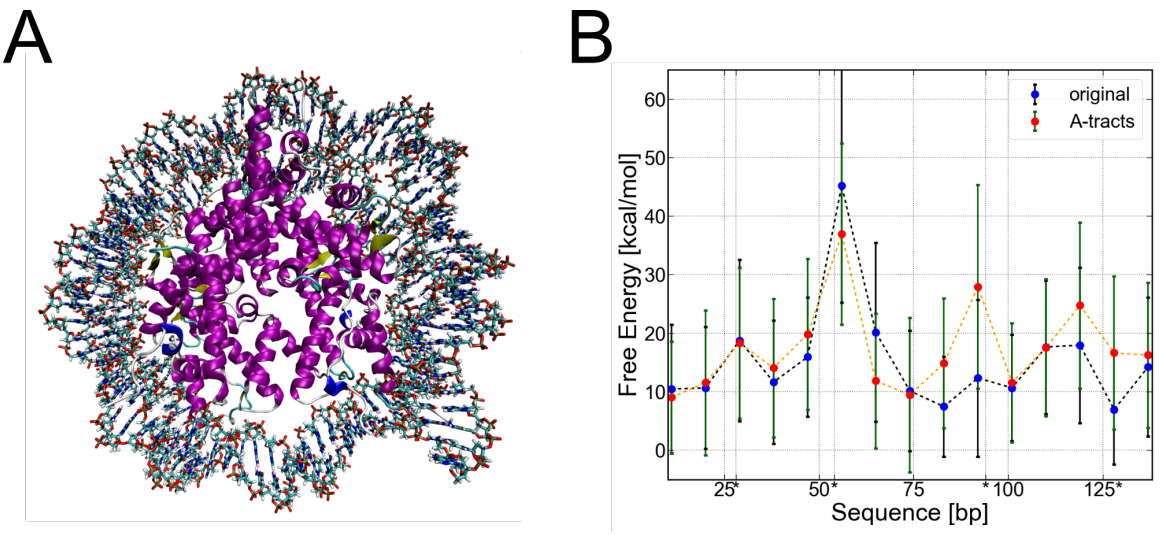


Figure 7.8: (A) Snapshot of the human nucleosome core particle [37] with heterogeneous DNA sequence (denoted as 'original'). (B) Deformation energies for the original DNA sequence and the A-tracts including sequence. Deformation energies have been calculated with our Ising model over 9 base-pair long DNA segments, and averaged over the last 10 ns of the MD trajectory. From the resulting energies, we have subtracted $\frac{k_B T}{2}$ for every degree of freedom. Positions of the A-tracts are labeled by asterisks, error bars represent standard deviations.

in the force field. Second, deformation energies for the A-tract sequence are higher. Summed over the total length of the DNA, the free energy is higher by 30.4 kcal/mol compared to the original sequence. Third, destabilization strongly depends on where the A-tracts are positioned, e.g. at the 24-33 base-pair position the placement of A-tracs is not unfavorable. Overall, our Ising model strongly indicates that A-tracs avoid nucleosome formation due to their stiffness and intrinsic structure. Thus, we further emphasize the role of DNA's sequence-dependent deformability for biological functions and the importance of a high-level theory modeling it.

7.4 Conclusion

Up to now, the development of coarse-grained models for DNA's free energy landscape has been challenging and from the mechanistic point of view incomplete. Base-pair stacking in DNA is strongly correlated to adjacent stacking interactions [22, 26], and base-pair stacking itself shows a high level of multimodality [21, 26, 28, 44]. Accounting for each of the two effects alone is simple, as the Hamiltonian can then be written in an additive, harmonic form. The difficulty arises when accounting for both effects simultaneously, as every substate in a base-pair step needs to be treated as correlated

to all nearest-neighbor substates. We have solved this problem by combining the harmonic model with the Ising theory. Thereby, we discretize DNA's conformational space into backbone-subspaces and introduce a harmonic interaction matrix between nearest-neighbor base-pair steps. In this way, we also account for all possible nearest-neighbor couplings. We have parameterized our Ising model for all 136 DNA tetranucleotides from MD simulations on the ABC-set of DNA sequences [21]. Notably, our Ising model does not only account for base-pair stacking but also for base-pairing deformabilities, which has been neglected in the coarse-grained model by Walther et al [23]. Evaluations of the free energy profiles for two dsDNA sequences reveal a substantially higher accuracy of our Ising model than the standard harmonic model. The overestimation of deformation free energies is strongly reduced in the Ising model, hence resulting in closer agreement with the Virial theorem. In a next step, we have studied protein-DNA systems. For five papillomavirus E2/DNA complexes, differing in the central spacer sequence, we have calculated reliable deformation energies for the DNA based on our Ising model. Furthermore, we have found excellent agreement between the sequence-dependent deformation energies and experimentally obtained binding affinities. To the best of our knowledge, this is the first time that reliable DNA deformation energies in correlation with binding affinities are reported, and we emphasize that these findings strongly support the indirect readout theory. Besides, we also point out that investigating this phenomenon requires a high level theory for DNA's deformability, as the standard harmonic model fails to accurately quantify deformation energies. In addition, we have analyzed sequence-dependent deformation energies in nucleosome complexes, whereby we have focused on A-tract depletion in nucleosomes. A-tracts are frequently occurring sequences in eukaryotic genomes and their pronounced reluctance to undergo nucleosome formation plays an important role in chromatin folding [31]. Here, we have shown that A-tracts mechanically destabilize nucleosome complex formation with respect to a heterogeneous DNA sequence. This further stresses the role of DNA's sequence-dependent deformability in chromatin folding and gene expression. Altogether, we conclude that calculation of realistic sequence-dependent deformation energies upon protein binding has become feasible through our Ising model. Moreover, our protocol is easy and efficient to use, it only requires preceding processing of the MD-trajectory/crystal structure with Curves+ and the DNA-sequence as input. Development of a Monte-Carlo procedure to study deformation energies for DNA under global stress will be a next step. This will also enable us to screen sequence-dependent effects in high throughput.

7.5 Bibliography

[1] Wilma K. Olson, Andrey A. Gorin, Xiang-Jun Lu, Lynette M. Hock, and Victor B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, 95(19):11163–11168, 1998.

[2] Filip Lankaš, Jiří Šponer, Jörg Langowski, and Thomas E. Cheatham. DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophysical Journal*, 85(5):2872–2883, 2003.

[3] Filip Lankaš, Jiří Šponer, Jörg Langowski, and Thomas E. Cheatham. DNA Deformability at the Base Pair Level. *J. Am. Chem. Soc.*, 126(13):4124–4125, 2004.

[4] Marcos J. Araúzo-Bravo, Satoshi Fujii, Hidetoshi Kono, Shandar Ahmad, and Akinori Sarai. Sequence-Dependent Conformational Energy of DNA Derived from Molecular Dynamics Simulations: Toward Understanding the Indirect Readout Mechanism in ProteinDNA Recognition. *J. Am. Chem. Soc.*, 127(46):16074–16089, 2005.

[5] Satoshi Fujii, Hidetoshi Kono, Shigeori Takenaka, Nobuhiro Go, and Akinori Sarai. Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.*, 35(18):6063–6074, 2007.

[6] Remo Rohs, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.*, 79(1):233–269, 2010.

[7] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–1253, 2009.

[8] Michael Y. Tolstorukov, Andrew V. Colasanti, David M. McCandlish, Wilma K. Olson, and Victor B. Zhurkin. A Novel Roll-and-Slide Mechanism of DNA Folding in Chromatin: Implications for Nucleosome Positioning. *Journal of Molecular Biology*, 371(3):725–738, 2007.

[9] Bo Duan, Pengfei Ding, Timothy R Hughes, William Wiley Navarre, Jun Liu, and Bin Xia. How bacterial xenogeneic silencer rok distinguishes foreign from self DNA in its resident genome. *Nucleic Acids Res.*, 46(19):10514–10529, 2018.

[10] Nils B. Becker, Lars Wolff, and Ralf Everaers. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res.*, 34(19):5638–5649, 2006.

[11] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thstrm, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[12] Horace R. Drew and Andrew A. Travers. DNA bending and its relation to nucleosome positioning. *Journal of Molecular Biology*, 186(4):773–790, 1985.

[13] Johanna Virstedt, Torunn Berge, Robert M. Henderson, Michael J. Waring, and Andrew A. Travers. The influence of DNA stiffness upon nucleosome formation. *Journal of Structural Biology*, 148(1):66–85, 2004.

[14] Tomáš Dršata, Nad'a Špačková, Petr Jurečka, Marie Zgarbová, Jiří Šponer, and Filip Lankaš. Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Res.*, 42(11):7383–7394, 2014.

[15] Wei Chen, Pengmian Feng, Hui Ding, Hao Lin, and Kuo-Chen Chou. Using deformation energy to analyze nucleosome positioning in genomes. *Genomics*, 107(2):69–75, 2016.

[16] Christina S. Hines, Colin Meghoo, Sanjay Shetty, Markus Biburger, Michael Brenowitz, and Rashmi S. Hegde. DNA structure and flexibility in the sequence-specific binding of papillomavirus E2 proteins. *Journal of Molecular Biology*, 276(4):809–818, 1998.

[17] Haim Rozenberg, Dov Rabinovich, Felix Frolow, Rashmi S. Hegde, and Zippora Shakked. Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets. *Proc. Natl. Acad. Sci. USA*, 95(26):15194–15199, 1998.

[18] Seung-Sup Kim, Jeffrey K. Tam, Ai-Fei Wang, and Rashmi S. Hegde. The Structural Basis of DNA Target Discrimination by Papillomavirus E2 Proteins. *Journal of Biological Chemistry*, 275(40):31245–31254, 2000.

[19] Jeff M. Zimmerman and L. James Maher III. Solution measurement of DNA curvature in papillomavirus E2 binding sites. *Nucleic Acids Res.*, 31(17):5134–5139, 2003.

[20] Z. Otwinowski, R. W. Schevitz, R.-G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, 335(6188):321–329, 1988.

[21] Marco Pasi, John H. Maddocks, David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham, III, Pablo D. Dans, B. Jayaram, Filip Lankaš, Charles Laughton, and et al. μABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. *Nucleic Acids Res.*, 42(19):12272–12283, 2014.

[22] Marie Zgarbová, Petr Jurečka, Filip Lankaš, Thomas E. Cheatham, Jiří Šponer, and Michal Otyepka. Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Inf. Model.*, 57(2):275–287, 2017.

[23] Jürgen Walther, Pablo D Dans, Alexandra Balaceanu, Adam Hospital, Genís Bayarri, and Modesto Orozco. A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res.*, 48(5):e29, 2020.

[24] Jasmin Cevost, Cédric Vaillant, and Sam Meyer. ThreaDNA: predicting DNA mechanics contribution to sequence selectivity of proteins along whole genomes. *Bioinformatics*, 34(4):609–616, 2017.

[25] Richard Lavery, Krystyna Zakrzewska, David Beveridge, Thomas C Bishop, David A Case, 3rd Cheatham, Thomas, Surjit Dixit, B Jayaram, Filip Lankas, Charles Laughton, John H Maddocks, Alexis Michon, Roman Osman, Modesto Orozco, Alberto Perez, Tanya Singh, Nad'a Špačková, and Jiří Šponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, 38(1):299–313, 2010.

[26] Korbinian Liebl and Martin Zacharias. How methyl-sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.*, 47(3):1132–1140, 2018.

[27] Brahim Heddi, Nicolas Foloppe, Nadia Bouchemal, Edith Hantz, and Brigitte Hartmann. Quantification of DNA BI/BII Backbone States in Solution. Implications for DNA Overall Structure and Recognition. *J. Am. Chem. Soc.*, 128(28):9170–9177, 2006.

[28] Tomáš Dršata and Filip Lankaš. Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3(4):355–363, 2013.

[29] Marie Zgarbová, F. Javier Luque, Jiří Šponer, Thomas E. Cheatham, Michal Otyepka, and Petr Jurečka. Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.*, 9(5):2339–2354, 2013.

[30] Lennart DeBruin and John H Maddocks. cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.*, 46(W1):W5–W10, 2018.

[31] Eran Segal and Jonathan Widom. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current opinion in structural biology*, 19(1):65–71, 2009.

[32] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[33] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[34] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham, Rodrigo Galindo-Murillo, and Petr Jurečka. Refinement of the Sugar-Phosphate Backbone

Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.*, 11(12):5723–5736, 2015.

[35] R Lavery, M Moakher, JH Maddocks, D Petkevičiutė, and K Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37(17):5917–5929, 2009.

[36] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[37] Yasuo Tsunaka, Naoko Kajimura, Shin-ichi Tate, and Kosuke Morikawa. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res.*, 33(10):3424–3434, 2005.

[38] Wilma K Olson, Manju Bansal, Stephen K Burley, Richard E Dickerson, Mark Gerstein, Stephen C Harvey, Udo Heinemann, Xiang-Jun Lu, Stephen Neidle, Zippora Shakked, Heinz Sklenar, Masashi Suzuki, Chang-Shung Tung, Eric Westhof, Cynthia Wolberger, and Helen M Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology*, 313(1):229–237, 2001.

[39] Xiang-Jun Lu and Wilma K. Olsen. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, 31(17):5108–5121, 2003.

[40] Tomáš Dršata, Alberto Pérez, Modesto Orozco, Alexandre V Morozov, Jiří Šponer, and Filip Lankaš. Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, 9(1):707–721, 2013.

[41] Korbinian Liebl and Martin Zacharias. Unwinding Induced Melting of Double-Stranded DNA Studied by Free Energy Simulations. *J. Phys. Chem. B*, 121(49):11019–11030, 2017.

[42] O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks. A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.*, 138(5):055102, 2013.

[43] Korbinian Liebl and Martin Zacharias. How global DNA unwinding causes non-uniform stress distribution and melting of DNA. *PLOS ONE*, 15(5):e0232976, 2020.

[44] Pablo Daniel Dans, Ignacio Faustino, Federica Battistini, Krystyna Zakrzewska, Richard Lavery, and Modesto Orozco. Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, 42(18):11304–11320, 2014.

[45] Brahim Heddi, Christophe Oguey, Christophe Lavelle, Nicolas Foloppe, and Brigitte Hartmann. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.*, 38(3):1034–1047, 2009.

[46] James C. Robertson and Thomas E Cheatham. Dna Backbone BI/BII Distribution and Dynamics in E2 Protein-Bound Environment Determined by Molecular Dynamics Simulations. *J. Phys. Chem. B*, 119(44):14111–14119, 2015.

8 Tumuc1: Parameterization of a New DNA Force Field

8.1 Introduction

Over the last decades, Molecular Dynamics simulations have strongly matured due to improvements in both, soft- and hardware, which promoted the field into the microsecond-timescale [1–5]. Efforts in the theoretical description (force field development), in contrast, are still mainly based on deprecated parameters. Although overstabilization of stacking interactions [5–13] and understabilization of hydrogen bonding [4, 14, 15] has been reported in several studies, the nonbonded parameters in Amber's DNA force fields [4, 16–21] have not been revised since the pioneering publication by Cornell et al in 1995 [22]. The same also holds largely true for RNA and proteins, over the last $\sim$ 25 years force field development has focused on refinements of the dihedral parameters. Probably, this approach is followed by the scientific community because of its simplicity: Changes in nonbonded parameters are way more challenging, as this instantaneously requires reparameterization of dihedral angles, but not vice versa. From the theoretical point of view, maintaining the nonbonded parameters (especially charges) from the '90s is very questionable: Charges were derived on a low level of QM-theory (Hartree Fock) and calculated on small DNA-fragments with a low grid resolution. These choices were reasonable $\sim$ 25 years ago due to a much lower computing capacity [22, 23]. From today's point of view, however, we strongly doubt that this can be considered as a parameterization procedure on a high level. In addition, state of the art DNA force fields also show deficiencies in the dihedral parameters, which results in sampling bugs.
In this project, we undertake a radical, yet constructive approach to establish a sound DNA force field. The parameters in our force field ('Tumuc1', for 1st generation of the Technical University Munich) are derived from high-level QM calculations (RI-MP2/def2-TZVP). Thereby, bond- and angle-terms are parameterized with the modified Seminario method [24]. All partial charges are calculated by fitting of electrostatic potentials based on the Merz-Kollman procedure [23, 25, 26]. In a next step, we have scanned the QM potential energy landscapes for all relevant dihedral angles, and then fitted the dihedral parameters to achieve best possible agreement between the MM and QM potential energy landscapes. For the Lennard-Jones parameters, the Tumuc1 force field relies on the old Cornell parameters. The Tumuc1 force field is as easy to use as all standard Amber force fields and has been tested on several DNA-systems: We achieve an excellent description of B-DNA's helical structure, and a significantly better sampling of

backbone states than the bsc1 force field, which is up to now considered as state of the art. Our force field shows precise population of the BI and BII states, good sugar puckering and a very robust α/γ population, which prevents serious, transient distortions in DNA's local structure as it occurs in the bsc1 force field. For the Dickerson-Drew-Dodecamer, the Tumuc1 force field shows an average rmsd of $1.52\mathring{A}$ with respect to the crystal structure and likewise the stacking-parameters are in excellent agreement with experiments. Here, Tumuc1 also maintains its good performance with respect to changes in the solvent conditions. Simulation with the OPC water model yields (compared to TIP3P) an even lower rmsd of $1.41\mathring{A}$. Importantly, (in vacuo) base-pair hydrogen bonding is stronger in the Tumuc1 description compared to classical Amber force fields, which results in lower base-pair breathing for the terminal base-pairs. Moreover, we have been able to simulate hybridization of DNA single-strands into the correct duplex structure with Tumuc1, and we demonstrate successful folding of a DNA hairpin in an unrestrained MD simulation. DNA quadruplex structures remain stable within the μs-timescale in the Tumuc1 force field, whereas it exhibits stronger base-pair breathing for Z-DNA but better conservation of the left-handed helicity than the bsc1 force field.
We conclude from the test simulations, that the Tumuc1 force field provides a sound standard for the simulation of DNA systems. In particular with respect to B-DNA, we state clear progress compared to the bsc1 force field. It is hence not only competitive with current force fields, but has also been designed in a way which allows future refinements: In contrast to other force fields, parameterization of Tumuc1 is based on just the QM theory. Additionally, we provide a library composed of more atom types, which makes changes in the Lennard-Jones parameters for instance easier. Thus, we are convinced that the Tumuc1 force field opens necessary long-term prospects. It represents a clear alternative to the conventional force fields, especially with respect to the electrostatic description, and creates strong competition in the field of force field development, which will prove highly beneficial in the long run. Parameterization of our DNA force field required elaboration of parameterization techniques. For the future, we are also going to make use of these for the force field development for other biomolecules, e.g. RNA.

8.2 Results and Discussion

8.2.1 Parameterization

General Outline

The Tumuc1 force field is a classical force field derived from QM calculations in a bottom-up approach. Our force field description relies on the additive Cornell-form [22, 27]:

$$E = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedral} V_n[1 + \cos(n\phi - \gamma)] + \\ + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{q_i q_j}{r_{ij}} + \epsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right]. \quad (8.1)$$

Thereby, we stick to the same Lennard-Jones parameters as the conventional Amber force fields for two reasons: First, these parameters are already physically sensible. Second, we observe the electrostatic contributions to be absolutely dominant in hydrogen bonding and stacking landscapes outside Pauli exclusion effects. We have therefore derived an entirely new set of partial charges as well as bond- and angle-parameters from QM calculations. Consequently, we have parameterized the relevant dihedral terms employing the previously established set of parameters. For trivial dihedral angles (e.g. dihedrals within a nucleobase) we use the standard parameters.

Partial Charges, Bond- and Angle-Parameters

We have computed RI-MP2 implicit solvent ($\epsilon = 80$) geometry optimizations [28–31] on model-systems slightly larger than a nucleotide. From these calculations, we have inferred partial charges with the Merz-Kollman procedure, whereby we incorporate charge constraints [23, 25, 26]. Given the size of our model-systems, we can enforce the partial charges of a whole nucleotide to sum up to -1 (Fig 8.1 A,B). Thus, our force field will maintain integer charges for any DNA-system, so that the latter can always be neutralized by counterions. Furthermore, we also include equivalence constraints, e.g. hydrogens bound to the C2'-atom should be physically identical. Note that the resulting electrostatic description shows significant differences to classical Amber force fields, such as substantially stronger (in vacuo) hydrogen bonding (Fig 8.2).
Parameters for bonded interactions (bonds, angles, dihedrals) have been determined from in vacuo QM calculations. In order to obtain bond- and angle-parameters, we have performed QM frequency calculations on slightly smaller model-systems than for the charges (overall four nucleoside- and two backbone-models). We then have applied the modified Seminario approach, recently developed by Allen and coworkers [24]. This method excels through improved mapping of the QM-derived Hessian matrix in Cartesian space to deformability in internal modes (bonds, angles). The derivation directly from the QM Hessian matrix is highly beneficial for our bottom-up approach, as the bond and angle parameters are then independent of all other force field parameters in contrast to interpolation approaches (Fig 8.1 C). Thus, we do not need to update them iteratively. While current Amber force fields still use the empirical bond- and angle-parameters established $\sim$ 25 years ago, we now provide a new set in consistency with QM calculations. This may also prove advantageous for the dihedral parameters, which are generally derived from fitting MM energies to QM scans.

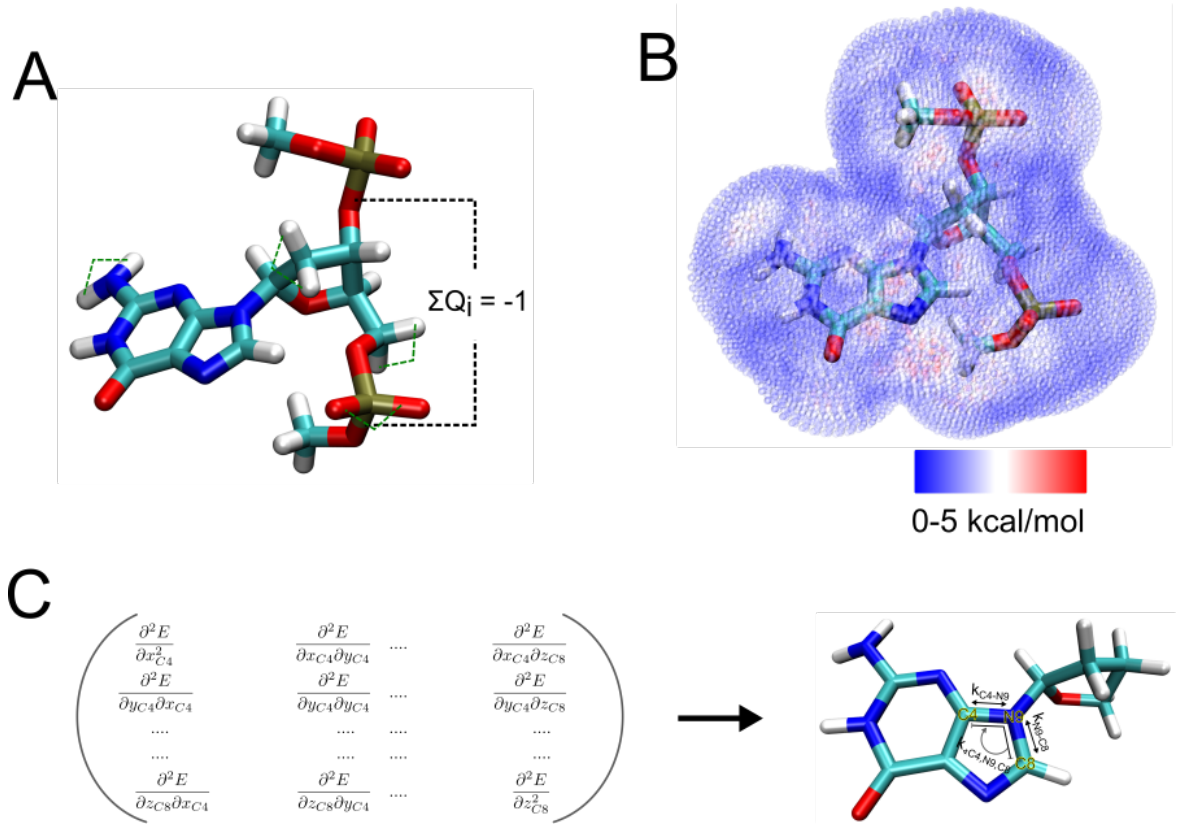


Figure 8.1: Derivation of charges, bond- and angle-parameters. (A) Model-system for the computation of charges. Equivalence constraints have been applied (dashed green lines). The sum of charges over a whole nucleotide has been constrained to -1. (B) Illustration of the fitting-procedure. Charges are fit to optimally reproduce the QM electrostatic potential on a grid around the molecule. Errors in the calculation are shown color-coded. Blue dots mean accurate description by the partial charges, red dots deviations by up to $\sim 5\,\mathrm{kcal/mol}$. (C) Bond- and angle-parameters are obtained with the modified Seminario method. In this method, an accurate projection of the Hessian matrix in Cartesian space (computed in QM calculations) on the internal parameters is achieved.

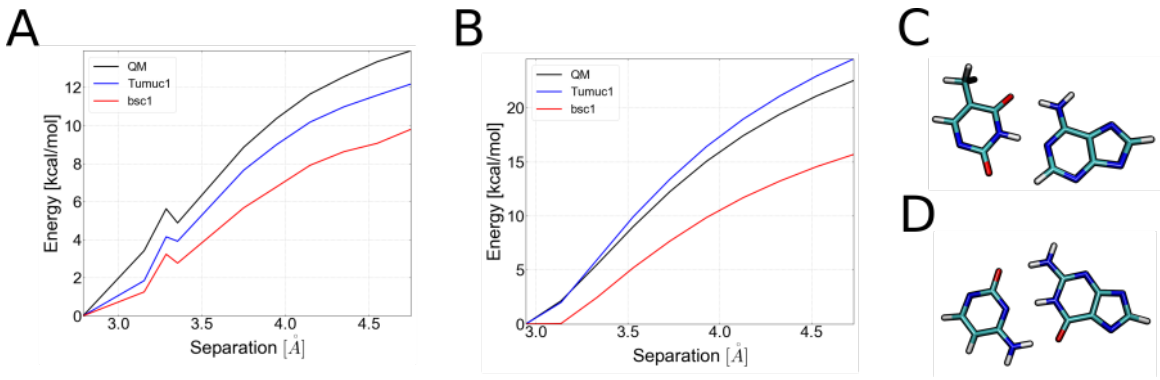


Figure 8.2: In vacuo hydrogen bonding for QM (RI-MP2/def2-TZVP), bsc1 and the Tumuc1 description. The scan has been performed by constraining interatomic distances between the base-pairs stepwise to higher values. (A) Hydrogen bonding energies for an A-T base-pair. (B) Hydrogen bonding energies for a G-C base-pair. (C,D) Systems used in the calculations. Note that we have used the same structures for QM and MM calculations. For the MM calculations, however, the hydrogen atoms on the glycosidic nitrogens have been removed.

Parameterization of Dihedral Angles

Parameters for all relevant dihedral angles have been deduced from three different types of model-systems (Fig 8.3). For each dihedral angle, we have performed a QM scan (RI-MP2/def2-TZVP) [29, 30] employing constraints on specific internal modes (see Materials and Methods section for details). Note that fitting of the dihedral parameters requires a force field description for each model-system in consistency with the final DNA force field. Thus, we have derived charges for the model-systems by multi-configurational fitting of the electrostatic potential using the same procedure as for the nucleotide-like system (see previous paragraph). Furthermore, we have also incorporated the bond- and angle-parameters which are used in the DNA force field. The description of the model-systems is then completed by setting the torsional potential for the scanned dihedral angles to zero. Subsequently, we have minimized the output-structures from the QM-scans with respect to our MM-descriptions. Thereby, we maintain the constraints of the QM calculations. In this way, we have generated MM-profiles along the relevant dihedrals. These MM-profiles can be understood as 'noise-profiles', because the actual torsional barriers are excluded and remain to be parameterized. We then have fitted the dihedral parameters to minimize the differences between QM and MM landscapes, schematically shown for the α, β and γ dihedrals:

$$f(\alpha, \beta, \gamma, V_n^\alpha, \delta_n^\alpha, V_n^\beta, \delta_n^\beta, V_n^\gamma, \delta_n^\gamma) =$$
$$\sum_{\gamma}^{(scan)} \sum_{\beta}^{(scan)} \sum_{\alpha}^{(scan)} ((E_{QM} - E_{MM}(V_\alpha = V_\beta = V_\gamma = 0) - V_\alpha - V_\beta - V_\gamma) \cdot w(\alpha, \beta, \gamma))^2. \quad (8.2)$$

Minimization of $f(\alpha, \beta, \gamma, V_n^\alpha, \delta_n^\alpha, V_n^\beta, \delta_n^\beta, V_n^\gamma, \delta_n^\gamma)$ is achieved by fitting of the torsional potentials V_α, V_β and V_γ. These functions depend on the dihedrals parameters $V_n^\alpha, \delta_n^\alpha, V_n^\beta, \delta_n^\beta$, $V_n^\gamma, \delta_n^\gamma$ and have the form of a truncated Fourier series:

$$V_\phi = \sum_{n=1} V_n^\phi \cdot [1 + \cos(n \cdot \phi - \delta_n^\phi)]. \quad (8.3)$$

We have usually expanded the torsional potentials to fourth order. In our approach, the relevant torsional potentials are fitted simultaneously. The rationale of this procedure is to account for the interdependency of dihedral angles (e.g. changes in α can affect the β and γ dihedrals). For this reason, we have also performed scanning of dihedral angles with coupled dihedrals constrained to specific values (see Materials and Methods section). Besides, we introduce a weighting function $w(\alpha, \beta, \gamma)$ to ensure a higher fitting-quality for important conformational spaces. Thereby, $w(\alpha, \beta, \gamma)$ is flexible: Overestimating the MM energy of unfavorable regions is penalized by low weights, whereas underestimation of such regions is strongly penalized. The motivation behind this design is that destabilization of conformational spaces, which are absent in DNA's structure, is practically without consequences. Stabilization of such regions, in contrast, may allow them to occur in MD simulations and hence have serious effects. The QM potential energy surfaces and fitted MM profiles for the backbone dihedrals are shown in Fig 8.4 and 8.5.

In our protocol, model-system C of Fig 8.3 represents a special case, as it includes four different nucleosides. Thereby, we have parameterized the χ dihedrals by restraining the sugar pucker to C2'-endo conformation (Fig 8.6). In a next step, we have parameterized the δ-dihedral by scanning of the cytosine nucleoside, while constraining it to C2'-endo conformation (Fig 8.7). Thereafter, we use the same δ-parameters for all nucleosides. The χ- and δ-parameters have been considered in the parameterization for the sugar pucker, which is defined by the dihedrals in the sugar ring. We describe sugar puckering for all nucleosides by the same parameter-set for $\nu_0, ..., \nu_4$ [32, 33]. For parameterization of the sugar puckering, we adjusted the weight function to allow destabilization of the C3'-endo with respect to the C2'-endo state. Based on test-runs with the new DNA force field, we have decided to destabilize the C3'-endo state by another 0.35 kcal/mol. This empirical modification has been made due to a high proportion of C3'-endo states for cytosine-bases. Intriguingly, similar decisions have also been made in other QM-based parameterization approaches [34].

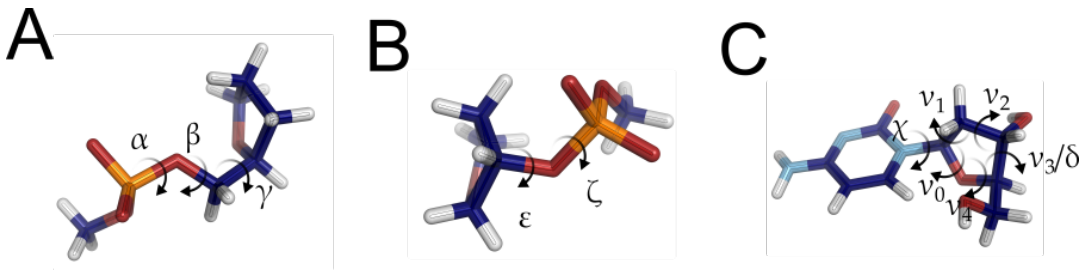


Figure 8.3: Model-systems for parameterization of the dihedral angles. (A) Parameterization of the $\alpha-, \beta-$ and γ-dihedrals. (B) Model-system for the $\epsilon-$ and $\zeta-$dihedrals. (C) Dihedral-parameters for the glycosidic bonds and sugar puckering are derived from nucleoside systems. All nucleosides are described by the same sugar puckering. The δ-dihedral is parameterized solely from the cytosine-nucleoside.

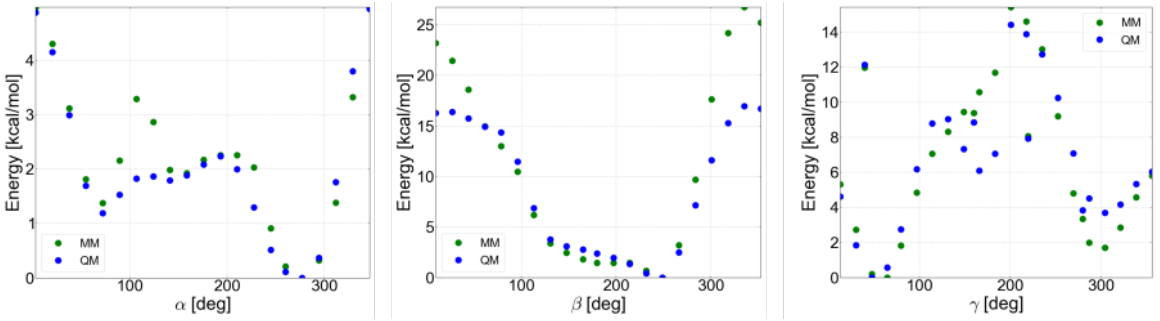


Figure 8.4: QM energy landscapes for scans of the $\alpha-, \beta-$ and $\gamma-$dihedrals are illustrated in blue. The fitted MM profiles are shown in green. The γ-profile also includes calculations with the $\beta-$dihedral constrained to 70°.

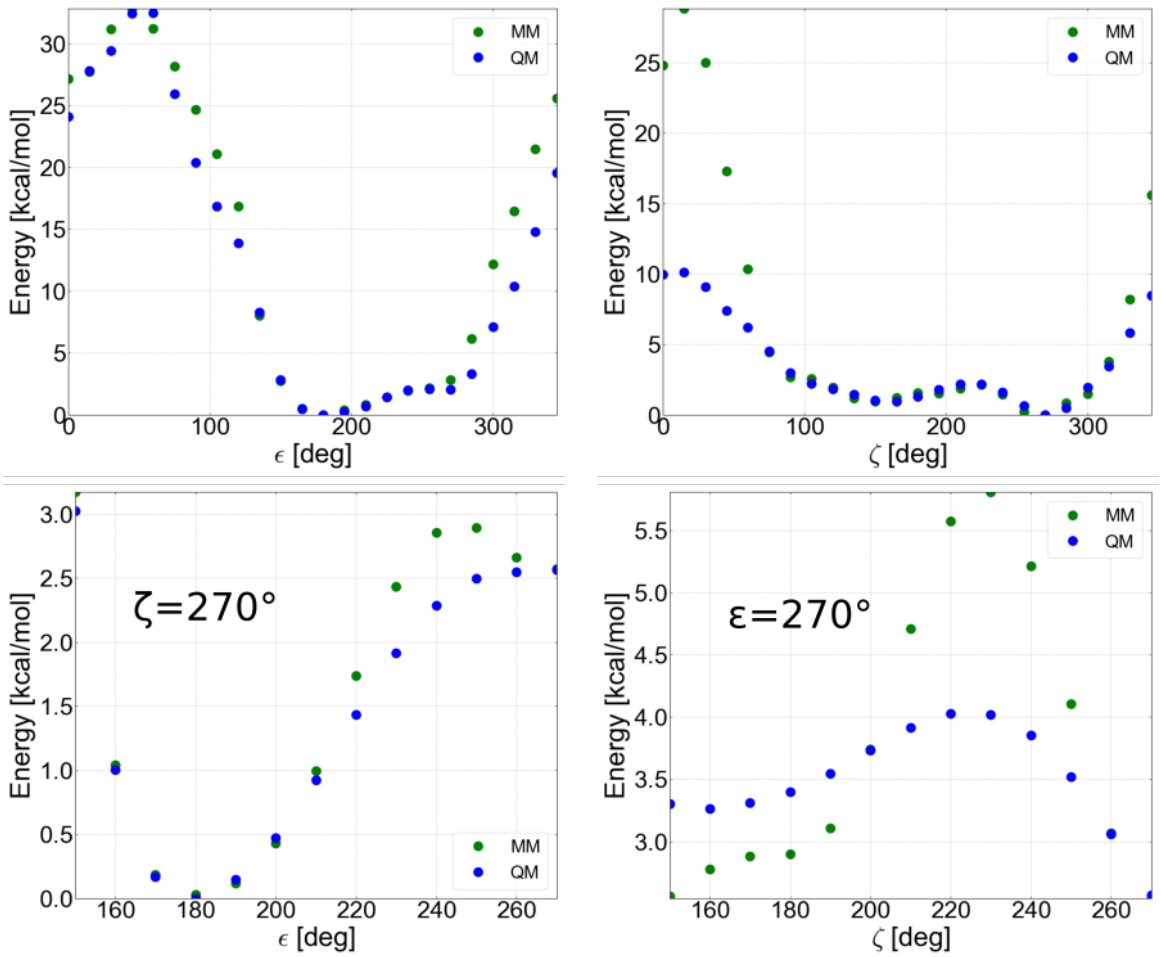


Figure 8.5: QM and fitted MM profiles for the $\epsilon-$ and ζ-dihedrals. Upper panels represent relaxed dihedral-scans, lower plots scans with ζ or ϵ constrained to 270°.

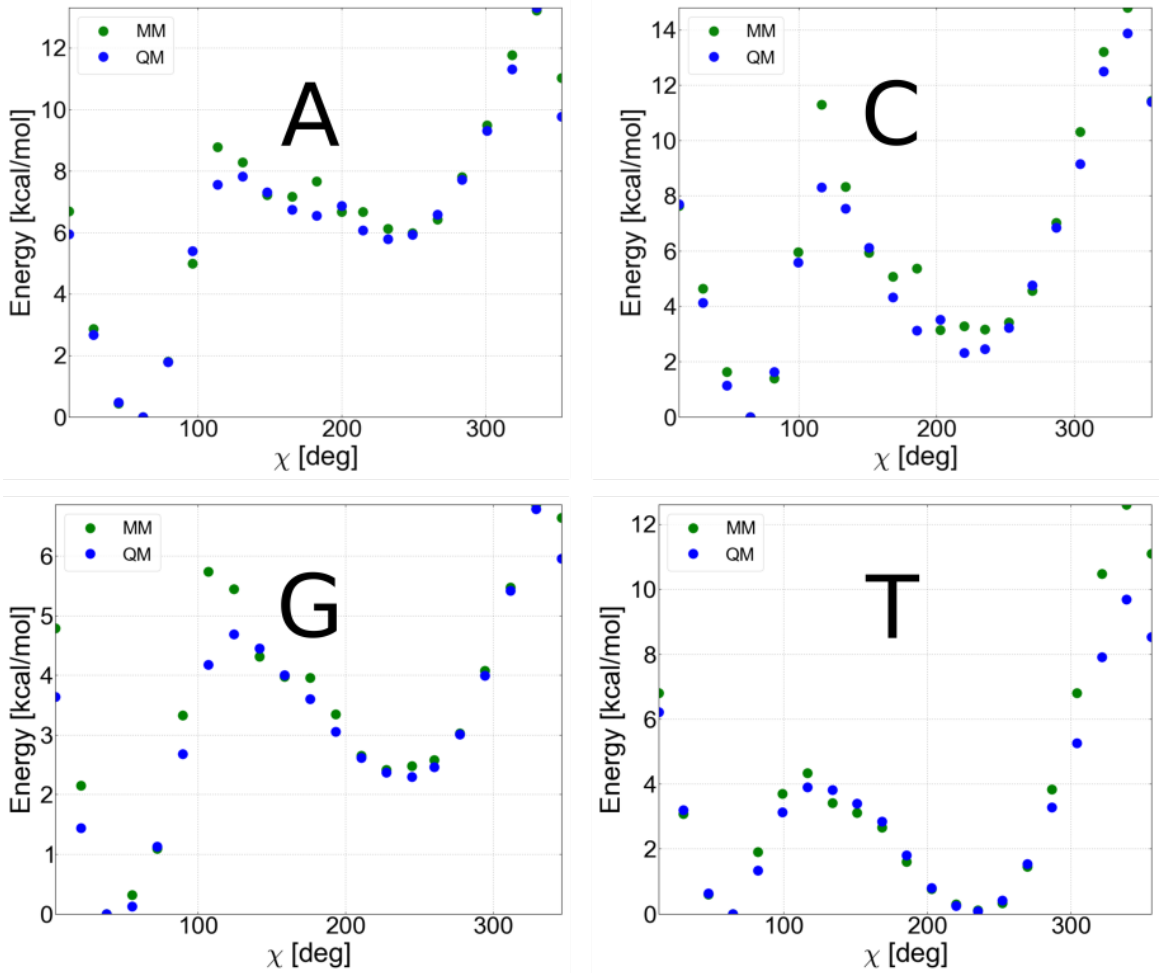


Figure 8.6: QM and fitted MM profiles for the χ-dihedral of the four nucleosides. The letters inside the panels indicate the base. The energy landscapes are obtained from scans with the sugar pucker constrained to C2'-endo conformation.

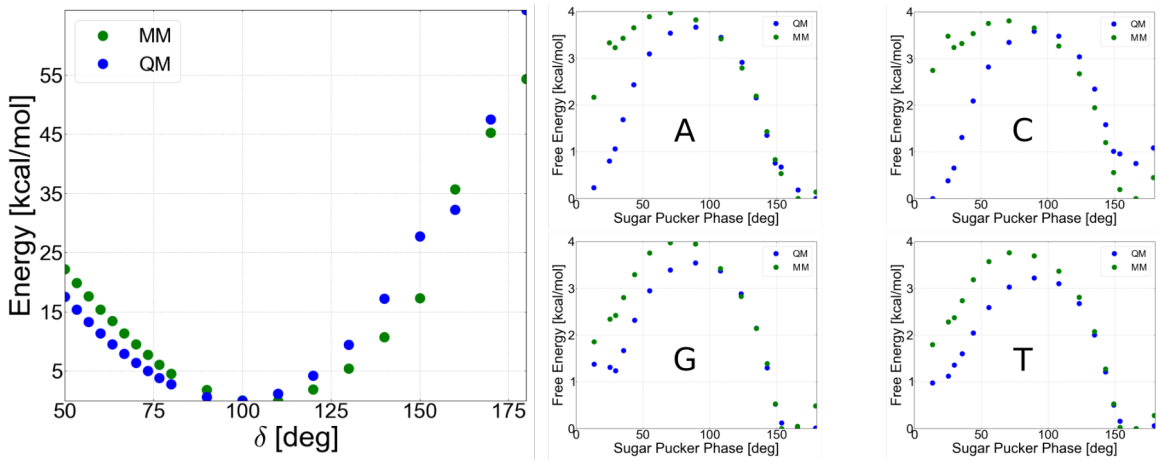


Figure 8.7: QM and fitted MM profiles for the δ-dihedral and the sugar pucker phase. Note that the δ-dihedral is strongly correlated to sugar puckering. The δ-dihedral (left panel) is therefore parameterized under constrained sugar pucker conditions (C2'-endo). Profiles for the sugar pucker phase angle are given in the right panels. Here, the previously derived $\chi-$ and δ-dihedrals have been used for the MM-description.

8.2.2 Simulation Results

B-DNA

Dickerson-Drew-Dodecamer We have checked the accuracy of the new Tumuc1 force field for several different DNA sequences.
The Dickerson-Drew-Dodecamer (DDD, d(CGCGAATTCGCG)) represents one of the most intensively studied DNA molecules [2, 5, 21, 35]. Note that the bsc1 force field is in large parts empirically parameterized to capture the DDD-structure accurately [21]. As shown in Fig 8.8 A, MD simulations with the Tumuc1 force field exhibit smaller deviations from the X-ray structure than with bsc1. This is reflected in an overall smaller rmsd, 1.52 $\mathring{A}$ for Tumuc1 and 1.69 $\mathring{A}$ for bsc1. Comparing the simulations to a NMR structure [36] yields substantially higher rmsds for both force fields (Fig 8.8 B). In this case, the performance of Tumuc1 and bsc1 is indistinguishable (rmsd of 3.42 vs 3.43 $\mathring{A}$). In a next step, we have checked the robustness of the Tumuc1 force field with respect to solvation. Upon increase of salt concentration (from neutralization only to 500mM NaCl) as well as change of the water model (TIP3P to OPC) [37, 38], Tumuc1 remains in excellent agreement with the X-ray structure. For the OPC water model, the accuracy apparently improves. The average rmsd in this case is 1.41$\mathring{A}$, for increased salt concentration it is 1.49$\mathring{A}$. The excellent agreement between simulations with the Tumuc1 force field and X-ray structure is visualized in Fig 8.9. Here, the average structure reveals stacking- and hydrogen bonding-patterns and backbone-geometries in close agreement with the X-ray structure. We have quantified all internal parameters of the DDD-molecule, and averaged them over the sequence. All parameters are summarized

in Table 8.1, whereby we draw comparison to the bsc1 force field as well as X-ray and NMR measurements [21, 35, 36]. With respect to base-pair, base-pair step and helical parameters, both the Tumuc1 and bsc1 force field are in good agreement with the experiments. For these parameters, far-reaching conclusions on the quality of the force fields are questionable due to the large differences between NMR- and X-ray measurements. Nevertheless, the Tumuc1 force field at least does not perform worse here. Concerning the dihedral angles, there is good consensus between NMR- and X-ray measurements. Thereby, the Tumuc1 force field performs clearly better than bsc1: For the α, δ, ϵ and ζ dihedrals, bsc1 deviates from experimental values by up to $\sim 15°$. Note that this has already been reported only shortly after the release of the bsc1 force field [5]. Average dihedral angles as captured by the Tumuc1 force field are generally in closer agreement to the experimental values. In addition, we have also computed elastic constants from the MD simulations (Table 8.2). Both force fields show good agreement with experimental values [39], except the stretch modulus, which is strongly overestimated by both force fields. However, overestimation of the stretch modulus by MD has been shown by the Noy group to be a length scale effect [40, 41]. For longer DNA sequences, existing force fields reproduce accurate stretch moduli. Calculating the stretch modulus as end-to-end distance (four base-pair steps longer than the previous scheme), yields already a significantly lower stretch modulus of 1854.0 pN with the Tumuc1 force field. As terminal base-pair fraying/breathing is very infrequent in the Tumuc1 force field ($4,0\%$, bsc1: 49.6%), this represents a reliable estimate. The substantially higher integrity of terminal base-pairs in the Tumuc1 force field is probably a result of the electrostatics, which reproduces a stronger (in vacuo) base-pairing. In these calculations, we have defined base-pairs to be frayed for a rmsd higher than 2.0 $\mathring{A}$.
Furthermore, we have verified the reproduction of local parameters along the DDD-sequence (Fig 8.10). For twist, roll and slide parameters, the Tumuc1 force field shows good agreement with X-ray data: Twisting around the central ApT step is in quantitative agreement with experiments. The roll angle is generally slightly underestimated, but the qualitative trend along the sequence matches with the X-ray profile quite well. The slide variable is underestimated by the bsc1 and the Tumuc1 force field. Nevertheless, the sequence dependence is captured accurately in both cases. Moreover, the Tumuc1 force field reproduces the sequence-dependent population of BI/BII states very well: In the central AT-segment, the backbone strongly prefers the population of the BI space, whereas CG-segments have a higher propensity for BII states. We have also investigated the distribution of all backbone dihedrals and sugar puckering to check the force field for possible artifacts. Importantly, the Tumuc1 force field does not populate non-canonical backbone states (see blue curves in Fig 8.11). In contrast, simulations with the bsc1 force field reveal sampling of a $60-70°$ state for the β-dihedral which is correlated with the γ dihedral flipping to $\sim 175°$. We are convinced that this is a serious force field artifact: First, the existence of this state in naked B-DNA is not supported by database research [42]. Second, it is not sampled in any other force field (bsc0, OL15, Tumuc1) [5]. Third, the β-dihedral has never been parameterized for bsc1. The overall population of this substate with bsc1 is $3-4\%$. However, here we have considered means over both backbone-strands along the entire DNA. For a 12 base-pair sequence, this means that

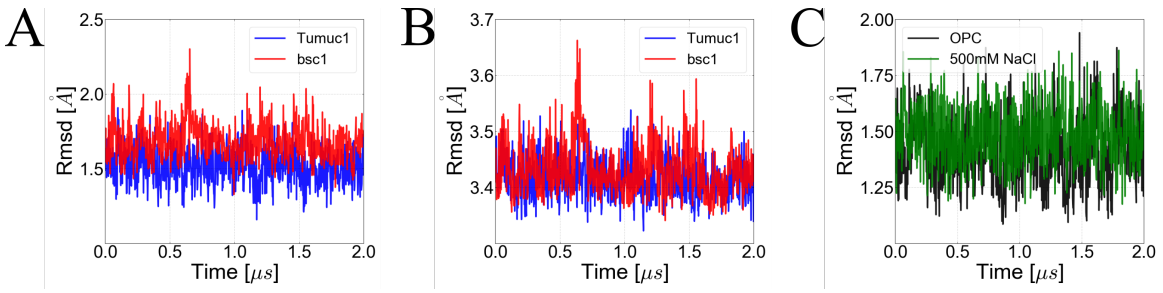


Figure 8.8: Simulations for the Dickerson-Drew-Dodecamer. (A) Rmsd-curves with respect to the X-ray structure (pdb:1BNA) for Tumuc1 and bsc1 simulations. (B) Rmsd-curves with respect to the NMR-structure (pdb:1NAJ). (C) Rmsd-curves of MD simulations with the Tumuc1 force field under different solvent conditions: OPC water model (black) and TIP3P water model at 500mM NaCl (green). The X-ray structure has been used as reference.

this artifact occurs very frequently at some position. Population of this artifact is also sequence-dependent, hence it is likely to be sampled more often in other sequences (e.g. in the second base-pair step of the Watson strand (GpC) its population amounts $\sim 15\%$). We notice that it occurs more frequently in C/G segments. This artifact has strong consequences for the local structure: The sugar ring is flipped which distorts base-pairing and base-stacking (Fig 8.11 D). In addition, it opens the minor groove widely (Fig 8.11 E), indicating fundamental problems when studying DNA-protein interactions. Thus, the structural integrity achieved with the Tumuc1 force field provides a huge advantage over bsc1. Apart from that, both force fields show reliable sugar puckering, whereby sampling of the Tumuc1 force field is more strongly confined to the C2'-endo state.

A-tract Conundrum A- and T-tracts display unique structural properties. The A-tract d(5'-AAAATTTT-3') is bent towards the minor groove (due to a high negative roll at the central ApT step), whereas the T-tract d(5'-TTTTAAAA-3') is bent towards the major groove with a high positive roll at the TpA step [46, 47]. MD simulations of the A-tract sequence reveal that the bsc1 and the Tumuc1 force field reproduce the sequence-dependent structure qualitatively correct (Fig 8.12 A-C). However, the central feature of this sequence is not covered very well. Both force fields underestimate bending at the ApT step remarkably. The rmsd-curves indicate equal performance of bsc1 and Tumuc1. In both cases, the DNA structure remains intact, for Tumuc1, short-lived plateaus at an rmsd of $\sim 3.8\mathring{A}$ are apparent (Fig 8.12 D). These plateaus correspond to slightly slided conformations of the terminal base-pairs, and are absent when the rmsd is calculated over the central segment (Fig 8.12 E,F). Note that we did not impose any restraints on the terminal base-pairs during our simulations.
For the T-tract sequence, we find similar performances. Twist-, roll- and slide-profiles are in qualitative agreement with the NMR-structure, but bending at the central base-pair step is again underestimated (Fig 8.13 A-C). Here, the rmsd-curves show a more reliable description for the Tumuc1 force field (Fig 8.13 D). After ~ 700 ns, the dsDNA

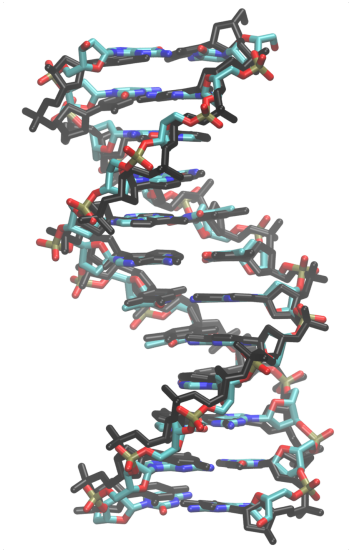


Figure 8.9: Average DDD-structure of the Tumuc1 simulation (colored) superposed to the X-ray structure (black). The average structure of the MD-trajectory has been computed as the Cartesian average with VMD [43].

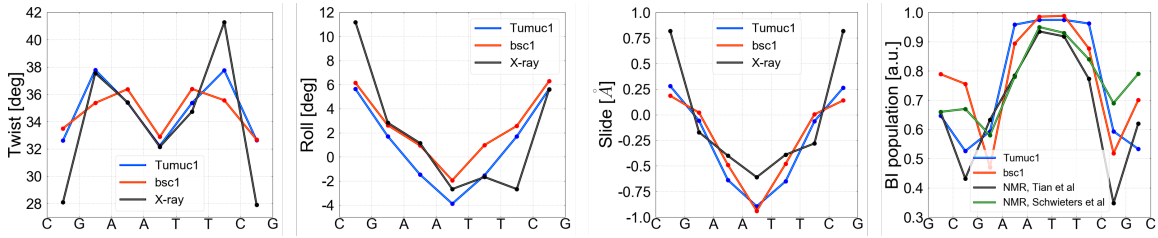


Figure 8.10: Profile of DNA parameters as a function of the DDD-sequence. Twist-, roll- and slide-profile for Tumuc1- (blue) and bsc1-simulations (red) and the X-ray structure (black). The right panel shows the BI population along the sequence for Tumuc1 (blue), bsc1 (red), NMR measurements from [44] (green) and NMR measurements from [45] (black). A higher BI population of the central AT-segment compared to flanking CG-sites is correctly sampled with bsc1 and Tumuc1.

	Tumuc1	bsc1	X-ray	NMR
shear [$\AA$]	0.00 ± 0.09	0.00 ± 0.10	0.01	0.00
stretch [$\AA$]	-0.04 ± 0.04	0.03 ± 0.04	-0.05	-0.33
stagger [$\AA$]	0.07 ± 0.14	0.07 ± 0.15	0.12	-0.16
buckle [deg]	-0.04 ± 4.58	-0.09 ± 4.40	1.09	-0.13
propeller [deg]	-10.32 ± 3.12	-10.30 ± 2.93	-13.19	-19.56
opening [deg]	2.11 ± 1.38	1.50 ± 1.69	3.45	2.33
shift [$\AA$]	0.00 ± 0.16	0.02 ± 0.19	0.01	-0.01
slide [$\AA$]	-0.25 ± 0.23	-0.22 ± 0.24	-0.03	-0.07
rise [$\AA$]	3.31 ± 0.07	3.30 ± 0.07	3.23	3.18
tilt [deg]	0.00 ± 1.05	0.15 ± 1.17	-0.07	0.01
roll [deg]	1.11 ± 2.16	2.52 ± 1.85	1.97	3.61
twist [deg]	34.83 ± 1.26	34.68 ± 1.40	33.88	35.73
x-displacement [$\AA$]	-0.54 ± 0.40	-0.78 ± 0.42	-0.15	-0.77
y-displacement [$\AA$]	0.00 ± 0.19	-0.04 ± 0.23	0.04	0.03
helical rise [$\AA$]	3.32 ± 0.08	3.31 ± 0.08	3.23	3.19
inclination [deg]	1.15 ± 2.87	3.80 ± 2.79	0.66	5.29
tip [deg]	-0.01 ± 1.53	-0.28 ± 1.75	-0.49	-0.11
helical twist [deg]	35.04 ± 1.25	34.93 ± 1.37	34.04	35.98
major groove width [$\AA$]	12.28 ± 0.75	11.86 ± 1.03	11.52	11.69
minor groove width [$\AA$]	5.19 ± 0.58	5.66 ± 0.59	4.42	4.70
sugar pucker phase [deg]	140.48 ± 22.19	114.36 ± 32.63	127.91	135.46
$\alpha[deg]$	292.87 ± 4.53	286.05 ± 7.99	298.64	301.18
$\beta[deg]$	168.57 ± 7.05	165.34 ± 8.11	174.16	170.49
$\gamma[deg]$	57.18 ± 2.42	54.56 ± 11.71	54.12	49.26
$\delta[deg]$	128.29 ± 4.14	134.26 ± 5.36	122.51	125.24
$\epsilon[deg]$	194.55 ± 7.50	201.08 ± 7.08	186.07	188.97
$\zeta[deg]$	248.57 ± 7.76	245.06 ± 8.60	256.36	257.21
$\chi[deg]$	252.63 ± 3.88	247.46 ± 5.17	246.11	249.84

Table 8.1: DNA parameters as obtained for DDD-simulations with Tumuc1 and bsc1 as well as X-ray and NMR measurements. Values for Tumuc1 and bsc1 have been computed from the $2\,\mu s$ MD simulations. The two terminal base-pairs of both ends have been omitted. The values represent averages over the DNA-structure and simulation time. Error estimates are standard deviations.

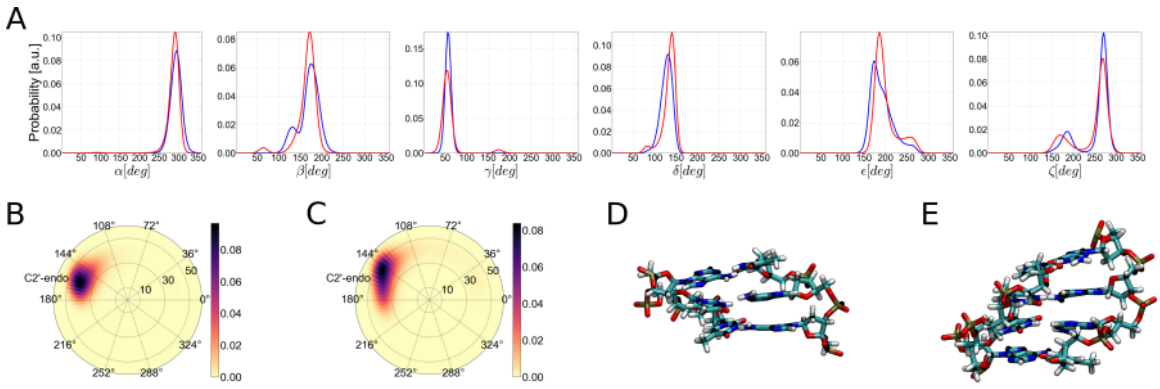


Figure 8.11: Population of backbone parameters sampled with bsc1 and Tumuc1 for the Dickerson-Drew-Dodecamer (last two base-pairs of both termini are excluded). (A) Probability-distributions shown for the backbone-dihedrals as obtained with bsc1 (red curves) and Tumuc1 (blue curves). (B) Sugar puckering of Tumuc1. Radial-component is the puckering amplitude and the azimuthal component is the phase-angle. (C) Sugar puckering of bsc1. (D,E) $\beta-$artifact in the bsc1 simulation. Transition to $\beta \sim 70°$ causes flipping of the γ-dihedral and an inclined position of the sugar ring (see central Adenine-nucleotide in (D). This also results in a strong local opening of the minor groove (E).

	Tumuc1	bsc1	experimental
Twist persistence length [nm]	97.6	78.6	109 ± 4
Bending persistence length [nm]	47.3 ± 8.4	43.9 ± 6.2	45 ± 2
Stretch modulus [pN]	3021.7	3045.5	1000 ± 200
Twist-stretch coupling $[\mathring{A}/deg]$	0.0094 ± 0.0002	0.0149 ± 0.0004	0.014 ± 0.003

Table 8.2: Elastic parameters calculated for the Dickerson-Drew-Dodecamer over the same segment as in Table 8.1. Experimental values are taken from [39]. Errors of the bending persistence length and twist-stretch coupling computed for Tumuc1 and bsc1 are standard deviations of the interpolation procedure. Stretch modulus is overestimated in the simulations due to length-scale effects [40, 41]. Calculating stretch modulus over the entire DNA-molecule (end-to-end) yields an apparently lower stretch modulus of 1854 pN for Tumuc1.

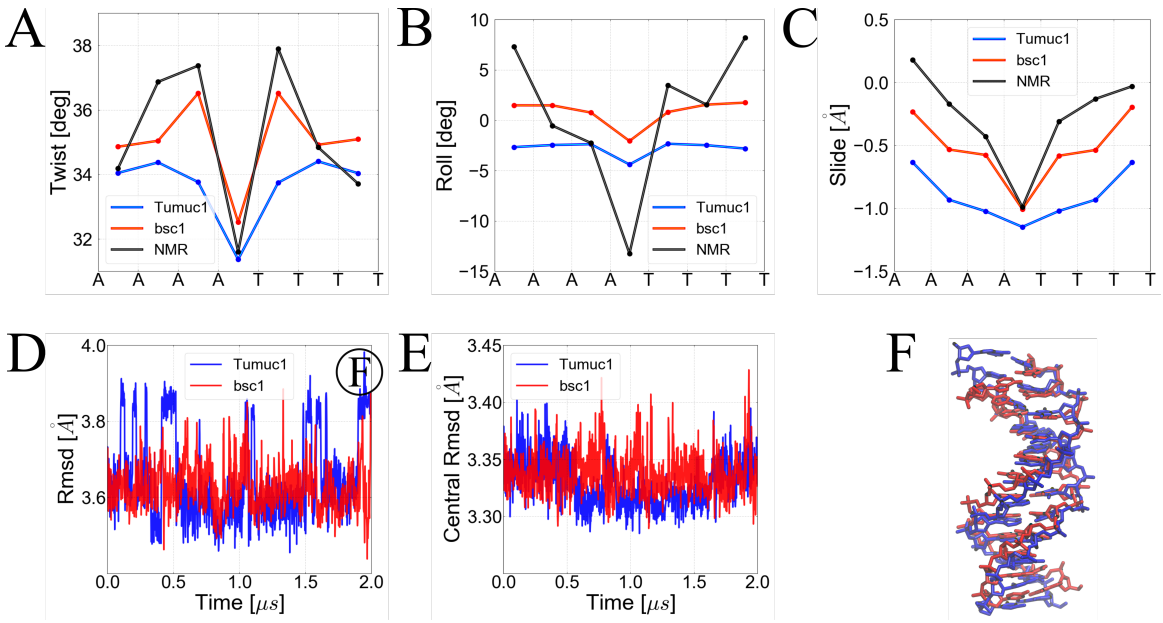


Figure 8.12: MD simulations of the A-tract sequence in comparison to the NMR-structure (pdb:1RVH) [46]. (A) Twist-profile along the central sequence for Tumuc1 (blue), bsc1 (red) and NMR (black). (B,C) Roll- and slide-parameters along the central sequence. (D) Rmsd-curves for Tumuc1 (blue) and bsc1 (red). Small plateaus at $\sim 3.85\,\mathring{A}$ are sampled in the Tumuc1-simulation. (E) Rmsd-curves excluding the three terminal base-pairs of both ends. No clear plateaus are visible in the Tumuc1-simulation. (F) Small plateaus for Tumuc1 at $\sim 3.85\,\mathring{A}$ correspond to sliding motions (blue). These deformations occur at the termini. Note that no restraints have been applied during the simulation. The superposed NMR-structure is shown in red.

helix breaks apart in the bsc1-simulation. These unstacking and flipping events are caused by a γ-flip in a central base-pair step (Fig 8.13 E,F). This broken conformation is stable for $\sim 20\,\text{ns}$ and reveals further problems in the bsc1 force field. Sampling of such strong distortions is not reasonable, as base-flipping occurs on the millisecond time-scale. We suppose that this is a force field artifact, and importantly Tumuc1 maintains the structural integrity substantially better (Fig 8.13 E).

Heterogeneous, 50 Base-Pair long Sequence We have performed a $2\,\mu s$ long MD simulation for the heterogeneous 50 base-pair long sequence (studied in chapter 6, [13]) with the Tumuc1 force field (Fig 8.14). Here, we record an excellent α/γ stability of 100% population of the g_-/g_+ basin [48], an accurate BI/BII balance (74%/26%) [49] and dominant population of the C2'-endo state (76%) [50]. These data demonstrate that DNA's structural integrity is conserved for various sequence contexts and excludes the scenario of small possible inaccuracies accumulating to severe artifacts in longer

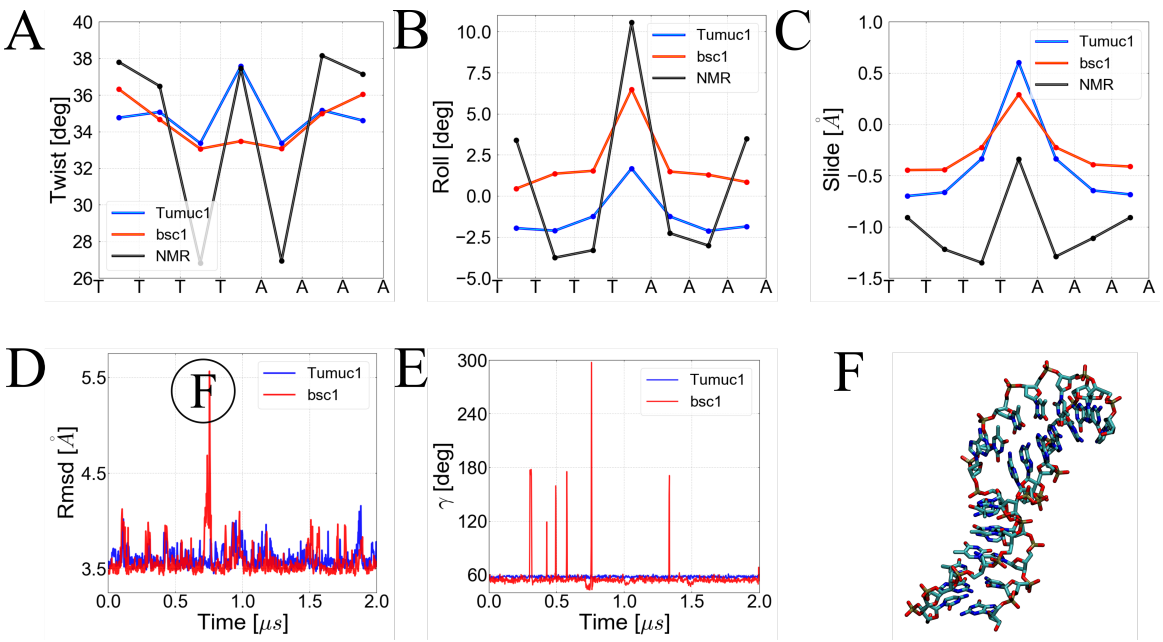


Figure 8.13: MD simulations of the T-tract sequence in comparison to the NMR-structure (pdb:1RVI) [46]. (A-C) Twist-, roll- and slide-profiles, similar as in Fig 8.12. (D) Rmsd-curves for Tumuc1 (blue) and bsc1 (red). (E) γ-dihedral in the Watson-strand of the central base-pair step. Full stability is maintained in the Tumuc1 force field. Flips in the bsc1-simulation reveal artifacts. (F) Snapshot of an artifact in the bsc1-simulation (at $\sim 750\,\mathrm{ns}$), caused by a γ-flip to $\sim 300°$. Complete rupture of the DNA double-helix and subsequent base-flippings are observed.

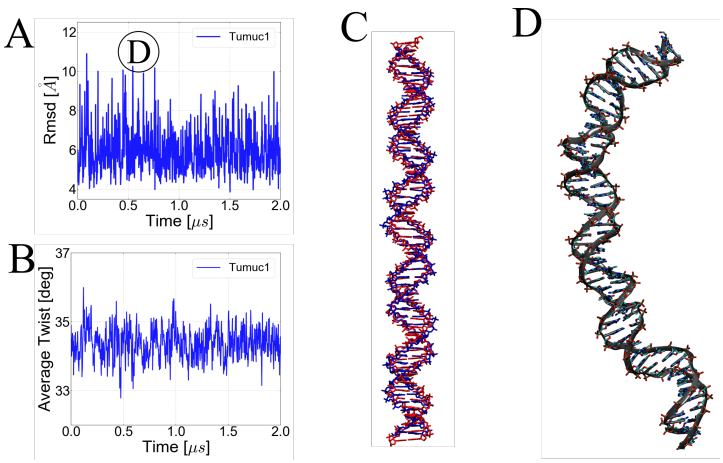


Figure 8.14: Simulation of a heterogeneous 50 base-pair long DNA sequence with the Tumuc1 force field. (A) Rmsd-curve. Larger values are associated with strong global bending. (B) Average twist over the central 43 base-pair steps as a function of time, revealing an accurate description of DNA's helicity by the Tumuc1 force field. (C) Average structure of the $2\,\mu s$ simulation (blue), generated in the same way as for Fig 8.9. The structure has been superposed to the experimentally expected structure built with the nab-module (red). The backbones and hence minor and major grooves align very well. Note that the average structure appears slightly shorter ($\sim$ 2 base-pair steps) due to bending fluctuations. (D) Snapshot taken from the simulation, the backbones are highlighted in cartoon representation. The DNA transiently undergoes strong global bending in the simulation with the Tumuc1 force field.

sequences. Larger rmsd values correspond to strong global bending, structural irregularities have not been observed. Furthermore, the Tumuc1 force field yields an excellent helicity: The average twist amounts to $34,4°$, which is equivalent to experimental values for heterogeneous sequences [51, 52]. Overlaying the average Cartesian structure from the MD simulation with the crystal structure shows a very good description of the backbone as well as major and minor groove. Note that the average structure appears slightly shorter than the straight crystal structure ($\sim$ 2 base-pair steps) due to the bending motions.

Thus, we conclude that the Tumuc1 force field represents a highly robust description for B-DNA's local and global structure. The description of backbone-dihedrals is apparently better than for bsc1. While the Tumuc1 force field does not capture all sequence-dependent parameters quantitatively correct, the qualitative behavior is generally accurate. This already represents a good success, as our parameterization procedure does not include any sequence-specific empirical modeling. We rather have built an atomistic, physical description for DNA based on quantum mechanical calculations, and this description widely agrees with experimental data on B-DNA.

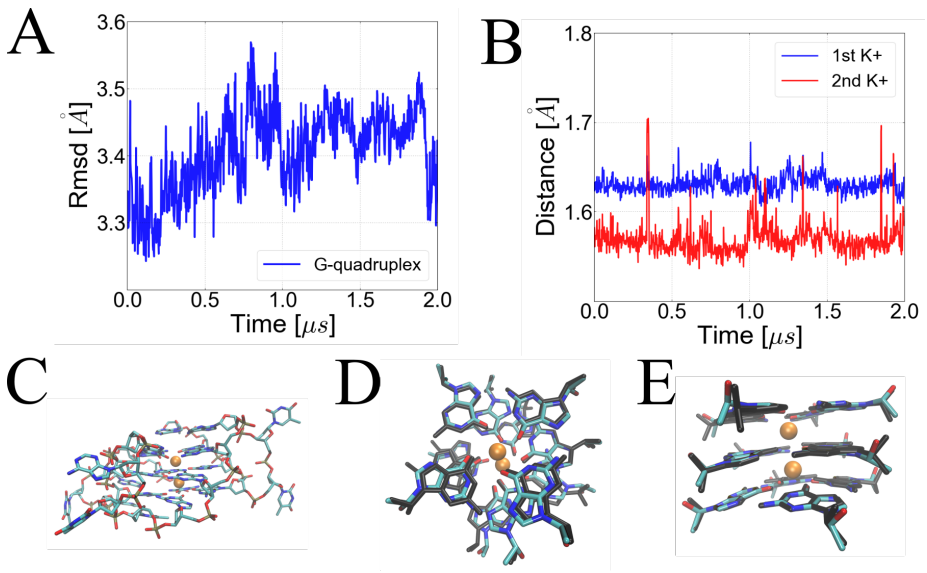


Figure 8.15: 2 μs MD simulation of a G4-quadruplex. (A) Rmsd-curve computed for the 12 guanine bases in reference to the crystal structure (pdb:1KF1) [57]. (B) Average distances between the potassium ions and interacting O6-guanine atoms. Both ions stay in between the two quadruplex-layers during the 2μ s-simulation. (C) Structure of the G4-quadruplex. Potassium ions are highlighted as orange spheres. (D,E) Average structure of the G4-layers (colored) superposed to the crystal structure (black) in top- and in side-view.

Stability of the G-Quadruplex Structure

Guanine quadruplex (G4) structures are often found in telomeric ends and are known to be of outmost biological relevance (e.g. changes of the telomeric state are associated with senescence) [53, 54]. G4 structures consist of multiple layers, whereby each layer is a quadratic arrangement of four guanine bases. The quadratic arrangement is formed through hydrogen bonding of the guanines via the Watson-Crick and Hoogsten edges. The overall structure is further stabilized by the capture of a cation between successive layers [54–57]. We have studied the stability of G4-structures in the Tumuc1 force field by an unrestrained 2 μs MD simulation (Fig 8.15). Throughout the simulation, the G4-structure remains stable, and, importantly, both potassium ions remain in between the quadratic layers. We have not observed distortions in the hydrogen bonding or stacking of the guanine bases during the whole simulation time. Indeed, the average structure of the guanine layers recorded from the Tumuc1 simulation is in good agreement with the crystal structure. We therefore state a sound description of guanine quadruplex structures with the Tumuc1 force field.

Hairpin Folding

As a next test-system, we have studied the 5'-GCGCAGC-3' DNA single-strand. This sequence forms a hairpin with the central cytosine-base in loop position, the juxtaposed G and A-base thereby form a mismatched base-pair [58]. We have carried out a $2\,\mu s$ MD simulation for this sequence with the Tumuc1 force field starting from an extended single-strand structure (Fig 8.16). After ~ 600 ns, the DNA strand folds into the correct hairpin-structure. For the rest of the simulation, the hairpin conformation remains stable. Intriguingly, the single-strand is folded into an intermediate state prior to the correct hairpin-structure. The intermediate state is characterized by correct loop formation but irregular hydrogen bonding at the terminal base-pair (Fig 8.16 D). This irregularity is caused by flipping of the 5'-Guanine base to syn-conformation ($\sim 60°$). Note that the syn-state of 5'-end Guanines is stabilized by an unconventional hydrogen-bond between the HO5' and the N3'-base atom [59, 60]. It is likely that such interactions are too attractive in current Amber force fields, hence overstabilizing the syn-state for 5'-terminal purines [60]. This imbalance can strongly impede structure formation of DNA. However, simulation of the DNA single-strand indicates that the Tumuc1 force field has a good balance between electrostatics and glycosidic dihedrals: The syn-state is visited twice for the 5'-terminal guanine, but the base is not caught in this state and undergoes transitions back to the native anti-state. Thus, we have been able to simulate correct hairpin folding in a standard, unrestrained MD simulation. To the best of our knowledge, hairpin folding has been reported so far only based on advanced replica exchange simulations [21, 58, 61].

Hybridization of two 5'-CGCG-3' Single-Strands

We have performed a $2\,\mu s$-long MD simulation for a solvated system which contains two 5'-CGCG-3' single-strands with the Tumuc1 force field (Fig 8.17). The two single-strands were placed $\sim 27\,\mathring{A}$ away from each other, and hence have not been in contact at the beginning of the simulation. The two strands then approach by stacking on top of each other (Fig 8.17 C). However, this complex does not remain stable and decays after ~ 30 ns. The two strands then contact each other by hydrogen bonding of the two 5'-terminal bases of both strands (Fig 8.17 D). After ~ 40 ns the two strands dissociate and adopt a complex characterized by a stronger hydrogen bonding network (Fig 8.17 E). Nevertheless, the two strands are aligned in parallel 5'-3' polarity. This complex is thermodynamically unstable and decays after a few nanoseconds. The two strands then approach each other in correct polarity, whereby the complementary bases pair accordingly, hence hybridizing the two strands into the correct duplex structure (Fig 8.17 F). This complex remains stable for the rest of the simulation. So far, hybridization of single-strands has not been introduced as a criterion to check the quality of the force fields. However, results with the Tumuc1 force fields are encouraging: Not only do we obtain the precise duplex structure, but also is the force field capable of overcoming meta-stable states. This indicates good kinetic behavior of the Tumuc1 force field.

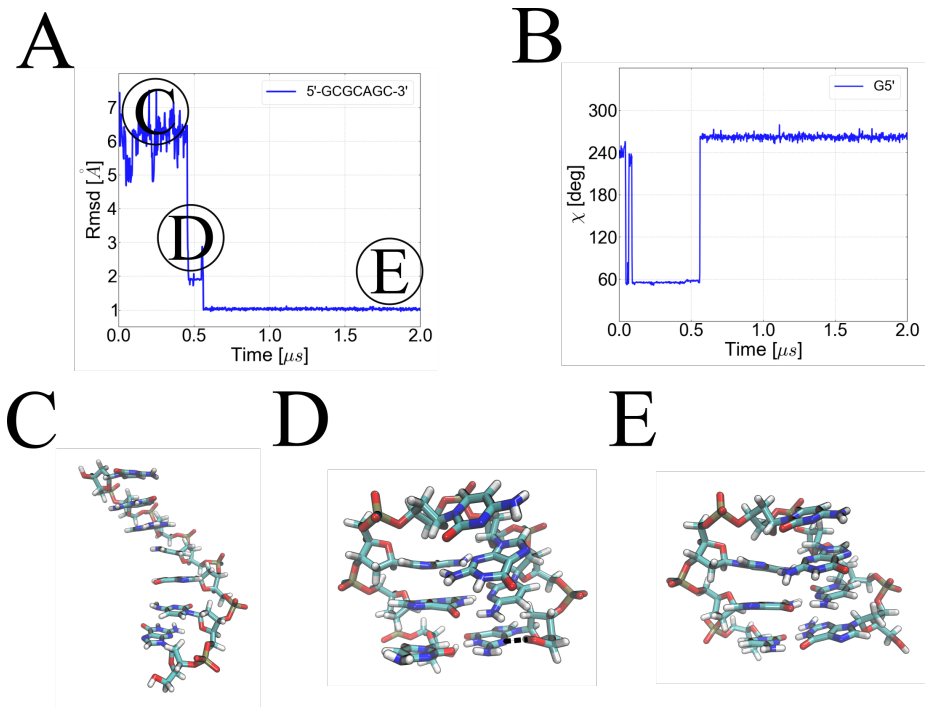


Figure 8.16: Folding of the 5'-GCGCAGC-3' single-strand. (A) Rmsd-curve with respect to the crystal structure shows folding after $\sim 600\,\text{ns}$. (B) χ-dihedral of the 5'-guanine base as a function of time. Transitions from anti- to syn-conformation and vice versa are observed. An accurate balance between both conformations is important for single-strand folding. (C) The simulation was started from an extended single-strand structure. (D) The hairpin-loop is correctly folded, but the 5'-guanine is in syn-conformation ($\chi \sim 60°$). This conformation is stabilized by an unconventional HO5'-N3 hydrogen-bond (highlighted by black dots), and prevents correct base-pairing and base-pair stacking at the termini. (E) Structure of the single-strand folded into the correct hairpin-conformation.

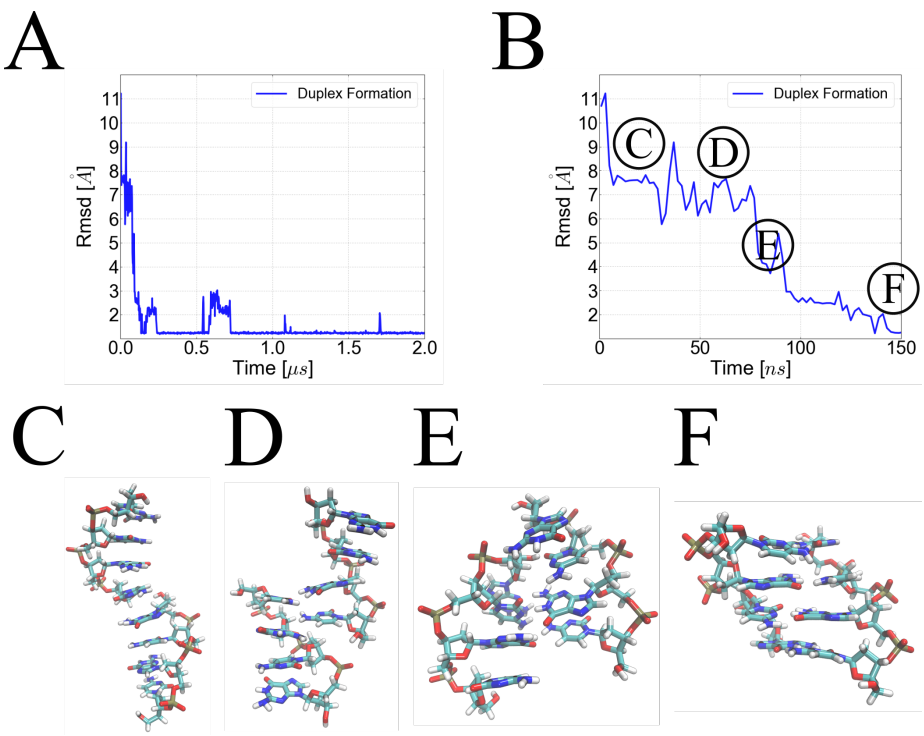


Figure 8.17: Hybridization of two 5'-CGCG-3' single-strands simulated with the Tumuc1 force field. (A,B) Rmsd-curves with respect to the duplex structure shown for the entire simulation time (A) and for the first 150 ns (B). (C) At the beginning of the trajectory, the two single-strands go into a stacked conformation. (D) Meta-stable state. Hydrogen bonding between terminal bases occurs. (E) Conformation characterized by stronger hydrogen bonding, but parallel strand polarity. (F) After $\sim$ 150 ns, the two single-strands have hybridized to the correct duplex structure.

Formation of Duplex Structure

Tumuc1's potential to capture formation of DNA duplex structures has been further investigated in ten $\sim 1.2\,\mu s$ MD simulations of the d(5'-CGTTGTTGG-3') DNA sequence. The ten simulations were started from the same structure, in which the first five base-pairs are in B-DNA configuration and the terminal four base-pairs are dissociated. Note that each simulation was run with different starting velocities, hence resulting in different trajectories. Thereby, the DNA-structure has folded correctly in five simulations, and is trapped in non-native conformations in the other five simulations (Fig 8.18). For the refolding of the DNA duplex structure, we thus find a success-rate of 50%, which is likely to increase upon extension of the simulation time.

Z-DNA

Finally, we have tested the Tumuc1 force field for Z-DNA. Thereby, we have simulated a Z-DNA hexamer [62] at 2M NaCl for $2\,\mu s$ with the Tumuc1 force field and comparatively with the bsc1 force field (Fig 8.19). Measured over the entire DNA-structure, the Tumuc1 force field shows a substantially higher rmsd, which can be explained by strong fraying of the terminal base-pairs. Excluding the terminal base-pairs results in a slightly lower rmsd compared to the bsc1-simulation. Intriguingly, we have observed only modest fraying in the bsc1 force field, revealing a reversed trend to the B-DNA simulations, where terminal base-pair fraying is significantly more pronounced for bsc1 than for Tumuc1. Here, information from the experimental side is scarce and challenging to compare to MD simulations due to the relatively short simulation times [5]. With respect to the helicity, however, the Tumuc1 force field performs better as it maintains Z-DNA's left-handed structure very well and reproduces base-pair stacking appropriately. The bsc1 force field, instead, overtwists the Z-DNA remarkably. Another advantage of the Tumuc1 force field is the accurate sampling of sugar puckering (Table 8.3): Guanine bases preferentially adopt the C3'-endo state, whereas cytosine bases exclusively occupy the C2'-endo state [62]. In the bsc1-simulations, all sugar rings adopt B-form puckering (around C2'-endo). In previous paragraphs, we have shown that the Tumuc1 force field captures B-DNA's backbone structure very accurately. For Z-DNA, in contrast, we find that the native g_+/t state is unstable. This is also largely the case for the bsc1 force field (Table 8.4). Z-DNA is known as a challenging system for force fields, and a reliable investigation of these structures with non-polarizable force fields has even been questioned [5, 20]. The performance of both, the bsc1 and the Tumuc1 force field is not optimal. Nonetheless, better reproduction of left-handed helicity might be a clear bonus for the Tumuc1 force field, as unwinding of double-stranded B-DNA can be absorbed by transition of G/C rich segments to Z-form [63]. For such phenomena, an accurate description of twisting is hence important.

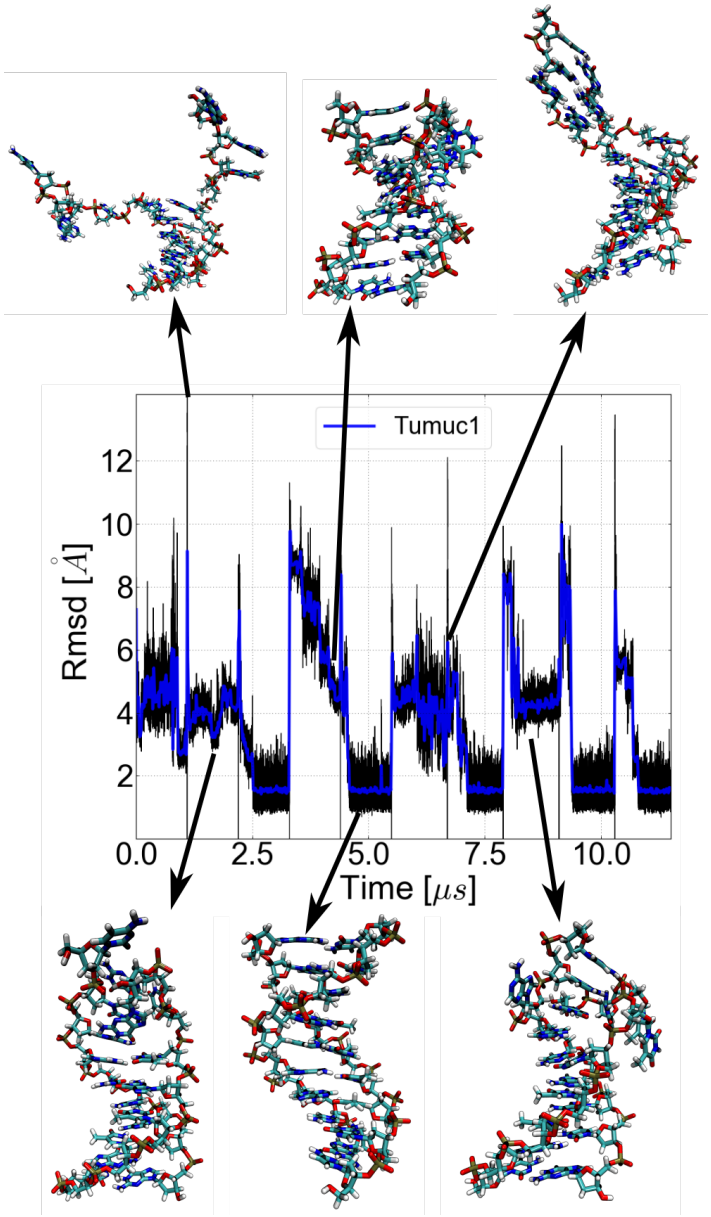


Figure 8.18: Refolding of the partially melted DNA duplex. The rmsd with respect to the correct B-DNA structure is shown in black and blue (binned in 8 ns). Restarts of the simulation are highlighted by lines to a rmsd of 0 $\AA$. From ten simulation runs, five folded into the correct structure, shown by a rmsd of $\sim 1.5\,\AA$.

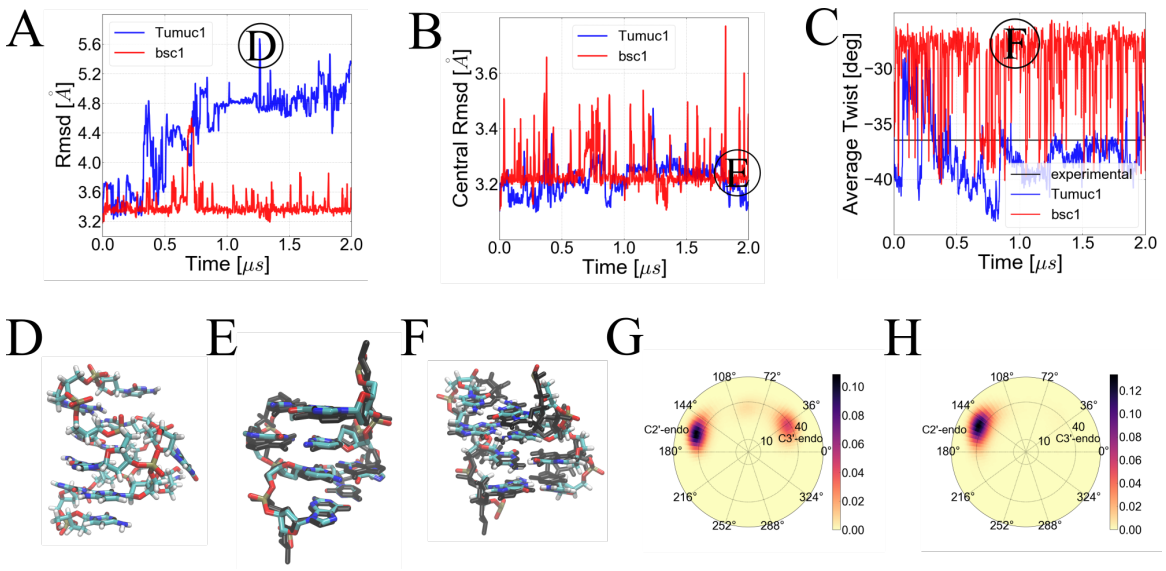


Figure 8.19: MD simulations of the Z-DNA hexamer (pdb:1ICK) with Tumuc1 and bsc1. (A) Higher rmsd-values for Tumuc1 are caused by terminal base-pair fraying. (B) For the central base-pairs, Tumuc1 shows on average a slightly lower rmsd. (C) Twist averaged over the three central base-pair steps. We report better agreement for Tumuc1 with the crystal structure than for bsc1, which overtwists Z-DNA's left-handed structure. (D) Snapshot taken from the Tumuc1 simulation. The two terminal guanine-bases are flipped out. (E) Snapshot of the four central base-pairs taken from the Tumuc1 simulation after $2\,\mu s$ (shown in color) superposed to the crystal structure (black). (F) Snapshot for the four central base-pairs taken from the bsc1 simulation (shown in color). Superposition to the crystal structure (black) illustrates reduced left-handedness. (G) Sugar puckering in the Tumuc1 simulation, the C3'-endo state is also sampled. (H) Sugar puckering in the bsc1 simulation, showing sampling only at B-DNA configuration (around C2'-endo).

	cytosine		guanine	
	C2'-endo [%]	C3'-endo [%]	C2'-endo [%]	C3'-endo [%]
Tumuc1	85.5	0	21.7	33.3
bsc1	58.5	0	69.7	1.7
experimental	100	0	33.3	50.0

Table 8.3: Sugar puckering for the cytosine and guanine bases in Z-DNA as obtained by the MD simulations and from the crystal structure [62]. Here, we have considered only the two most prominent pucker conformations.

	g_-/g_+[%]	t/g_+[%]	g_+/g_+[%]	g_+/t[%]
Tumuc1	38	55	6	0
bsc1	14	68	0	17
experimental	0	40	0	60

Table 8.4: Population of α/γ-substates recorded from MD simulations and the crystal structure [62].

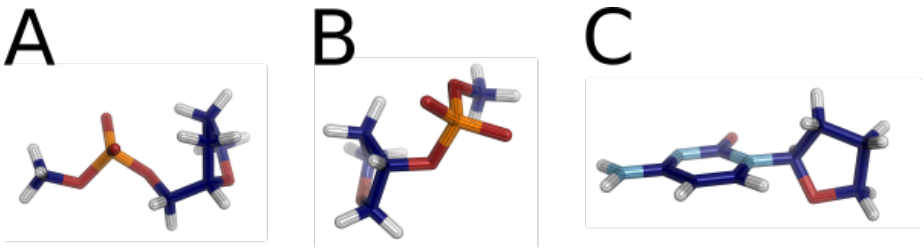


Figure 8.20: Model-systems for the parameterization of bonds and angles. (A,B) Model-systems used to parameterize the backbone and sugar ring. (C) Model-system for the derivation of base and base-to-sugar parameters.

8.3 Materials and Methods

8.3.1 Parameterization

Bonds and Angles

Bond- and angle-parameters for the Tumuc1 force field have been determined based on QM calculations and the modified Seminario method [24]. For this purpose, QM-optimization and frequency-calculations have been performed on the model-systems shown in Fig 8.20. The cc-pVTZ basis set [64, 65] and B3LYP exchange correlation functional [66–68] in combination with the empirical D3 version of the Grimme dispersion-correction was used [69]. These QM calculations were carried out with the Gaussian09 program [70]. From the QM calculations, bond- and angle-parameters have been extracted with the modified Seminario method. Thereby, the Matlab-script provided by Allen and coworkers was used and the scaling factor was set to 1.0 [24]. In cases where this method yielded slightly different parameters for equivalent atom-type combinations, we selected the average values for the Tumuc1 force field.

Charges

The model-system for calculation of the charges is shown in the Results section (Fig 8.1). This model-system is slightly larger than a whole nucleotide and was used for the four different bases: Adenine, cytosine, guanine and thymine. For each of the bases, a

RI-MP2 optimization was performed with the orca 4.2.0 software [31]. The def2-TZVP basis-set and the def2/JK auxiliary basis-set were used [29, 30]. Calculations were performed in implicit solvent (CPCM, $\epsilon = 80$) [28]. The output was processed with the Multiwfn 3.6 software to fit the partial charges to reproduce the electrostatic potential from the QM calculation [26]. Fitting was done by means of the Merz-Kollman scheme [23, 25], where a grid is spanned around the molecule. We spanned grids composed of 6 data points per $\mathring{A}^2$, and four layers were placed around each atom, whereby each layer was placed in multiples of 1.4, 1.6, 1.8 and 2.0 of the vdW-radii to the atoms. This resulted in overall $\sim$ 25000 data-points, for which the electrostatic potential was fitted (Fig 8.1). Furthermore, charge- and equivalence-constraints were employed. A charge of -1 was enforced for a nucleotide, and charges of equivalent atoms were constrained to have the same value. As a first step, charges were derived for the Adenine-system. For the remaining three molecules, charges on backbone and sugar atoms were constrained to the same values as in the Adenine-system. Finally, charges for the 5'- and 3'-hydroxyl groups were derived. This was done based on QM-calculations for a cytosine-nucleoside, whereby all non-hydroxyl-charges were constrained to the previously derived cytosine-, sugar- and backbone-charges.

QM Energy Profiles along the Dihedrals

For the parameterization of dihedral angles, QM-scans were performed for all relevant dihedrals with orca 4.2.0 [31]. The QM calculations are based on the RI-MP2/def2-TZVP, def2/JK level of theory [29, 30]. Thereby, we applied constraints on specific other dihedrals and angles to reduce the 'noise' in the MM-energy landscapes. In this way, the energetic contribution of the scanned dihedral angle could be identified more easily. Scanning of the dihedral angles was carried out by equally spaced samples over the corresponding intervals.

$\alpha/\beta/\gamma$ Parameters for the α, β and γ dihedral angles were derived by scanning each dihedral angle from 0° to 360° in 21 steps. Scanning was carried out for model-system A in Fig 8.3. Thereby, sugar puckering was constrained as well as the O5'-P-O3'-C3' dihedral angle. Note that the C3' atom represents a methyl-capping carbon, and we did not derive dihedral parameters therefore. This problem is circumvented by constraining it, as this dihedral angle does then not contribute to the relative MM energy-landscape. In addition, we conducted QM calculations for the γ dihedral at $40°, 160°$ and 280°, whereby the β-dihedral was constrained to 70° (the α-dihedral was unrestrained). These additional calculations were necessary to eradicate the β-artifact which is present in the bsc1 force field.

ϵ/ζ The ϵ and ζ dihedrals were scanned from 0° to 360° through 24 samples in each case. The calculations were performed for model-system B in Fig 8.3. Additional scans from 150° to 270° with 13 samples were conducted for each of the two dihedrals with one

of them being constrained to 270°. In this way, the BI and BII conformational spaces were extensively sampled. Puckering and the C5'-O5'-P-O3' dihedral were constrained in all calculations, for similar reasons as in the $\alpha/\beta/\gamma$ case: The C5' atom is a capping methyl-carbon. In addition, the P-O3'-C3', O3'-C3'-C2' and O3'-C3'-C4' angles were constrained. From previous, initial scans we have observed that these angles can undergo large changes, which may not be covered accurately by the uncoupled, harmonic form of our force field.

δ Parameters for the δ-dihedral were obtained from scanning of the cytosine-nucleoside with constraints on the χ-dihedral and sugar pucker. The scans comprise 10 samples in the range from 90° to 180°, and another 10 samples from 50° to 80°.

χ χ-dihedrals were scanned for the four nucleosides by 21 samples in the range from 0° to 360°. Sugar puckering was constrained to the C2'-endo state.

$SugarPucker$ The sugar pucker was scanned for all four nucleosides without any further constraints. Thereby, the sugar pucker phase angle was scanned from 0° to 180° with 16 samples for each nucleoside. Scanning of the sugar pucker was achieved by constraining the ν_1 and ν_3 dihedrals to corresponding values.

MM Energy Profiles along the Dihedrals

Fitting of the dihedral-parameters requires an MM-description (topology) for the model-systems. Topology-files for the model-systems were built with the antechamber module [27]. For model-systems A and B of Fig 8.3, we have derived charges similar to paragraph 8.3.1. Here, we considered multiple conformations along the scanned dihedral-profiles. Charges of the nucleoside systems were directly taken from the previously derived charges for the Tumuc1 force field. Analogously, also the same bond- and angle-parameters as in the Tumuc1 force field were used. The bond-, angle- and charge-parameters were incorporated into the topology-files with the parmed module of the Amber16 package [27]. We also used the parmed-module to set parameters to zero for the dihedrals which we fitted.

Fitting of the Dihedral-Parameters

Generation of MM Trajectories Coordinate-outputs of the QM-scans were transformed to Amber-compatible format with VMD [43]. Resulting structures were then energy-minimized with respect ot the previously established topologies. Minimization included 2500 steps of steepest-descent and the same constraints as in the QM calculations were applied. In this way, reliable MM trajectories for the dihedral-scans are obtained.

Fitting Procedure MM energy profiles along the dihedrals are derived from the MM trajectories and the constructed topologies. Dihedral parameters were now fitted to offset differences between the QM and MM energy landscapes. As discussed in the results section, dihedral-potentials are usually expanded to 4th order, and weighting functions were used to guarantee higher fitting quality for thermodynamically more relevant conformational spaces. Note that puckering is described by the same parameters for all four nucleosides, and energy-overestimation of the C3'-endo state has been allowed by at least 0.35 kcal/mol. Large energetic changes in the δ-dihedral are a consequence of the applied constraints. In order to avoid overfitting, boundaries have been applied to fitting of the δ-dihedral. Potential barriers were confined to intervals of (-2.25,2.25), (-1.75, 1.75) and (-1.25, 1.25) kcal/mol for the three orders (higher potential barriers are allowed for lower orders). Similarly, phase-angles were confined to $(70°, 170°)$, $(75°, 160°)$ and $(75°, 160°)$. Fitting was then carried out with the Basin-hopping technique [71]. For each model-system, we performed fitting at least 10 times, and then chose the parameter-set for the Tumuc1 force field, which yields the best agreement between MM and QM (i.e. the lowest value for the function f in equation 8.2).

8.3.2 MD Simulations

Starting structures for all systems were generated either from pdb-structures or with the nab-module of the Amber16 package (system dependent details are given in the following paragraphs) [27]. All systems were prepared with the xleap-module, where the Tumuc1 force field files can easily be accessed. The systems were neutralized by sodium ions and solvated with explicit solvent described by the TIP3P water model [37] if not stated otherwise. The distance between box-boundary and DNA was set to 10.0 $\mathring{A}$. After preparation, the systems were energy minimized, heated up and equilibrated for the NPT ensemble ($p = 1\,\text{bar}, T = 300\,\text{K}$) for overall 1 ns. The production runs lasted 2 μs, all simulations were performed with the pmemd.cuda module of the Amber16 package. If not explicitly mentioned, no restraints were applied in these simulations.

Dickerson-Drew-Dodecamer

The sequence of the Dickerson-Drew-Dodecamer is d(5'-CGCGAATTCGCG-3'). The crystal structure (pdb:1BNA) [35] served as starting structure for the simulations.

A- and T-tract

The 1rvh- and 1-rvi-pdb structures were used as input for simulations of the A- and T-tract, respectively [46].

Heterogeneous 50 Base-Pair Sequence

The same sequence as in chapter 6, [13] was studied, see Table 6.1 label AT. During the $2\,\mu s$ production run, positional restraints were applied on the first two base-pairs. The force constant was $0.1\,\frac{\text{kcal}}{\text{mol}\text{Å}^2}$.

Guanine-Quadruplex

The 1kf1 pdb-structure [57] was used as starting structure.

Hairpin Folding

The 5'-GCGCAGC-3' single-strand was studied. A B-form like extended single-strand was generated as a starting structure with the nab module [27].

Hybridization

Two 5'-CGCG-3' single-strand structures were generated with the nab-module [27]. The two single-strands were displaced manually using VMD (center of mass distance is $\sim 27\mathring{A}$) [43].

Formation of Duplex Structure

Refolding of the d(5'-CGTTGTTGG-3') DNA duplex structure has been simulated at a temperature of $T = 320\,\text{K}$. Positional restraints were applied on the first two base-pairs with a force constant of $k = 0.25\,\frac{\text{kcal}}{\text{mol}\text{Å}^2}$. Ten simulations were started from the same partially melted structure, but with different velocities. Hydrogen-mass-repartitioning was activated, which allowed a time step of 4 fs. Atom coordinates were written out every 80 ps. The total simulation time was $\sim 12\,\mu s$.

Z-DNA

The 1ick-pdb structure [62] was used. This DNA structure consists of six guanine-cytosine base-pairs. 2M NaCl were added.

8.3.3 Analysis of MD Trajectories

RMSD and DNA-Parameters

The rmsd is defined by:

$$rmsd = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{r}_i - \mathbf{r}_{i,eq}\|^2}. \tag{8.4}$$

Rmsd-curves of the MD trajectories were calculated with VMD [43]. All other internal parameters for the DNA molecules (base-pair and base-pair step parameters, dihedrals and sugar pucker) were computed with Curves+ [72].

Global Elasticities

For the Dickerson-Drew-Dodecamer, twist persistence length C, bending persistence length A, stretch modulus S and twist-stretch coupling D were determined based on the Tumuc1 and bsc1 force fields. Twist persistence length and stretch modulus were obtained as inverse of the variance, $C = \frac{L}{\mathrm{var}(\phi)}$ and $S = \frac{k_B T \cdot L}{\mathrm{var}(L)}$. ϕ denotes the total twist over the corresponding segment and L the contour-length measured as sum of the rise-parameters. Twist-stretch coupling D was extracted from linear interpolation between helical-twist and helical-rise. We used our rigid-body protocol introduced in chapter 6, to calculate the bending persistence length A. This allowed us to compute a mean axis for every base-pair. Thus, bending as a function of contour-length could be recorded. The bending persistence length is defined by [41]:

$$< \cos(\theta) >= e^{-\frac{L}{A}}. \tag{8.5}$$

From series expansion, $< \cos(\theta) >= 1 - < \frac{\theta^2}{2} >$ and $e^{-\frac{L}{A}} = 1 - \frac{L}{A}$, we find:

$$< \theta^2 >= \frac{2 \cdot L}{A}. \tag{8.6}$$

The bending persistence length A was then derived from linear interpolation of equation 8.6.

8.4 Conclusion

We have derived a new DNA force field ("Tumuc1") in a bottom-up approach. Charges, bonds, angles and dihedral angles have been parameterized based on high-level QM calculations. Charges have been fitted with the Merz-Kollman scheme [23, 25] using a high grid resolution, which results in a significantly different electrostatic landscape compared to the classical Amber force fields. We have employed the recently developed modified Seminario method [24] to compute bond- and angle-parameters. These techniques represent novelties for the development of force fields, classical Amber force fields

still rely on the charge-,bond- and angle-parameters established by Cornell et al in 1995 [22]. The Tumuc1 force field contains the same Lennard-Jones parameters as classical Amber force fields. The force field is completed by parameters for the critical dihedral angles, which we have obtained from scanning and fitting of the quantum mechanical energy landscapes. Parameter- and library-files have been set up to make the Tumuc1 force field as simple to use as standard force fields.
We have tested the Tumuc1 force field for several systems: MD simulations on the Dickerson-Drew-Dodecamer overall reveal a better performance of Tumuc1 compared to bsc1. Here, we have made remarkable improvements for the dihedral angles, and sampling with the Tumuc1 force field is in excellent agreement with the X-ray structure. Furthermore, we have also obtained decent values for global elasticities. For the A- and T-tract sequences, the Tumuc1 force field exhibits decent sampling as shown by the correct qualitative trends for the internal parameters. Importantly, the Tumuc1 force field guarantees a highly robust description for the backbone: Accurate populations of the BI/BII states, α/γ stability and dominant C2'-endo puckering demonstrate a high quality of our force field. This is confirmed by a $2\,\mu s$ MD simulation of a heterogeneous 50 base-pair sequence. We have not observed an artifact in any of the B-DNA simulations, in contrast to the bsc1 force field, which suffers from a too limited approach. Our main criticism here is, that parameterization of the β- and δ-dihedrals has been ignored, hence sticking to deprecated, rough estimates.
A $2\,\mu s$ MD simulation of a G-quadruplex structure records good performance of the Tumuc1 force field: The stacking and hydrogen bonding pattern remains stable, and the potassium ions stay inside the quadruplex-layers throughout the whole simulation.
In an unrestrained MD simulation, the 5'-GCGCAGC-3' single-strand folded from an extended starting-structure into the correct hairpin-structure with the Tumuc1 force field. This simulation further indicates a good syn-anti balance for the glycosidic-dihedrals. To the best of our knowledge, hairpin folding has only been achieved by means of replica exchange simulations so far. Precise folding of a single-strand during an unrestrained simulation hence gives further trust in the accuracy of Tumuc1.
Likewise, we have simulated folding of two 4 bases long single-strands folding into the correct duplex-structure with Tumuc1. Thereby, several meta-stable states are overcome, which indicates a good description of the kinetics.
Finally, we have also simulated a Z-DNA hexamer at 2M NaCl employing either the Tumuc1 or bsc1 force field. Simulations with Tumuc1 show a substantially stronger fraying of the terminal base-pairs. However, stacking of the inner sequence and sugar puckering is better captured by Tumuc1. On the other hand, both force fields show a lack in the description of α/γ substates for Z-DNA.
Altogether, we find that the Tumuc1 force field performs on a high level and in many ways outperforms bsc1, which is widely considered as state of the art. The development of the Tumuc1 force field is therefore completed and will be released to stand the test of time. Nevertheless, expecting fully satisfying results for all applications without further room for improvements is naive. A main concept in the construction of Tumuc1 has been to making changes in the force field easily achievable without the danger of causing inconsistencies. For this reason, we have developed Tumuc1 from just one

theory, generated a new library with more atom-types and realized several parameterization techniques. Thus, we claim the best fundament for the long-term competition between DNA force fields and are optimistic to also contribute to the theory of other biomolecules. As a next step, we are intending to develop a RNA force field.

8.5 Bibliography

[1] Alberto Pérez, F. Javier Luque, and Modesto Orozco. Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.*, 129(47):14739–14745, 2007.

[2] Pablo D Dans, Linda Danilane, Ivan Ivani, Tomáš Dršata, Filip Lankaš, Jürgen Walther, Ricard Illa Pujagut, Federica Battistini, Josep Lluis Gelpí, and Richard Lavery. Long-Timescale Dynamics of the Drew–Dickerson Dodecamer. *Nucleic Acids Res.*, 44(9):4052–4066, 2016.

[3] Marco Pasi, John H. Maddocks, David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham, III, Pablo D. Dans, B. Jayaram, Filip Lankaš, Charles Laughton, and et al. μABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. *Nucleic Acids Res.*, 42(19):12272–12283, 2014.

[4] Jiří Šponer, Giovanni Bussi, Miroslav Krepl, Pavel Banáš, Sandro Bottaro, Richard A. Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simn Poblete, Petr Jurečka, Nils G. Walter, and Michal Otyepka. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chem. Rev.*, 118(8):4177–4338, 2018.

[5] Rodrigo Galindo-Murillo, James C. Robertson, Marie Zgarbová, Jiří Šponer, Michal Otyepka, Petr Jurečka, and Thomas E. Cheatham. Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.*, 12(8):4114–4127, 2016.

[6] Christopher Maffeo, Binquan Luan, and Aleksei Aksimentiev. End-to-end attraction of duplex DNA. *Nucleic Acids Res.*, 40(9):3812–3821, 2012.

[7] Reid F. Brown, Casey T. Andrews, and Adrian H. Elcock. Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment. *J. Chem. Theory Comput.*, 11(5):2315–2328, 2015.

[8] Pavel Banáš, Arnošt Mládek, Michal Otyepka, Marie Zgarbová, Petr Jurečka, Daniel Svozil, Filip Lankaš, and Jiří Šponer. Can We Accurately Describe the Structure of Adenine Tracts in B-DNA? Reference Quantum-Chemical Computations Reveal Overstabilization of Stacking by Molecular Mechanics. *J. Chem. Theory Comput.*, 8(7):2448–2460, 2012.

[9] Florian Häse and Martin Zacharias. Free energy analysis and mechanism of base pair stacking in nicked DNA. *Nucleic Acids Res.*, 44(15):7100–7108, 2016.

[10] Jejoong Yoo and Aleksei Aksimentiev. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Phys. Chem. Chem. Phys.*, 20(13):8432–8449, 2018.

[11] Dazhi Tan, Stefano Piana, Robert M Dirks, and David E Shaw. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. USA*, 115(7):E1346–E1355, 2018.

[12] Holger Kruse, Pavel Banáš, and Jiří Šponer. Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *J. Chem. Theory Comput.*, 15(1):95–115, 2019.

[13] Korbinian Liebl and Martin Zacharias. How global DNA unwinding causes non-uniform stress distribution and melting of DNA. *PLOS ONE*, 15(5):e0232976, 2020.

[14] Petra Kührová, Robert B. Best, Sandro Bottaro, Giovanni Bussi, Jiří Šponer, Michal Otyepka, and Pavel Banáš. Computer Folding of RNA Tetraloops: Identification of Key Force Field Deficiencies. *J. Chem. Theory Comput.*, 12(9):4534–4548, 2016.

[15] Petra Kührová, Vojtěch Mlýnský, Marie Zgarbová, Miroslav Krepl, Giovanni Bussi, Robert B. Best, Michal Otyepka, Jiří Šponer, and Pavel Banáš. Improving the Performance of the Amber RNA Force Field by Tuning the Hydrogen-Bonding Interactions. *J. Chem. Theory Comput.*, 15(5):3288–3305, 2019.

[16] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiří Šponer, 3rd Cheatham, Thomas E, Charles A Laughton, and Modesto Orozco. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophysical Journal*, 92(11):3817–3829, 2007.

[17] Marie Zgarbová, Michal Otyepka, Jiří Šponer, Arnošt Mládek, Pavel Banáš, Thomas E. Cheatham, and Petr Jurečka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*, 7(9):2886–2902, 2011.

[18] Miroslav Krepl, Marie Zgarbová, Petr Stadlbauer, Michal Otyepka, Pavel Banáš, Jaroslav Koča, Thomas E. Cheatham, Petr Jurečka, and Jiří Šponer. Reference Simulations of Noncanonical Nucleic Acids with Different χ Variants of the AMBER Force Field: Quadruplex DNA, Quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.*, 8(7):2506–2520, 2012.

[19] Marie Zgarbová, F. Javier Luque, Jiří Šponer, Thomas E. Cheatham, Michal Otyepka, and Petr Jurečka. Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.*, 9(5):2339–2354, 2013.

[20] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham, Rodrigo Galindo-Murillo, and Petr Jurečka. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.*, 11(12):5723–5736, 2015.

[21] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jrgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem

Portella, Federica Battistini, Josep Lluis Gelp, Carlos Gonzlez, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55–58, 2016.

[22] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[23] Piotr Cieplak, Wendy D. Cornell, Christopher Bayly, and Peter A. Kollman. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.*, 16(11):1357–1377, 1995.

[24] Alice E. A. Allen, Michael C. Payne, and Daniel J. Cole. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.*, 14(1):274–281, 2018.

[25] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[26] Tian Lu and Feiwu Chen. Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.*, 33(5):580–592, 2012.

[27] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[28] Vincenzo Barone and Maurizio Cossi. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A*, 102(11):1995–2001, 1998.

[29] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

[30] Florian Weigend. Hartree-Fock exchange fitting basis sets for H to Rn . *J. Comput. Chem.*, 29(2):167–175, 2008.

[31] Frank Neese. Software update: the ORCA program system, version 4.0. *WIREs Comput Mol Sci*, 8(1):e1327, 2018.

[32] C. Altona and M. Sundaralingam. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. A New Description Using the Concept of Pseudorotation. *Journal of the American Chemical Society*, 1972.

[33] Bohdan Schneider and Helen M. Berman. Basics of Nucleic Acid Structure. In *Computational Studies of RNA and DNA (Challenges and Advances in Computational Chemistry and Physics vol 2)*, chapter 1, pages 1–44. Springer, 2006.

[34] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W. Ponder, and Pengyu Ren. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.*, 14(4):2084–2108, 2018.

[35] H R Drew, R M Wing, T Takano, C Broka, S Tanaka, K Itakura, and R E Dickerson. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. USA*, 78(4):2179–2183, 1981.

[36] Zhengrong Wu, Frank Delaglio, Nico Tjandra, Victor B. Zhurkin, and Ad Bax. Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and 31P chemical shift anisotropy. *Journal of Biomolecular NMR*, 26(4):297–315, 2003.

[37] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[38] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.*, 5(21):3863–3871, 2014.

[39] Jan Lipfert, Gary M. Skinner, Johannes M. Keegstra, Toivo Hensgens, Tessa Jager, David Dulin, Mariana Kber, Zhongbo Yu, Serge P. Donkers, Fang-Chieh Chou, Rhiju Das, and Nynke H. Dekker. Double-stranded RNA under force and torque: Similarities to and striking differences from double-stranded DNA. *Proc. Natl. Acad. Sci. USA*, 111(43):15408–15413, 2014.

[40] Victor Velasco-Berrelleza, Matthew Burman, Jack W. Shepherd, Mark C. Leake, Ramin Golestanian, and Agnes Noy. SerraNA: a program to determine nucleic acids elasticity from simulation data. *bioRxiv*, page 004945, 2020.

[41] Agnes Noy and Ramin Golestanian. Length Scale Dependence of DNA Mechanical Properties. *Phys. Rev. Lett.*, 109(22):228101, 2012.

[42] Daniel Svozil, Jan Kalina, Marek Omelka, and Bohdan Schneider. Dna conformations and their sequence preferences. *Nucleic Acids Res.*, 36(11):3690–3706, 2008.

[43] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. *J. Molec. Graphics*, 14(1):33–38, 1996.

[44] Ye Tian, Michael Kayatta, Katharine Shultis, Alejandro Gonzalez, Leonard J. Mueller, and Mary E. Hatcher. 31P NMR Investigation of Backbone Dynamics in DNA Binding Sites. *J. Phys. Chem. B*, 113(9):2596–2603, 2009.

[45] Charles D. Schwieters and G. Marius Clore. A Physical Picture of Atomic Motions within the Dickerson DNA Dodecamer in Solution derived from Joint Ensemble Refinement against NMR and Large-Angle X-ray Scattering Data. *Biochemistry*, 46(5):1152–1166, 2007.

[46] Richard Stefl, Haihong Wu, Sapna Ravindranathan, Vladimír Sklenář, and Juli Feigon. DNA A-tract bending in three dimensions: Solving the $dA_4T_4vs.dT_4A_4$ conundrum. *Proc. Natl. Acad. Sci. USA*, 101(5):1177–1182, 2004.

[47] Tali E. Haran and Udayan Mohanty. The unique structure of A-tracts and intrinsic DNA bending. *Quaterly Reviews of Biophysics*, 42(1):41–81, 2009.

[48] Péter Várnai, Dragana Djuranovic, Richard Lavery, and Brigitte Hartmann. α/γ Transitions in the BDNA backbone. *Nucleic Acids Res.*, 30(24):5398–5406, 2002.

[49] Brahim Heddi, Christophe Oguey, Christophe Lavelle, Nicolas Foloppe, and Brigitte Hartmann. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.*, 38(3):1034–1047, 2009.

[50] Marie Zgarbová, Petr Jurečka, Jiří Šponer, and Michal Otyepka. A- to B-DNA Transition in AMBER Force Fields and Its Coupling to Sugar Pucker. *J. Chem. Theory Comput.*, 14(1):319–328, 2018.

[51] Richard Wing, Horace Drew, Tsunehiro Takano, Chris Broka, Shoji Tanaka, Keiichi Itakura, and Richard E. Dickerson. Crystal structure analysis of a complete turn of B-DNA. *Nature*, 287(5784):755–758, 1980.

[52] J C Wang. Helical repeat of DNA in solution. *Proc. Natl. Acad. Sci. USA*, 76(1):200–203, 1979.

[53] Jan Karlseder, Agata Smogorzewska, and Titia de Lange. Senescence induced by Altered Telomere State, Not Telomere Loss. *Science*, 295(5564):2446–2449, 2002.

[54] Jochen Spiegel, Santosh Adhikari, and Shankar Balasubramanian. The Structure and Function of DNA G-Quadruplexes. *Trends in Chemistry*, 2(2):123–136, 2020.

[55] Jiří Šponer, Arnošt Mládek, Nad'a Špačková, Xiaohui Cang, Thomas E. Cheatham, and Stefan Grimme. Relative Stability of Different DNA Guanine Quadruplex Stem Topologies Derived Using Large-Scale Quantum-Chemical Computations. *J. Am. Chem. Soc.*, 135(26):9785–9796, 2013.

[56] Till Siebenmorgen and Martin Zacharias. Origin of Ion Specificity of Telomeric DNA G-Quadruplexes Investigated by Free-Energy Simulations. *Biophysical Journal*, 112(11):2280–2290, 2017.

[57] Gary N. Parkinson, Michael P. H. Lee, and Stephen Neidle. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, 417(6891):876–880, 2002.

[58] Srinivasaraghavan Kannan and Martin Zacharias. Folding of a DNA Hairpin Loop Structure in Explicit Solvent Using Replica-Exchange Molecular Dynamics Simulations. *Biophysical Journal*, 93(9):3218–3228, 2007.

[59] Xiaohui Cang, Jiří Šponer, and III Cheatham, Thomas E. Explaining the varied glycosidic conformational, G-tract length and sequence preferences for anti-parallel G-quadruplexes. *Nucleic Acids Res.*, 39(10):4499–4512, 2011.

[60] Vojtěch Mlýnský, Petra Kührová, Tomáš Kühr, Michal Otyepka, Giovanni Bussi, Pavel Banáš, and Jiří Šponer. Fine-Tuning of the AMBER RNA Force Field with a New Term Adjusting Interactions of Terminal Nucleotides. *J. Chem. Theory Comput.*, 16(6):3936–3946, 2020.

[61] Guillem Portella and Modesto Orozco. Multiple Routes to Characterize the Folding of a Small DNA Hairpin. *Angewandte Chemie International Edition*, 49(42):7673–7676, 2010.

[62] Zbigniew Dauter and Dorota A. Adamiak. Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Cryst. D*, 57(7):990–995, 2001.

[63] E Di Capua, A Stasiak, T Koller, S Brahms, R Thomae, and F M Pohl. Torsional stress induces left-handed helical stretches in DNA of natural base sequence: circular dichroism and antibody binding. *EMBO J.*, 2(9):1531–1535, 1983.

[64] David E. Woon and Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. iii. The atoms aluminum through argon. *J. Chem. Phys.*, 98(2):1358–1371, 1993.

[65] Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. i. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90(2):1007–1023, 1989.

[66] Axel D. Becke. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J. Chem. Phys.*, 96(3):2155–2160, 1992.

[67] K. Kim and K. D. Jordan. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.*, 98(40):10089–10094, 1994.

[68] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

[69] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010.

[70] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji,

M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[71] David J. Wales and Harold A. Scheraga. Global Optimization of Clusters, Crystals, and Biomolecules. *Science*, 285(5432):1368–1372, 1999.

[72] R Lavery, M Moakher, JH Maddocks, D Petkeviciute, and K Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37(17):5917–5929, 2009.

9 The Geometry of Blunt-end and Base-Pair Stacking

9.1 Introduction

In this chapter, we study the conformation of DNA blunt-ends. Although experimental data and computational insights are overall scarce on this subject, a few crystal structures suggest that blunt-end stacking adopts a negatively-twisted geometry [1–3]. Furthermore, a left-handed conformation was reported for the stacking of poly(dA:dT) fragments in a MD study by the Aksimentiev group [4]. We have investigated the stacking geometry of DNA blunt-ends by unrestrained MD simulations for all ten possible blunt-end stacks starting from B-DNA like conformations. Our simulations indicate that a configuration with a twist of $\phi \sim -30°$ is indeed the native state for all sequences, except for the GA and GC sequences, where we find equal stabilities between the right- and left-handed conformation. In most cases, a transition to the left-handed state occurs within $\sim 10\,\text{ns}$, and the stability ranges from $\sim 2-4\,\text{kcal/mol}$. We characterize blunt-end stacking by a canonical B-DNA rise of $\sim 3.3-3.4\,\mathring{A}$ and sliding close to $\sim 0\,\mathring{A}$, which reflects negligible lateral displacement for blunt-end base-pair stacking.
The finding of a completely different state compared to regular B-DNA is of high relevance for the following reasons: First, blunt-end stacks represent DNA double-strand breaks and are for instance induced by the Cas9 endonuclease, which is utilized in the CRISPR/Cas-method [5–7]. Double-strand breaks are repaired through the HDR or NHEJ mechanism [8–10]. The large structural difference of blunt-ends to conventional DNA base-pair stacking is likely to be strongly implicated in these processes. Second, blunt-end motives are used in DNA-nanotechnology [11, 12]. Knowledge of their structural arrangement is hence important for applicative purposes. Third, blunt-end stacking also occurs in DNA crystallization. We have found three different blunt-end sequences from DNA crystal structures (TA,GC and TG step), which are in good agreement with our computational results [1–3, 13]. The TA and TG blunt-end steps show a negative twist ($\phi \sim -25°$), whereas GC is twisted by $\phi \sim 26°$. Fourth, we learn a fundamental thing: What is the origin of DNA's right-handed structure? Based on this study, we argue it is not a consequence of base-stacking. It is rather caused by the backbone constraining the base-pairs into the right-handed conformation. This understanding has already been useful for the development of Tumuc1, as it shows that an accurate description of B-DNA's helical structure can hardly be achieved by modulation of nonbonded interactions. Parameterization of the bonded backbone-terms is central therefore.

We have complemented our MD simulations of blunt-end structures by extensive DFT-calculations on base-pair stacks. Thereby, we have scanned the quantum mechanical potential energy for twisting of all ten base-pair stacks. These calculations reveal that B-DNA like stacking conformations ($\phi \sim 34.5°$) do not represent the energetically most favorable base-pair stacking configuration. In general, base-pair stacking turns out to become more stable for modest twisting ($\phi \sim 0°$). However, the energetically most stable state in the MD simulations on the blunt-end systems as well as in the QM calculations on the base-pair stacks is characterized by translational base-pair step parameters corresponding to B-DNA ($rise \sim 3.3 - 3.4\,\mathring{A}$ and $slide \sim 0\,\mathring{A}$).
For the blunt-end systems, we obtain a negative twist, as the $\phi \sim 0°$ state is unfavorable due to clashing of the sugar rings, hence they adopt a stacking geometry with a twist of $\phi \sim -30°$. Therefore, we point out that blunt-end stacking prefers a negatively twisted state and that the right-handed helical structure of B-DNA may be a constraint-effect of the backbone-strands.

9.2 Materials and Methods

9.2.1 MD Protocol

Preparation of Blunt-end Systems

Starting structures have been constructed from complementary d(5'-CGCXYCGC-3') B-DNA structures, which were generated with the nab-module [14]. X and Y denote the base-pairs between which double-strand breaks were induced to built blunt-end stacked DNA duplexes (Fig 9.1 A). All ten possible base-pair steps (XY) were studied. The double-strand breaks were incorporated by removing the phosphate atoms of the Watson- and Crick-strand in the fourth base-pair step (between X and Y base-pairs). Residue names of the bases in the X/Y base-pairs were changed to 3'- and 5'-ends, respectively. The structures are then automatically completed at the junction with the xleap-module. The bsc1 force field [15] and TIP3P water model [16] was used ($10\,\mathring{A}$ between DNA and box-boundary). The systems were neutralized with sodium ions.

MD Simulations

The systems were energy-minimized in 2500 steps with the sander-module and afterwards stepwise heated up and equilibrated for the NPT-ensemble (p=1bar, T=300K) in consecutive MD simulations with the pmemd.cuda module of the Amber16 package [14]. This protocol covers a simulation time of 2 ns. The output-structures served as input for the subsequent production runs, which lasted 200 ns. No restraints were applied in the production runs.

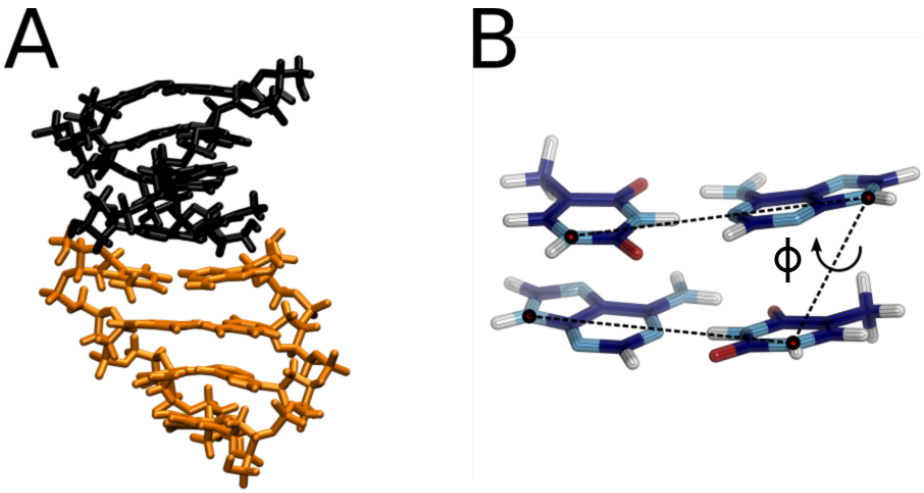


Figure 9.1: Studied Structures: (A) MD simulations were performed on blunt-end stacked duplexes. (B) The base-pair stacking geometry has been scanned with QM calculations by employing a dihedral angle on the glycosidic nitrogens as reaction coordinate.

9.2.2 QM Calculations

DFT calculations were conducted for all ten base-pair stacks. Note that these systems do not include sugar- or backbone-atoms (Fig 9.1 B). The B3LYP functionals [17–19] and def2-TZVP basis-set with an auxiliary basis-set for the Coulomb-integrals were used [20, 21]. The Becke-Johnson damping function (D3BJ) [22] was employed to account for dispersion correction, and the QM calculations were performed at implicit solvent (CPCM, $\epsilon = 80$) [23]. Stacking of the base-pairs was scanned along the dihedral $\angle(N_{W1}, N_{C1}, N_{C2}, N_{W2})$ in 5°-steps between 35° and −35°. N_{W1} (and other atoms accordingly) denotes the glycosidic-nitrogen atom of the Watson-base in the first base-pair. Scanning of the dihedral angle was done by means of geometry optimization at the constrained dihedral angles. In cases where the optimization procedure was not successful, geometry optimization with BLYP-functionals [17, 18] was performed. The output of these calculations was then used as input for new calculations with the B3LYP-functionals.

9.3 Results and Discussion

9.3.1 MD Simulations of Blunt-end Base-Pair Steps

We have started the unrestrained 200 ns MD simulations from blunt-end stacking geometries similar to the right-handed B-DNA geometry. However, structural transitions

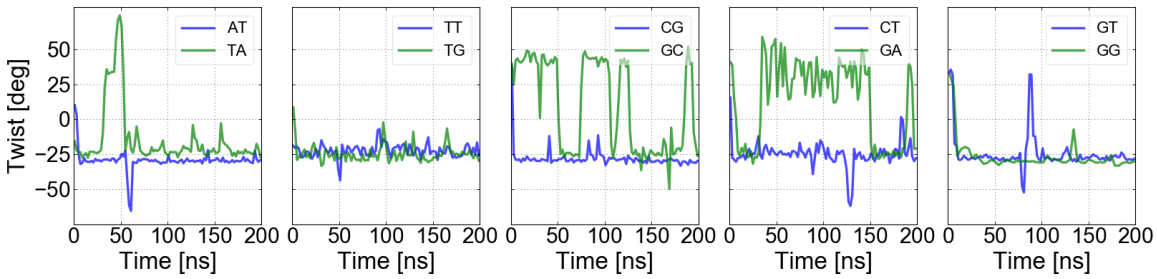


Figure 9.2: Twist of the DNA blunt-end stacks as recorded from the MD simulations for all sequences. Note that the simulations were started from a right-handed conformation ($\phi > 0$). The plot was generated by averaging of 10 ps bins.

to a left-handed state ($\phi \sim -30°$) occur after already ~ 10 ns (Fig 9.2). Note that we have computed the geometric parameters with our protocol presented in chapter 6 [24].

Only the GC and GA sequences behave differently, and show a strong population of the canonical B-DNA state ($\phi \sim 34.5°$). We have quantified the stabilities of the states by recording the probability as a function of twist and subsequently calculating relative free energy profiles by Boltzmann inversion, $F = -k_B T \cdot \ln(p(\phi))$. The resulting free energy curves illustrate a stability of the left-handed state of $\sim 2 - 4$ kcal/mol, except for the GC and GA steps (Fig 9.3, 9.4). This state is stabilized by unconventional hydrogen bonds between HO5' and phosphate bound oxygens (Fig 9.4). Otherwise, our results indicate that base-pair stacking in a B-DNA twisted configuration is energetically unfavorable. Thus, we argue that B-DNA's helical structure is not primarily shaped by stacking interactions, but by the backbone-strand which induces large restrictions on the conformational space for base-pair stacking.

From database queries, we have found three different blunt-end sequences: TA, GC and TG (Fig 9.5). Intriguingly, these structures also show negative twisting for the TA and TG blunt-end stacks ($\phi = -21.6°/-25.8°$), whereas the GC-step is in right-handed configuration ($\phi = 25.8°$) [1, 2, 13]. Although these values are not in perfect quantitative agreement with the free energy curves obtained from the MD simulations, this confirms our key message: Blunt-end stacking prefers a left-handed geometry. Furthermore, the sequence-dependent behavior is also in line with the MD simulations, as we have sampled a large proportion of right-handed states for the GC-step.

Moreover, every blunt-end stack exhibits instability for non-twisted conformations ($\phi \sim 0°$) in our simulations, which is a consequence of steric repulsion of the sugar-groups (Fig 9.4 B). This conformation is only accessible for high axial displacement (rise) between the blunt-end base-pairs (see maxima in Fig 9.6). For the $\phi \sim -30°$ state, the axial displacement is analog to B-DNA, $\sim 3.3 - 3.4\,\mathring{A}$ (Fig 9.6). The lateral displacement (slide), in contrast, approaches to $\sim 0\,\mathring{A}$, hence causing strong overlap between the blunt-end base-pairs (Fig 9.7).

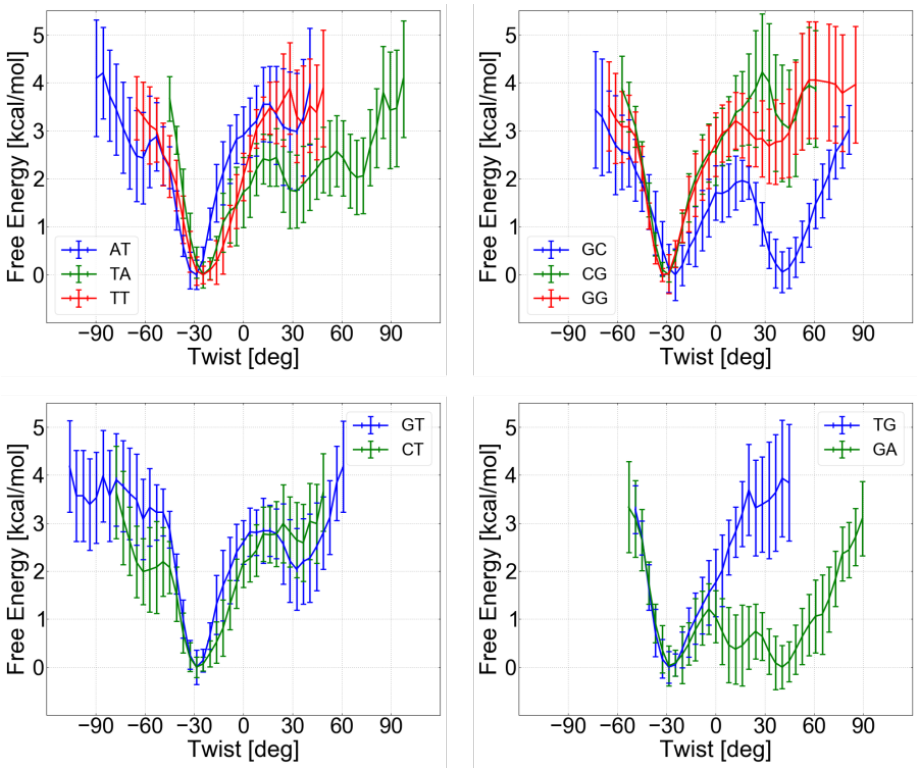


Figure 9.3: Relative free energy profiles as a function of twist for all sequences. Free energy estimates were obtained by Boltzmann inversion of the probabilities. The probabilities have been computed by dividing the covered twist-range into 60 equally spaced bins and counting the population of each bin. The error bars represent standard deviations which have been obtained by slicing the trajectory into 100 subintervals. For each subinterval, a free energy curve is computed, and the final standard deviations are then calculated over the free energy curves of all subintervals.

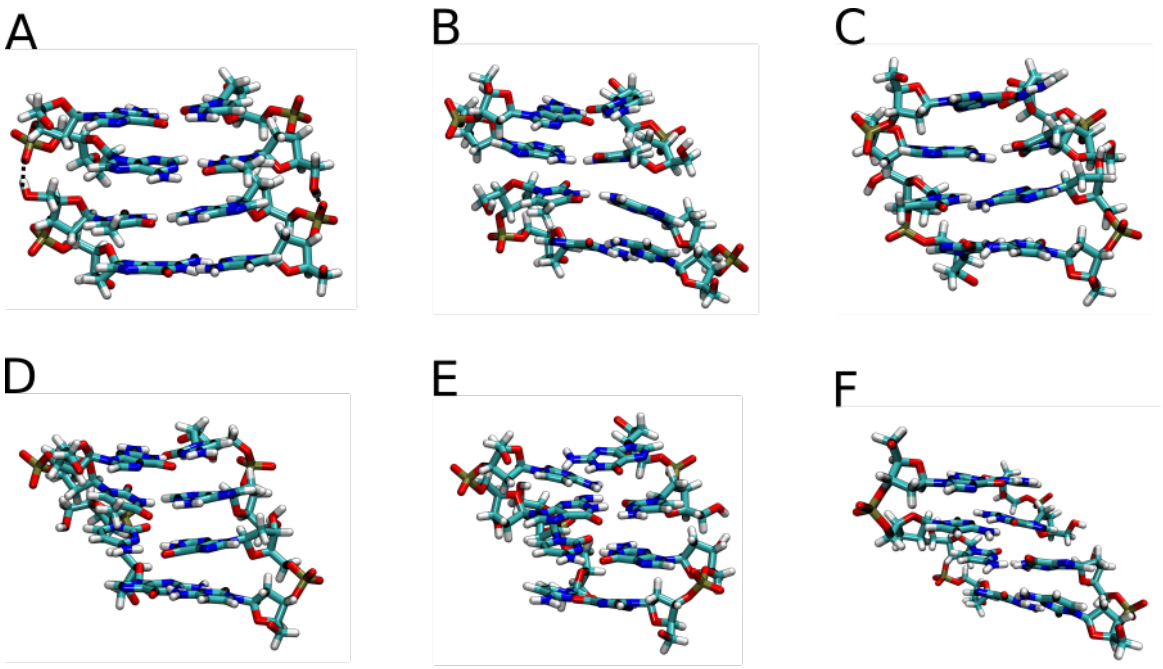


Figure 9.4: Structural snapshots taken from the MD simulations: (A) AT-step at $\phi \sim -30°$. This configuration is stabilized by hydrogen bonds between the HO5' atom and a phosphate-bound oxygen, see dashed lines. (B) AT-step at $\phi \sim 0°$. The blunt-end stack shows strong axial displacement in order to avoid sterical clashing of the sugar and backbone groups. (C) AT-step at $\phi \sim 35°$. (D,E) The GA- and GC-step are meta-stable at $\phi \sim 40°$. (F) The right-handed conformation is unfavorable for the GG-step.

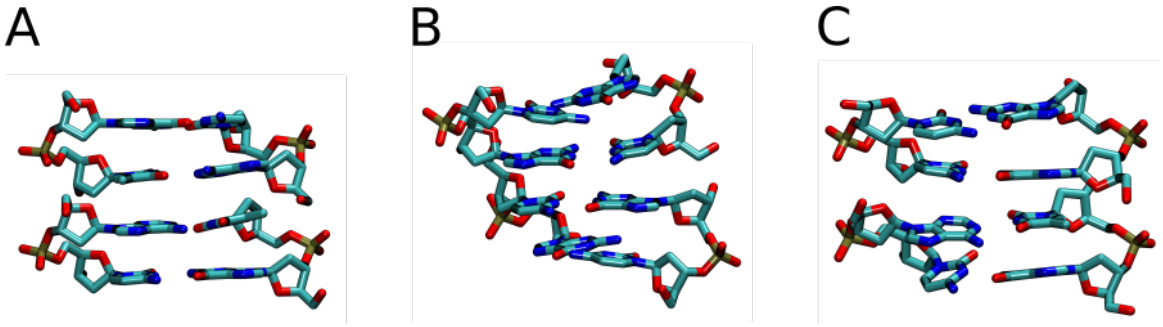


Figure 9.5: Blunt-end structures found from database: (A) TA-step in left-handed conformation ($\phi = -21.6°$, pdb:1ilc) [2]. (B) GC-step in right-handed conformation ($\phi = 25.8°$, pdb:1d23) [13]. (C) TG-step in left-handed conformation ($\phi = -25.8°$, pdb:330d) [1].

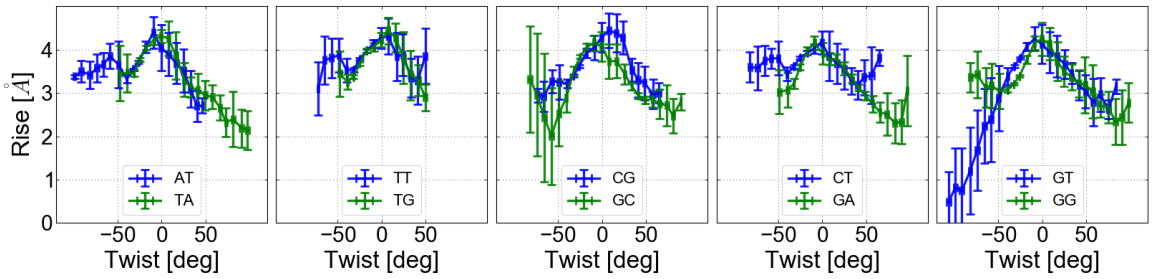


Figure 9.6: Twist-rise profiles obtained from the MD simulations for all ten blunt-end sequences. The maxima positions at $\phi \sim 0°$ reflect evasive motions due to sugar-clashing. The plots were generated in a similar way as for Fig 9.3.

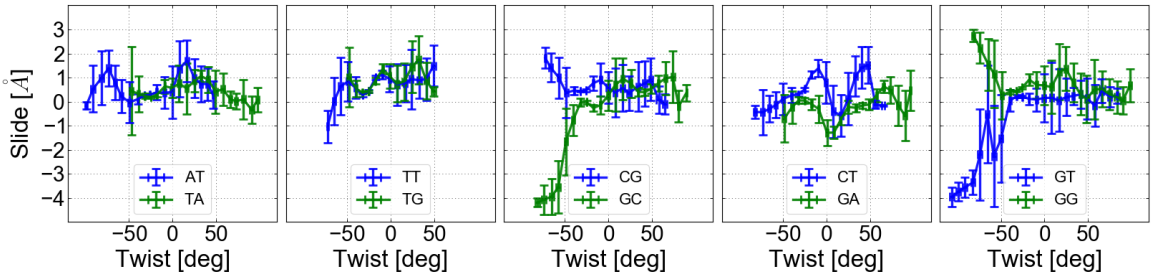


Figure 9.7: Twist-slide profiles obtained from the MD simulations on the blunt-end structures. The plots were generated in a similar way as for Fig 9.3.

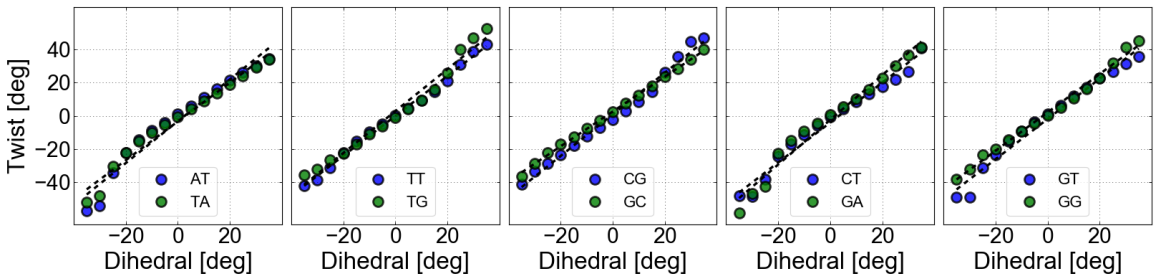


Figure 9.8: Correlation between the reaction-coordinate, the dihedral angle, and the twist variable, obtained from the QM scans of all ten possible DNA base-pair stacks. The dashed lines represent linear interpolation between the dihedral and the twist angles. The slopes are in the range of 1.0 to 1.3.

9.3.2 QM Calculations of the Potential Energy of Base-Pair Stacking

We have scanned the potential energy surface for all ten possible DNA base-pair stacks in a series of DFT-calculations. Thereby, we have employed a dihedral angle (defined by the glycosidic nitrogens) which shows a good correlation with the twist variable (also used in the analysis of MD simulations), see Fig 9.8. Thus, our setup facilitates stepwise untwisting of base-pair stacking. Structural snapshots are depicted in Fig 9.9 and demonstrate sound sampling with our methodology.

The computed energies as a function of twist are shown for all sequences in Fig 9.10. These profiles reveal that a B-form stacking conformation is not the optimal state for plain base-pairs. In fact, a conformation characterized by modest twisting ($\phi \sim 0°$) is mostly favored. Exceptions therefore are the AT, CG, TG and GC sequence. The TG step prefers twisting of $\sim -20°$ and the GC step twisting of $\sim 20°$. Note that this corresponds very well with the blunt-end crystal structures for the two sequences which occupy a very similar state. From the computed QM energy landscapes, we conclude that base-pair stacking alone does not give rise to B-DNA's right-handed helicity, but the backbone plays a key role. Consequently, this also suggests conformational transitions upon backbone damages.

The relationship between twist and rise as obtained from the QM calculations is given in Fig 9.11. At the ground state, the base-pair steps have a rise characteristic for B-DNA ($\sim 3.3 - 3.4\,\mathring{A}$). In our QM calculations, the slide variable undergoes remarkable fluctuations (Fig 9.12), but the energy minima coincide with a slide of $\sim 0\,\mathring{A}$ for many sequences (compare Fig 9.10 and 9.12). This is especially apparent for the CG and GC sequence. Notably, the slide for base-pair stacking is also very small in regular B-DNA (see Table 8.1 in chapter 8). Thus, our QM calculations reveal that the ground state in base-pair stacking has compared to regular B-DNA very similar values for translational but not for rotational degrees of freedom. The twist of $\sim 34.5°$ for B-DNA may hence be enforced by accommodation of the backbone atoms. Note that a different conclusion can

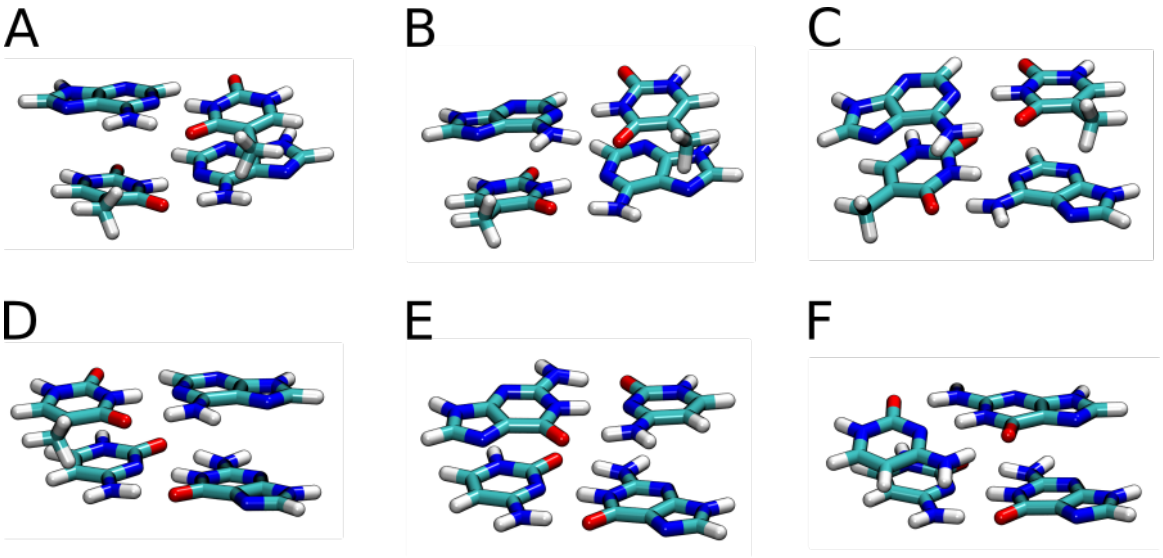


Figure 9.9: Geometry-optimized structures extracted from the QM calculations: (A) AT-step at $\phi = -30°$. (B) AT-step at $\phi = 0°$. (C) AT-step at $\phi = 35°$. (D) GA-step at $\phi \sim 40°$. (E) GC-step at $\phi \sim 40°$. (F) GG-step at $\phi \sim 40°$.

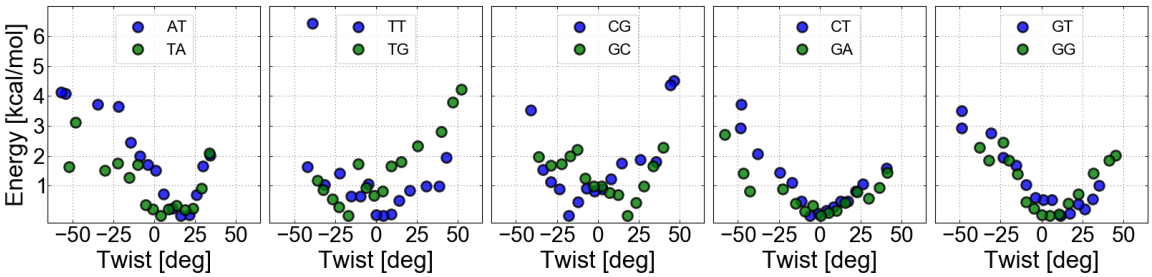


Figure 9.10: Quantum mechanical potential energy as a function of twist for all ten base-pair steps.

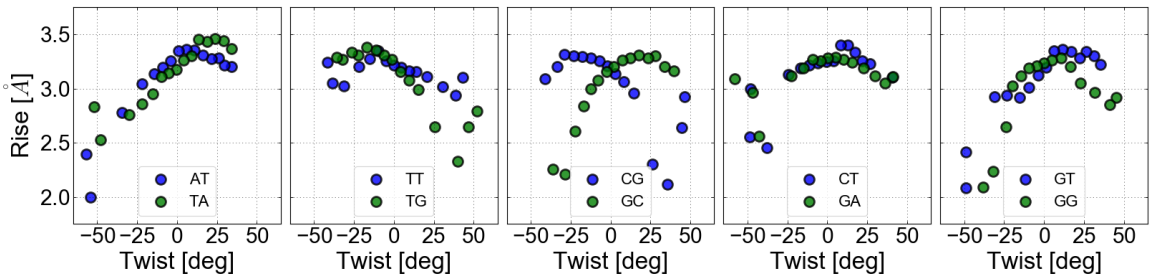


Figure 9.11: Relation between twist and rise as obtained from the QM calculations for all base-pair steps.

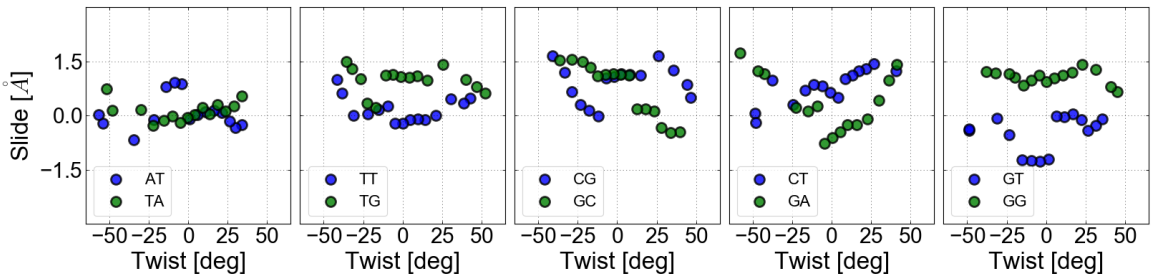


Figure 9.12: Relation between twist and slide as obtained from the QM calculations for all base-pair steps.

be inferred from the QM calculations by Parker et al [25]. However, these calculations have been performed with the SAPT0 method, exclude solvent effects and have scanned twisting over shorter ranges while including constraints on orthogonal degrees of freedom.

9.4 Conclusion

Blunt-end stacked DNA structures represent double-strand breaks and can occur in vivo both, inadvertently and deliberately [26]. Resolving the structure of blunt-end stacks is therefore important for understanding the recognition and repair of double-strand breaks [8–10]. Furthermore, this motif is also used in DNA-nanotechnology and occurs in DNA crystals [11, 12].
In this study, we have carried out unrestrained 200 ns long MD simulations for all ten possible DNA blunt-end stacks exposed to explicit solvent. While the blunt-end structures adopt a B-DNA like geometry with respect to translational parameters ($rise \sim 3.3 - 3.4\,\mathring{A}$ and $slide \sim 0\,\mathring{A}$), they undergo a rapid transition to a left-handed conformation ($\phi \sim -30°$) in our MD simulations. The MD simulations allowed us to calculated free energy profiles as a function of the twist-parameter through Boltzmann inversion. Calculated stabilities for the left-handed state range from $2 - 4\,$kcal/mol, except for the GA and GC sequence. For these two sequences the left-handed state shows equal stability compared to the canonical B-DNA conformation ($\phi \sim 34.5°$). From database

queries, we have found three different blunt-end sequences (TA, GC and TG) [1, 2, 13]. These crystal structures widely agree with our MD simulations: The TA and TG steps are in a negatively twisted conformation, whereas the GC-step is twisted by 25.8°. Thus, we give strong evidence that a left-handed state is the preferred conformation for DNA blunt-end steps.

In order to seek a better understanding of base-pair stacking, we have performed extensive QM calculations (DFT, implicit solvent) for the twisting of base-pair steps. Resulting energy landscapes show that a canonical twist of $\phi \sim 34.5°$ does not represent the optimal conformation for base-pair stacking. Indeed, the ground state structures adopt similar translational parameters as in B-DNA and the base-pair steps have a twist of $\phi \sim 0°$ for most sequences. Based on these results, we argue that base-pair stacking alone does not cause B-DNA's right-handed helicity. We suppose that adaption of the backbone structure to the base-pair geometries plays a key role.

For the future, our study might be extended by focusing on the following two aspects: First, our MD simulations indicate that blunt-end stacking is stabilized by unconventional hydrogen bonds between the HO5' atoms and a phosphate-bound oxygen. We will quantify the energetic contribution of this hydrogen bond by MD simulations during which we switch off these interactions (employing the same technique as proposed in chapter 5, [27]). Second, more extensive, higher-dimensional QM scans might be carried out for base-pair stacking, in order to clearly dissect the energy profiles for rotational and translational degrees of freedom.

9.5 Bibliography

[1] Youri Timsit, Eric Vilbois, and Dino Moras. Base-pairing shift in the major groove of (CA)n tracts by B-DNA crystal structures. *Nature*, 354(6349):167–170, 1991.

[2] Jenny Hizver, Haim Rozenberg, Felix Frolow, Dov Rabinovich, and Zippora Shakked. DNA bending by an adenine-thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. USA*, 98(15):8490–8495, 2001.

[3] Lourdes Campos, Núria Valls, Lourdes Urpí, Catherine Gouyette, Trinidad Sanmartín, Michael Richter, Elida Alechaga, Alicia Santaolalla, Roberto Baldini, Marc Creixell, Ruth Ciurans, Petr Skokan, Joan Pous, and Juan A. Subirana. Overview of the Structure of All-AT Oligonucleotides: Organization in Helices and Packing Interactions. *Biophysical Journal*, 91(3):892–903, 2006.

[4] Christopher Maffeo, Binquan Luan, and Aleksei Aksimentiev. End-to-end attraction of duplex DNA. *Nucleic Acids Res.*, 40(9):3812–3821, 2012.

[5] Jian-Ping Zhang, Xiao-Lan Li, Guo-Hua Li, Wanqiu Chen, Cameron Arakaki, Gary D. Botimer, David Baylink, Lu Zhang, Wei Wen, Ya-Wen Fu, Jing Xu, Noah Chun, Weiping Yuan, Tao Cheng, and Xiao-Bing Zhang. Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biology*, 18(1):1–18, 2017.

[6] Jonathan M Geisinger, Sren Turan, Sophia Hernandez, Laura P Spector, and Michele P Calos. In vivo blunt-end cloning through CRISPR/Cas9-facilitated non-homologous end-joining. *Nucleic Acids Res.*, 44(8):e76, 2016.

[7] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096):816–821, 2012.

[8] Swetha Parvathaneni, Alexei Stortchevoi, Joshua A Sommers, Jr Brosh, Robert M, and Sudha Sharma. Human RECQ1 interacts with Ku70/80 and modulates DNA end-joining of double-strand breaks. *PLOS ONE*, 8(5):e62481, 2013.

[9] PR Blier, AJ Griffith, J Craft, and JA Hardin. Binding of Ku protein to DNA. Measurement of affinity for ends and demonstration of binding to nicks. *The Journal of Biological Chemistry*, 268(10):7594–7601, 1993.

[10] Michael R Lieber. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry*, 79:181–211, 2010.

[11] Fabian Kilchherr, Christian Wachauf, Benjamin Pelz, Matthias Rief, Martin Zacharias, and Hendrik Dietz. Single-molecule dissection of stacking forces in DNA. *Science*, 353(6304), 2016.

[12] Thomas Gerling and Hendrik Dietz. Reversible Covalent Stabilization of Stacking Contacts in DNA Assemblies. *Angew. Chem. Int. Ed.*, 58(9):2680–2684, 2019.

[13] K Grzeskowiak, K Yanagi, G G Priv, and R E Dickerson. The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-G. The effect of base pair reversals. *Journal of Biological Chemistry*, 266(14):8861–8883, 1991.

[14] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[15] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jrgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep Lluis Gelp, Carlos Gonzlez, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55–58, 2016.

[16] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[17] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

[18] Axel D. Becke. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J. Chem. Phys.*, 96(3):2155–2160, 1992.

[19] K. Kim and K. D. Jordan. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.*, 98(40):10089–10094, 1994.

[20] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

[21] Florian Weigend. Hartree-Fock exchange fitting basis sets for H to Rn . *J. Comput. Chem.*, 29(2):167–175, 2008.

[22] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, 2011.

[23] Vincenzo Barone and Maurizio Cossi. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A*, 102(11):1995–2001, 1998.

[24] Korbinian Liebl and Martin Zacharias. How global DNA unwinding causes non-uniform stress distribution and melting of DNA. *PLOS ONE*, 15(5):e0232976, 2020.

[25] Trent M. Parker, Edward G. Hohenstein, Robert M. Parrish, Nicholas V. Hud, and C. David Sherrill. Quantum-Mechanical Analysis of the Energetic Contributions to π Stacking in Nucleic Acids versus Rise, Twist, and Slide. *J. Am. Chem. Soc.*, 135(4):1306–1316, 2013.

[26] Wendy J Cannan and David S Pederson. Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin. *Journal of cellular physiology*, 231(1):3–14, 2016.

[27] Korbinian Liebl and Martin Zacharias. How methyl-sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.*, 47(3):1132–1140, 2018.

10 Calculation of the Absolute Binding Free Energy for the Cren7-DNA Complex

10.1 Introduction

The binding of proteins to DNA regulates the activity and architecture of the genome [1–6]. For this reason, a large number of MD studies have been devoted to protein-DNA interactions [7–11]. While these studies give valuable insight into the conformational flexibility of protein-DNA complexes, the computation of absolute binding free energies has mostly been approached with oversimplifying methods leading to rather qualitative conclusions [12–16]. Calculating absolute binding free energies from MD simulations is challenging as it requires advanced sampling methods, which guarantee a proper dissociation pathway and convergence. In this study, we have applied an advanced free energy method [17], which is commonly employed only for small ligands, to the Cren7-DNA complex. The Cren7-protein sharply bends the DNA upon binding and in this way compacts the global DNA structure [4, 6]. From experiments, absolute binding free energies of ~ -7.0 to $-9.4\,\mathrm{kcal/mol}$ have been determined for this system, with a slight preference for AT-rich sequences [18, 19].
Here, we demonstrate that the free energy method developed by Woo and Roux is principally applicable to the relatively large Cren7-DNA complex. Our setup is technically sound, however, we obtain a significant overestimation of the binding free energy, $G_{bind} = -23.8\,\mathrm{kcal/mol}$. We discuss that this likely stems from classical force fields (we have used bsc1 for DNA) overstabilizing attractive electrostatic interactions. This possible, major force field deficiency has also been pointed out in a very recent review from the Aksimentiev group [20], for the sliding of PCNA along DNA even an underestimation of the diffusion coefficient by a whole order of magnitude has been reported [8]. Thus, we emphasize that further investigations and optimizations of the force field accuracy for DNA/protein complexes are urgent.
Besides, this method allows us to compute the energetic costs required to keep the protein as well as the DNA in the configurational subspace of their bound forms. For the Cren7 protein, we obtain a configurational contribution of only $\sim 0.9\,\mathrm{kcal/mol}$. Substantially higher costs arise for the DNA, where we find a configurational contribution of $\sim 6.3\,\mathrm{kcal/mol}$. The large difference in the configurational terms for the two binding partners reflects that the Cren7 protein undergoes only little structural adaption in the

binding process, whereas the DNA is strongly deformed (bent). We have compared the configurational cost for the DNA to the deformation energy calculated from a harmonic stiffness model, which is based on the two bending modes of the binding site. The harmonic stiffness model indicates a deformation energy for the DNA of $\sim 8.6\,\mathrm{kcal/mol}$. While this model hence yields an overestimation of more than 2 kcal/mol compared to the configurational term, we nevertheless can draw qualitative conclusions from the resulting energy landscape, which depends on the two bending modes. Here, we show that the DNA is not bent along its principal axis in the complex. Thus, our findings indicate an active role of the Cren7 protein: Upon binding, the DNA is strongly bent, whereby its own elasticity is not determinant for the deformation. Moreover, this process does not require energetic-expensive, structural changes of the Cren7-protein.
Future efforts may be directed to the impact of global restraints on the binding affinity, in order to mimic cooperative effects. Finally, repeating simulations with the Tumuc1 force field is of high interest. By reweighting the trajectories with Tumuc1, we find a reduction of the binding affinity by 12.5 kcal/mol. Although this must be considered as a rough estimate, it gives slight optimism for a better performance of protein/DNA systems when the DNA is described with Tumuc1.

10.2 Materials and Methods

10.2.1 Advanced Sampling Method for the Computation of Absolute Binding Free Energies

In the following, we summarize the advanced free energy method developed by Woo and Roux. Assuming low ligand concentration and neglecting ligand-ligand interactions, the equilibrium binding constant can be written as [17]:

$$K_{eq} = \frac{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta U]}{\int_{bulk} \mathrm{d}1 \delta(r_1 - r_1^*) \int \mathrm{d}X \exp[-\beta U]}. \tag{10.1}$$

The degrees of freedom are denoted by 1 for the ligand and X for the remaining molecules (solvent and receptor). r_1^* is a large distance between receptor and ligand, i.e. the ligand is somewhere in the bulk region [17]. The main idea behind the Woo-Roux method is to enhance sampling by decomposing receptor-ligand binding into restrained binding, orientational, axial and configurational contributions which is achieved by extension of equation 10.1.

$$K_{eq} = \overbrace{\frac{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta U]}{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c)]}}^{\exp[\beta G_c^{site}]} \cdot \overbrace{\frac{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c)]}{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c+U_o)]}}^{\exp[\beta G_o^{site}]} \cdot$$
$$\cdot \overbrace{\frac{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c+U_o)]}{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c+U_o+U_a)]}}^{\exp[\beta G_a^{site}]} \cdot \overbrace{\frac{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c+U_o+U_a)]}{\int_{bulk} \mathrm{d}1\delta(r_1-r_1^*) \int \mathrm{d}X \exp[-\beta(U+U_c+U_o)]}}^{\exp[-\beta G_{restr}^{bind}]/C^0} \cdot$$
$$\cdot \overbrace{\frac{\int_{bulk} \mathrm{d}1\delta(r_1-r_1^*) \int \mathrm{d}X \exp[-\beta(U+U_c+U_o)]}{\int_{bulk} \mathrm{d}1\delta(r_1-r_1^*) \int \mathrm{d}X \exp[-\beta(U+U_c)]}}^{\exp[-\beta G_o^{bulk}]} \cdot \overbrace{\frac{\int_{bulk} \mathrm{d}1\delta(r_1-r_1^*) \int \mathrm{d}X \exp[-\beta(U+U_c)]}{\int_{bulk} \mathrm{d}1\delta(r_1-r_1^*) \int \mathrm{d}X \exp[-\beta U]}}^{\exp[-\beta G_c^{bulk}]} \cdot \tag{10.2}$$

As shown by Woo and Roux, the restrained binding free energy can be split into two terms [17]:

$$\exp[-\beta G_{restr}^{bind}]/C^0 = \frac{\int_{site} \mathrm{d}1 \int \mathrm{d}X \exp[-\beta(U+U_c+U_o+U_a)]}{\int_{bulk} \mathrm{d}1\delta(r_1-r_1^*) \int \mathrm{d}X \exp[-\beta(U+U_c+U_o)]} = S \cdot I, \tag{10.3}$$

with

$$S = (r_1^*)^2 \int_0^{\pi} \sin(\theta_1)\mathrm{d}\theta_1 \int_0^{2\pi} \mathrm{d}\phi_1 \exp[-\beta U_a(\theta_1, \phi_1)] \tag{10.4}$$

and

$$I = \int_{site} \mathrm{d}r_1 \exp[-\beta(W(r_1) - W(r_1^*))]. \tag{10.5}$$

C_0 is the standard concentration of $1/1661\mathring{A}^3$, θ and ϕ are axial angles, and hereafter we denote the orientational angles with α, γ, ϵ [17]. Introducing restraints on the axial angles (U_a) allows us to simplify sampling in the way that ligand and receptor are separated in radial direction. Motions along the spherical surface are hence suppressed. The orientation of the ligand with respect to the receptor is restrained by a potential depending on the orientational angles, $U_o(\alpha, \gamma, \epsilon)$. Furthermore, we make use of restraints in which the internal flexibility of both, receptor and ligand, are restricted. In our case, these are rmsd-restraints on the Cren7 protein and on the DNA-molecule. Due to the isotropy of the bulk, we calculate the orientational bulk term analytically [17]:

$$G_o^{bulk} = -\frac{1}{\beta}\ln\left[\frac{1}{8\pi^2}\int_0^{\pi} \mathrm{d}\alpha\, \sin(\alpha)\int_0^{2\pi}\mathrm{d}\gamma\int_0^{2\pi}\mathrm{d}\epsilon \exp[-\beta U_o(\alpha,\gamma,\epsilon)]\right]. \tag{10.6}$$

The remaining contributions are computed by means of Free energy perturbation. Thus, we can calculate the equilibrium constant [17]:

$$K_{eq} = S \cdot I \cdot \exp[-\beta(G_o^{bulk} + G_c^{bulk} - G_a^{site} - G_o^{site} - G_c^{site})]. \tag{10.7}$$

The absolute binding free energy is [17]:

$$G_{bind} = -\frac{1}{\beta}\ln(K_{eq}C^0) = G_{restr}^{bind} + G_o^{bulk} + G_c^{bulk} - G_a^{site} - G_o^{site} - G_c^{site}. \tag{10.8}$$

10.2.2 MD Simulations

System-Preparation and Equilibration

The Amber16 package was used for the preparation of the systems and all subsequent MD simulations [21]. The starting structure was built based on the crystal structure (pdb:3LWH) [22]. The dsDNA structure was extended on the 5' Watson-end with a complementary d(5'-CGCGC-3') and on the 3' end with a d(5'-GCGCG-3') sequence using a VMD protocol developed by Dr Coles. The resulting DNA-Cren7 complex structure was neutralized with Na+ ions and exposed to explicit solvent (TIP3P water model [23], the distance between the octahedral box boundary and the complex was set to $25\,\mathring{A}$) with the xleap-module [21]. Additional Cl- and Na+ ions were added to achieve a salt concentration of 250mM. The ff14SB [24] and bsc1 force field [25] was used for the protein and the DNA, respectively. Afterwards, the systems were energy-minimized in 60000 steps with the sander-module [21]. The energy-minimized systems were then equilibrated for the NPT-ensemble ($p = 1\,\mathrm{bar}, T = 300\,\mathrm{K}$) which includes gradual removement of positional restraints. The simulation time for the whole equilibration phase amounts to $\sim 2\,\mathrm{ns}$. Finally, we conducted a 350 ns long unrestrained MD simulation. Note that all simulations were carried out with the pmemd.cuda module [21]. Analogously, a system for the DNA as well as the Cren7-protein alone was set up and equilibrated. Besides, we have also conducted MM/GBSA calculations based on the unrestrained MD simulation for salt concentrations of 250mM, 500mM, 750mM and 1M NaCl. Thereby, the gb=5 model was employed [26].

Free Energy Simulations

Virtual sites defined for the restraints are given in Table 10.1, and the restraints used in the free energy simulations are defined in Table 10.2 and are illustrated in Fig 10.1.

Virtual sites	D1	D2	D3	C1	C2	C3
residue number	68,87	69,70	85,86	39,49	51,52,53	36,37,38

Table 10.1: Defined virtual sites are the center of mass of the corresponding residues.

Restraint	Definition	Type
r_1	C3-D1	distance
ϕ	$\angle(C2, C1, D1)$	angle (axial)
θ	$\angle(C3, C2, C1, D1)$	dihedral (axial)
α	$\angle(C3, D1, D2)$	angle (orientational)
γ	$\angle(C2, C1, D1, D2)$	dihedral (orientational)
ϵ	$\angle(C1, D1, D2, D3)$	dihedral (orientational)

Table 10.2: Restraints employed in the advanced sampling setup.

Umbrella Sampling (US) simulations along the distance r_1 were performed in order to evaluate equation 10.5. The US simulations consisted of 48 equally spaced windows with an increase in the reference-distance by $0.60\mathring{A}$ and a force constant of 25 $\frac{\text{kcal}}{\text{mol}\cdot\text{Å}^2}$. Initially, each window was simulated for 2.5 ns. The first 10 windows were simulated for another 20 ns to improve convergence. In cases of insufficient overlap between the probability distributions, additional simulations were carried out to cover sampling gaps. All US simulations included axial, orientational and configurational restraints. The free energy profile $W(r_1)$ was calculated with WHAM [27] and all integrals were computed with the trapezoidal rule. In a next step, we evaluated the axial term, G_a^{site}. Hereto, we performed two simulations: One with a force constant for the axial restraints of 240, and one with 0 $\frac{\text{kcal}}{\text{molrad}^2}$. Each simulation still included orientational and configurational restraints and covered 0.5 ns. In a similar way, the G_o^{site}-term was derived. These simulations included configurational restraints and were otherwise analog to the G_a^{site}-related simulations. Configurational free energy terms were calculated from two MD simulations whereby the rmsd-restraints were reduced from 500 to 0 $\frac{\text{kcal}}{\text{molÅ}^2\text{N}}$ (N denotes number of atoms). This procedure was carried out for three systems: The Cren7-DNA complex, the Cren7-protein and the DNA alone. Simulations in the unrestrained case (0 $\frac{\text{kcal}}{\text{molÅ}^2\text{N}}$), were performed for 50 ns for the Cren7-protein and the Cren7-DNA complex. For the DNA alone, this window was simulated for 100 ns, as the DNA relaxes from the strongly bent configuration to regular B-DNA under these conditions. Based on all those simulations, the $G_a^{site}, G_o^{site}, G_c^{site}$ and $G_c^{bulk} (= G_c^{bulk,DNA} + G_c^{bulk,Cren7})$ terms were determined from free energy perturbation.

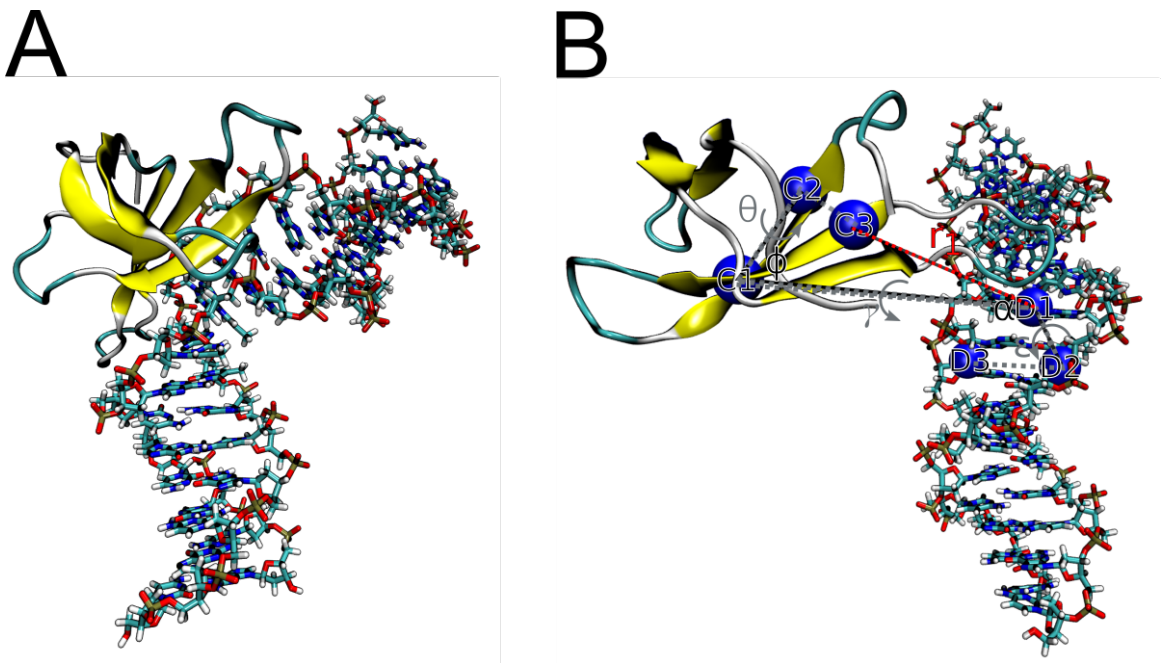


Figure 10.1: Cren7-DNA system. (A) Crystal structure (pdb:3LWH) [22]. The Cren7-protein is shown in Cartoon representation. (B) Illustration of the advanced sampling setup. The defined virtual sites are highlighted by blue spheres.

10.3 Results and Discussion

10.3.1 Binding Free Energy

The calculated free energy contributions for binding of the Cren7 protein to DNA are given in Table 10.3. With equation 10.8, we hence obtain a binding free energy of -23.8 ± 7.3 kcal/mol. Note that this estimate strongly exceeds the experimentally determined binding free energy of $G_{bind}^{exp} \sim -7.0$ to -9.4 kcal/mol [18, 19]. The overestimation of the binding affinity as computed from our MD simulations is mainly caused by the high restrained binding free energy, $G_{restr}^{bind} = -33.89 \pm 7.32$ kcal/mol, which is derived from the distance PMF (Fig 10.2). Most studies rely on MM/PBSA or Steered Molecular Dynamics simulations and yield binding free energies in the order of ~ -50 to -100 kcal/mol, hence having only qualitative meaning [12–16]. By means of MM/GBSA calculations, we have also studied the influence of salt concentration on the binding affinity (Fig 10.3). At a salt concentration of 250mM NaCl, we obtain a binding free energy of -67.7 kcal/mol, which reflects the lower accuracy of the MM/GBSA simulations compared to advanced Umbrella Sampling simulations. Nevertheless, we find only small decreases in the binding affinity upon increase of the salt concentration in our MM/GBSA computations: Calculations at 1M NaCl reveal a binding free energy of -62.3 kcal/mol.
We rather interpret the overestimation of the binding free energy obtained from our

advanced sampling simulations as an overstabilization in the force field. Thus, we have reweighted the trajectories from the first and from the last Umbrella window of the distance PMF shown in Fig 10.2. We have evaluated the energies for every frame of the second half of the trajectories using an implicit solvent model (gb=5). We then have calculated the free energy difference between both Umbrella windows, which yields a new, rough estimate for G_{restr}^{bind}:

$$\Delta F = -\frac{1}{\beta} \cdot \ln \left(\frac{\sum_i \exp(-\beta E_i^1)}{\sum_j \exp(-\beta E_j^{49})} \right), \tag{10.9}$$

where we sum over the frames i or j of the first and last (49th) window, respectively. This calculation yields $\Delta F = -77.6\,\text{kcal/mol}$ for the force fields used so far (bsc1 for DNA). If we describe the DNA with our Tumuc1 force field, we obtain a substantially lower binding affinity, $\Delta F = -65.1\,\text{kcal/mol}$. This finding indicates that the bsc1 force field overstabilizes the protein-DNA complex, and a more realistic binding affinity may be inferred with Tumuc1. We emphasize that this comparison remains vague, as we rely on an implicit solvent description and only reevaluate the energies of the trajectories. In order to compare the performance of the force fields precisely, the MD simulations should be repeated with the Tumuc1 force field. For small minor-groove-binding ligands, accurate absolute binding free energies have been determined in two MD studies, which followed a very similar approach as in our study [28, 29]. However, this does not necessarily mean that accurate binding free energies can also be obtained for substantially larger protein-DNA complexes with the Amber force fields. Indeed, You and coworkers have reported very recently that the sliding-diffusion of proteins along DNA is underestimated by a whole order of magnitude in the classical Amber force fields [8], and an overstabilization of Homeodomain-DNA complexes by $\sim 5-15\,\text{kcal/mol}$ has been found by means of standard Umbrella Sampling simulations [30].

Besides, we find a relatively small configurational contribution for the Cren7-protein, $G_c^{bulk,Cren7} = 0.93\,\text{kcal/mol}$, which reveals only little structural adaption of the Cren7-protein upon binding. For the DNA, instead, we obtain $G_c^{bulk,DNA} = 6.32 \pm 0.03\,\text{kcal/mol}$. This is caused by the strong bending deformation the DNA undergoes during binding of the Cren7 protein, and can therefore be directly related to estimates for DNA's deformation energy based on a harmonic model.

G_{restr}^{bind}	-33.89 ± 7.32
G_{a}^{site}	0.76 ± 0.02
G_{o}^{site}	0.57 ± 0.03
G_{o}^{bulk}	5.13
G_{c}^{site}	0.94 ± 0.00
$G_{c}^{bulk,DNA}$	6.32 ± 0.02
$G_{c}^{bulk,Cren7}$	0.93 ± 0.00

Table 10.3: Computed free energy terms. All values are in kcal/mol.

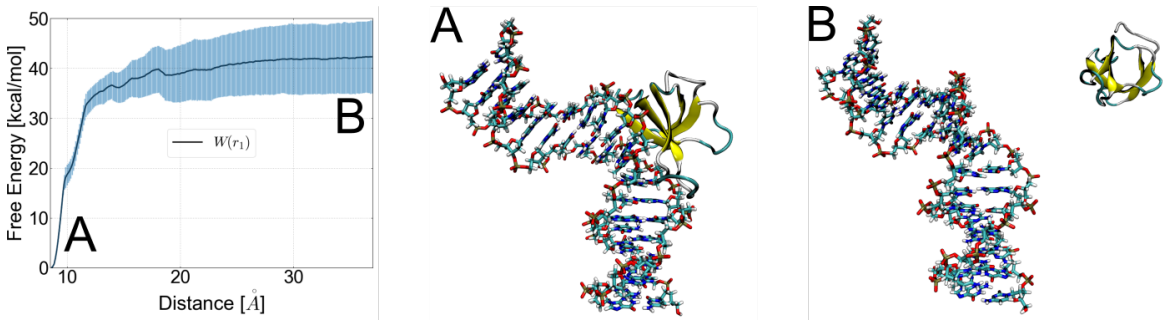


Figure 10.2: Computation of G_{restr}^{bind}. Left panel shows the distance PMF $W(r_1)$. Representative snapshots for the Cren7-DNA system are depicted for the first Umbrella window (A) and for the last Umbrella window (B).

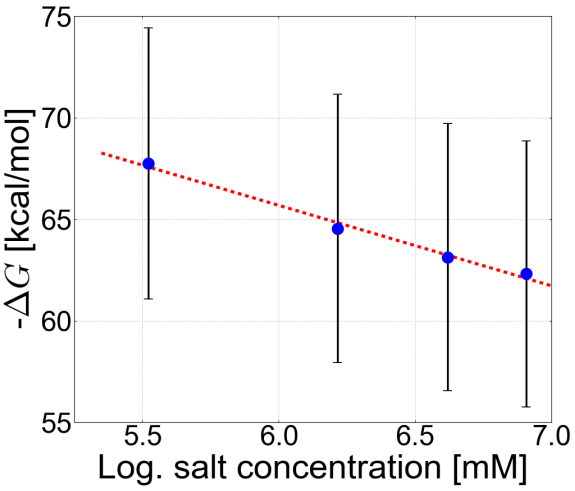


Figure 10.3: Binding free energy as a function of the logarithm of salt concentration. The binding free energies have been calculated with MM/GBSA.

10.3.2 Deformation Energy of the DNA upon Binding

We have studied the deformability of the DNA's central six base-pair steps, as they represent the binding site for the Cren7 protein (Fig 10.4 A). Here, we consider the three rotational degrees of freedom, which describe the rigid body transformation from the orientation of axis system A1 (composed of the x_1, y_1 and z_1 vectors) to A2 (defined by x_2, y_2 and z_2): $\mathbf{R} = \mathbf{A_2 A_1^T}$ [31]. In this way, we have found that binding of the Cren7 protein induces only very modest untwisting ($\sim 4.4\,\mathrm{deg}$), but pronounced bending (Fig 10.4 B). Thus, we limit the discussion on DNA's deformation energy to the two bending modes (ϕ, θ). We have extracted the stiffness matrix and equilibrium parameters from the 100 ns long MD simulation for the DNA molecule, the deformation energy is then given by [32–34]:

$$E_{def} = \frac{1}{2} \begin{pmatrix} \theta - \theta_0 & \phi - \phi_0 \end{pmatrix} \begin{pmatrix} k_{\theta\theta} & k_{\theta\phi} \\ k_{\phi\theta} & k_{\phi\phi} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \phi - \phi_0 \end{pmatrix}. \tag{10.10}$$

Based on this harmonic approximation we can quantify deformation energies as a function of θ and ϕ. We hence evaluate DNA's deformation energy upon binding by the Cren7 protein to 8.59 kcal/mol (see red cross in Fig 10.4). Note that this overestimates the configurational term for the DNA by more than $2\,\mathrm{kcal/mol}$ ($G_c^{bulk,DNA} = 6.32\,\mathrm{kcal/mol}$). This is probably due to a deviation from the quadratic behavior in the severely bent/kinked subspace. Nevertheless, the deformation energy profile in (θ, ϕ) -space (Fig 10.4 B) reveals two interesting results: First, bending by $\sim 25 - 30\,\mathrm{deg}$ can already increase the binding affinity by 2 kcal/mol. Second, the complex structure is very distant from the principal axis. In the complex, the DNA is actually bent into the stiffer direction (larger deformation of the θ-mode), which is reflected by a lower variance for θ, $\mathrm{var}(\theta) = 116.0\,\mathrm{deg}^2$ and $\mathrm{var}(\phi) = 145.5\,\mathrm{deg}^2$. This indicates an active mechanism of the Cren7 protein, but is also compatible with allosteric effects. Previous studies have indeed suggested a cooperative mechanism as a result of protein binding induced changes in DNA's global elasticity [6, 35].

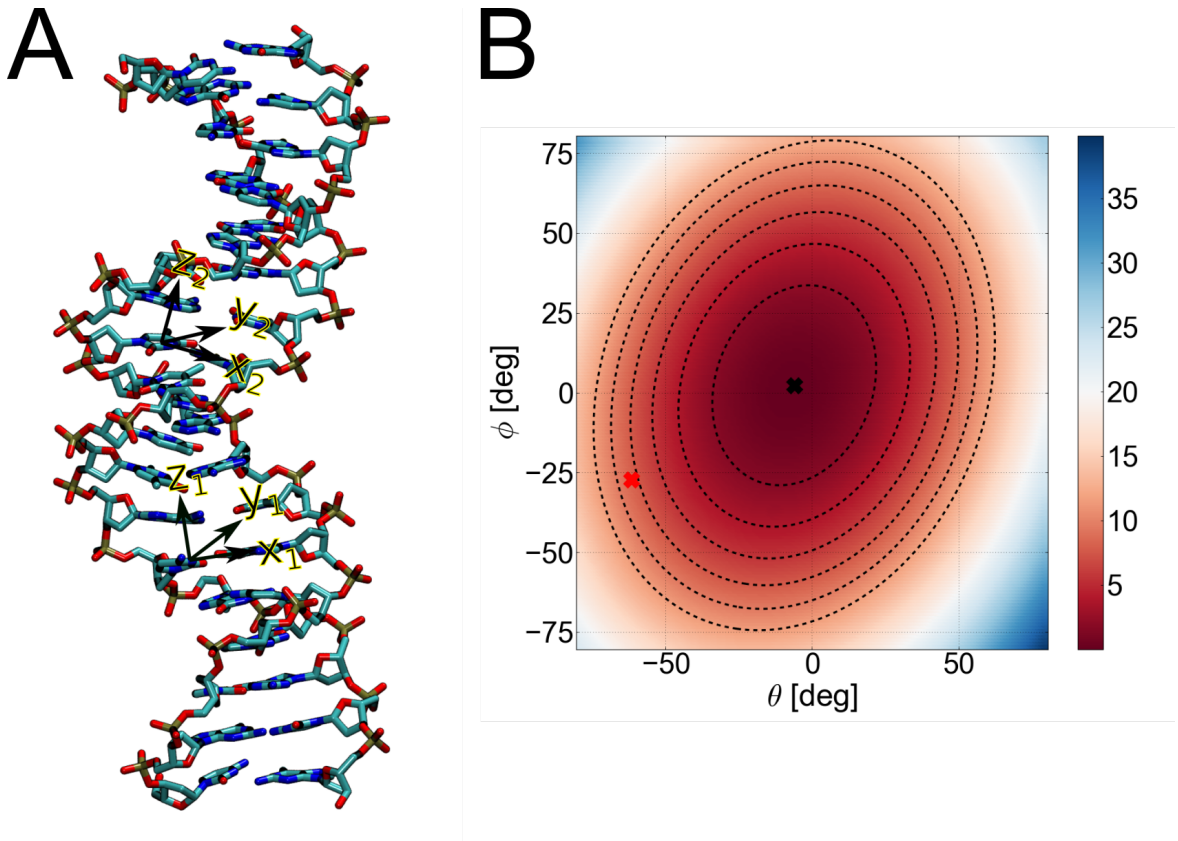


Figure 10.4: Deformability of DNA's binding site. (A) Two axis-systems (x_1, y_1, z_1) and (x_2, y_2, z_2) have been defined, which allowed us to compute bending parameters for the binding site. (B) Deformation free energy landscape of the binding site with respect to bending. Reference values for the DNA in unbound and in complex conformation are marked by a black and red cross, respectively. The deformation energy is shown color-coded in units of kcal/mol.

10.4 Conclusion

The Cren7 protein packages chromosomal DNA in archaea by inducing strong local bending in DNA's binding site [6, 18, 19]. Experimentally, binding free energies of ~ -7.0 to -9.4 kcal/mol have been measured with a slight preference for AT-rich sequences [18, 19]. In this chapter, we have quantified free energy contributions for binding of the Cren7 protein to DNA with advanced free energy simulations. Overall, we obtain a significant overestimation of the binding affinity, $G_{bind} = -23.8 \pm 7.3$ kcal/mol. We suspect that this demonstrates further deficiencies of the classical force fields. Reweighting the trajectory with the Tumuc1 force field for DNA makes the binding less stable by ~ 12.5 kcal/mol. Note that this represents a rough estimate, and by repeating simulations with Tumuc1 we will be able to give a more precise comparison. However, deficiencies of current force

fields for protein-DNA complexes have also been reported in previous studies [8, 20, 30]. You and coworkers, for instance, have shown that the sliding diffusion of PCNA along DNA is underestimated by a whole order of magnitude [8]. Thus, our observed overstabilization of protein/DNA complexes is conceivable.
In addition, we have quantified configurational contributions of the Cren7 protein and the DNA. Here, we have found only small conformational changes for the Cren7 protein in contrast to the DNA, for which we have obtained a configurational contribution of $G_c^{bulk,DNA} = 6.32\,\mathrm{kcal/mol}$. This term can be directly compared to a deformation energy computed from a harmonic model. From a harmonic model, we have estimated that bending the DNA into its complex structure requires a free energy cost of 8.59 kcal/mol, which overestimates the configurational term possibly due to deviations from the quadratic landscape for higher bending. Note that theoretical and experimental studies have proposed that protein binding alters DNA's elastic behavior, hence triggering cooperative binding effects [6, 35]. The investigation of such effects is out of scope for the present MD study. However, the analysis of DNA's deformability has shown that DNA is not bent along the principal axis. This indeed points to an active or cooperative mechanism of the Cren7 protein, as it means that DNA's elasticity itself is not determinant. Altogether, one can hence think of the binding process as a strong deformation of the genetic material, which is caused by a relatively rigid machine.
The large difference between experimentally and computationally determined binding free energy demands further focus on the methodology of our MD simulations. Though computationally expensive and no sampling issues were noted, the simulation times might be extended. Moreover, carrying out the advanced free energy simulation with Tumuc1 will elucidate the impact of the electrostatic description by the force field on the binding free energy. For the mechanistic understanding of the Cren7-DNA binding process, it would be interesting to include global restraints on the DNA during the MD simulations and hence quantify the binding free energy as a function of DNA's global topology (similarly to [9]).

10.5 Bibliography

[1] Remo Rohs, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.*, 79(1):233–269, 2010.

[2] Michael Y. Tolstorukov, Andrew V. Colasanti, David M. McCandlish, Wilma K. Olson, and Victor B. Zhurkin. A Novel Roll-and-Slide Mechanism of DNA Folding in Chromatin: Implications for Nucleosome Positioning. *Journal of Molecular Biology*, 371(3):725–738, 2007.

[3] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thstrm, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[4] Rosalie P. C. Driessen, He Meng, Gorle Suresh, Rajesh Shahapure, Giovanni Lanzani, U Deva Priyakumar, Malcolm F. White, Helmut Schiessel, John van Noort, and Remus Th. Dame. Crenarchaeal chromatin proteins Cren7 and Sul7 compact DNA by inducing rigid bends. *Nucleic Acids Res.*, 41(1):196–205, 2012.

[5] Howard Cedar. DNA methylation and gene activity. *Cell*, 53(1):3–4, 1988.

[6] Zhenfeng Zhang, Zhengyan Zhan, Bing Wang, Yuanyuan Chen, Xiuqiang Chen, Cuihong Wan, Yu Fu, and Li Huang. Archaeal Chromatin Proteins Cren7 and Sul7d Compact DNA by Bending and Bridging. *mBio*, 11(3), 2020.

[7] David Winogradoff and Aleksei Aksimentiev. Molecular Mechanism of Spontaneous Nucleosome Unraveling. *Journal of Molecular Biology*, 431(2):323–335, 2019.

[8] Seonju You, Hong-Guen Lee, Kimoon Kim, and Jejoong Yoo. Improved Parameterization of Protein-DNA Interactions for Molecular Dynamics Simulations of PCNA Diffusion on DNA. *J. Chem. Theory Comput.*, 16(7):4006–4013, 2020.

[9] Giuseppe LaRosa and Martin Zacharias. Global deformation facilitates flipping of damaged 8-oxo-guanine and guanine in DNA. *Nucleic Acids Res.*, 44(20):9591–9599, 2016.

[10] Loc Etheve, Juliette Martin, and Richard Lavery. Dynamics and recognition within a protein-DNA complex: a molecular dynamics study of the SKN-1/DNA interaction. *Nucleic Acids Res.*, 44(3):1440–1448, 2016.

[11] Danilo Roccatano, Andre Barthel, and Martin Zacharias. Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation. *Biopolymers*, 85(56):407–421, 2007.

[12] Lin Chen, Ji-Long Zhang, Li-Ying Yu, Qing-Chuan Zheng, Wen-Ting Chu, Qiao Xue, Hong-Xing Zhang, and Chia-Chung Sun. Influence of Hyperthermophilic Protein Cren7 on the Stability and Conformation of DNA: Insights from Molecular Dynamics Simulation and Free Energy Analysis. *J. Phys. Chem. B*, 116(41):12415–12425, 2012.

[13] Lin Chen, Qing-Chuan Zheng, and Hong-Xing Zhang. Insights into the effects of mutations on Cren7-DNA binding using molecular dynamics simulations and free energy calculations. *Phys. Chem. Chem. Phys.*, 17(8):5704–5711, 2015.

[14] Dhanusha Yesudhas, Muhammad Ayaz Anwar, Suresh Panneerselvam, Prasannavenkatesh Durai, Masaud Shah, and Sangdun Choi. Structural Mechanism behind Distinct Efficiency of Oct4/Sox2 Proteins in Differentially Spaced DNA Complexes. *PLOS One*, 11(1):e0147240, 2016.

[15] Dhanusha Yesudhas, Muhammad Ayaz Anwar, Suresh Panneerselvam, Han-Kyul Kim, and Sangdun Choi. Evaluation of Sox2 binding affinities for distinct DNA patterns using steered molecular dynamics simulation. *FEBS Open Bio*, 7(11):1750–1767, 2017.

[16] Changhao Wang, D'Artagnan Greene, Li Xiao, Ruxi Qi, and Ray Luo. Recent Developments and Applications of the MMPBSA Method. *Frontiers in Molecular Biosciences*, 4(87), 2018.

[17] Hyung-June Woo and Benoît Roux. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA*, 102(19):6825–6830, 2005.

[18] Zhenfeng Zhang, Yong Gong, Yuanyuan Chen, Hongbin Li, and Li Huang. Insights into the interaction between Cren7 and DNA: the role of loop β3-β4. *Extremophiles: life under extreme conditions*, 19(2):395–406, 2015.

[19] Zhenfeng Zhang, Mohan Zhao, Yuanyuan Chen, Li Wang, Qinghua Liu, Yuhui Dong, Yong Gong, and Li Huang. Architectural roles of Cren7 in folding crenarchaeal chromatin filament. *Mol. Microbiol.*, 111(3):556–569, 2019.

[20] Jejoong Yoo, David Winogradoff, and Aleksei Aksimentiev. Molecular dynamics simulations of DNA-DNA and DNA-protein interactions. *Current Opinion in Structural Biology*, 64:88–96, 2020.

[21] D.A. Case, R.M. Betz an D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. Kovalenko J. Kaus, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *Amber 16*. University of California, San Francisco, 2016.

[22] Zhenfeng Zhang, Yong Gong, Li Guo, Tao Jiang, and Li Huang. Structural insights into the interaction of the crenarchaeal chromatin protein Cren7 with DNA. *Molecular Microbiology*, 76(3):749–759, 2010.

[23] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[24] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[25] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jrgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep Lluis Gelp, Carlos Gonzlez, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55–58, 2016.

[26] Alexey Onufriev, Donald Bashford, and David A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55(2):383–394, 2004.

[27] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[28] Fabian Zeller and Martin Zacharias. Adaptive Biasing Combined with Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations. *J. Chem. Theory Comput.*, 10(2):703–710, 2014.

[29] Hong Zhang, Hugo Gattuso, Elise Dumont, Wensheng Cai, Antonio Monari, Christophe Chipot, and Franois Dehez. Accurate Estimation of the Standard Binding Free Energy of Netropsin with DNA. *Molecules*, 23(2):228, 2018.

[30] David Jakubec and Jiří Vondrášek. Can All-Atom Molecular Dynamics Simulations Quantitatively Describe Homeodomain-DNA Binding Equilibria? *J. Chem. Theory Comput.*, 15(4):2635–2648, 2019.

[31] Korbinian Liebl and Martin Zacharias. How global DNA unwinding causes non-uniform stress distribution and melting of DNA. *PLOS ONE*, 15(5):e0232976, 2020.

[32] O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks. A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.*, 138(5):055102, 2013.

[33] Tomáš Geršata, Alberto Pérez, Modesto Orozco, Alexandre V Morozov, Jiří Šponer, and Filip Lankaš. Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, 9(1):707–721, 2013.

[34] Tomáš Geršata and Filip Lankaš. Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3(4):355–363, 2013.

[35] Joseph Rudnick and Robijn Bruinsma. DNA-Protein Cooperative Binding through Variable-Range Elastic Coupling. *Biophysical Journal*, 76(4):1725–1733, 1999.

So eine Arbeit wird eigentlich nie fertig, man muss sie für fertig erklären, wenn man nach Zeit und Umständen das Möglichste getan hat.

- Goethe, Italienische Reise

11 Outlook

Within my PhD thesis, I have used Molecular Dynamics simulations, Quantum Mechanical calculations and other theoretical concepts to explore the mechanics of DNA. In the beginning of my PhD, I was much concerned with how torsional stress affects DNA's structure, what is the impact of global restraints, what sort of phase transition does it trigger and does the stress at a certain level localize at specific sequences? Here, we have clearly seen the importance to study DNA's mechanics at atomistic resolution. The switching between local BI and BII states regulates to a far extent DNA's deformability. This results from a sequence-dependent population of the backbone states, a phenomenon for which we have been able to give a far-reaching atomistic explanation: Sterical clashing between thymine's methyl group and the C2' atom with its hydrogens leads to a significant stabilization of the BI conformation.
In chapter 7 we have accounted for the role of the backbone states for DNA's mechanics. We have derived a new method to compute deformation free energies for double-stranded DNA structures: The DNA is described as an Ising model, whereby each base-pair step is discretized to be either in the BI or BII conformation. Each configuration is approximated by a multivariate harmonic ansatz and we thereby account for nearest-neighbor coupling. Thus, we have been able to develop the first theory for DNA's elasticity which captures both local effects, nearest-neighbor coupling and multimodality. The success of our model is not only shown by the evaluation of single-molecule MD simulations for DNA but also demonstrated in protein-DNA systems: With our Ising model we find excellent agreement between experimentally obtained binding affinities and computed deformation energies for the papillomavirus E2-protein/DNA complexes, whereas the standard harmonic model falls behind.
However, I hypothesize that by focusing on atomistic interactions we may lose the essence of some cellular processes. In the past years, much has been published on DNA-protein interactions from the MD community, many of them deal with the binding of repair enzymes to DNA damages. And here, we always do the same: Take the crystal structure of the protein bound to a small fragment of DNA and let our computers go. But maybe things work out entirely differently in our cells as there is much more allosteric regulation than we can anticipate right now? For the recognition of DNA damages, for example, I imagine the following: Enzymatic processes like transcription and replication generate high torsional stress for the DNA, which leads to supercoiling. Now, where does the tip of the supercoil localize? The tip of the supercoil is bent by approximately 180° and hence absorbs the largest amount of mechanical stress. Thus, the softest sequence, which most likely is a damage if present, should be in tip position. Note that this mechanism underlies the same principle as TATA-box melting which we have elaborated in

chapter 6. The DNA damage would then be structurally exposed and in addition, the compacted supercoil may provide a distinct electrostatic signal for enzymes. This represents one of many scenarios why I think that a combination of coarse-grained modeling and atomistic simulation is an important future step. In this way, we would be able to study cellular processes in atomistic detail while capturing long-range, allosteric effects. The plausibility of such effects has also been discussed in chapter 10

Another point of concern is the theory we employ in our MD simulations, the force fields. In my opinion, the quality of force fields has not been examined sufficiently. To me, this is not very understandable, as the parameterization methods behind them are in large parts clearly outdated, which is enough reason to adopt a critical perspective on the standard force fields. Bsc1 represents the most prominent DNA force field. In chapter 8, we have revealed several problems of this force field. My concerns about current force fields and recent development approaches have motivated me to construct a new DNA force field based on QM calculations. By the end of my PhD, a first version has been completed. The force field is named after my alma mater, 'Tumuc1' for Technical University of Munich. We have tested it on several different DNA systems: For B-DNA, we obtain good agreement with experimental data, and also more complex events such as hairpin folding or hybridization of single strands have been captured accurately. Despite its good performance so far, I do not think that it can be the final solution. The final solution that is just what we are constantly striving for - the way is the aim. In the coming years, we will closely follow the performance and reputation of Tumuc1 and by time take action and refine it. A central aspect in the development of Tumuc1 has actually been to generate a description which is easily modifiable and does not have to suffer from Amber's tight corset of generalization. Besides, it is my very ambition that Tumuc1 will also enter the RNA and protein field. Thereby, we will greatly benefit from the techniques and experiences we have developed during parameterization of the DNA force field. However, in the future, we shall go to more complex force fields, which include polarization effects. By now, polarization effects are generally not considered in MD studies, but in this way, we neglect interactions that may be central to certain events. For instance, methylation of pyrimidines at the 5' position, which represents a major epigenetic modification, increases the polarizability. It is therefore of interest to clarify in how far polarization effects can alter nucleosome stability and affect protein-DNA binding in general.

My outlook is very optimistic. The great advances in hardware provide us with plenty of possibilities for the future. We can simulate longer timescales in MD, and perform QM calculations more efficiently. What we need to do now is to improve our theories and develop new methods.

Wenn ich weiter geblick habe, so deshalb, weil ich auf den Schultern von Riesen stehe.

- Isaac Newton

Acknowledgments

First of all, special thanks are due to Prof. Dr. Martin Zacharias, whose advice and doings for me greatly exceed the role of a supervisor. The countless discussions about ideas and academic perspectives were very interesting, helpful, encouraging and, very often, quite funny. This put me not only into the position of (hopefully) obtaining a PhD and learning very much, but also constantly reinforced my curiosity and made me overcome pitfalls straightforwardly. Martin, that you always just want the best for others in an honest way and that you always have an open door leaves me with a great impression.
Sonja: Thank you for all your help on administrative stuff in your always direct and friendly spirit, and for the organization of the winterschools. Your efforts for the staff are far beyond the matter of course.
I pay tribute to all proofreaders, as well as current and former members of the T38 chair. Many thanks to you, Manu, Max, Fabi, Flo, Danial, Till, Brianda, Julian, Richard, Arien, Shu-Yu, Johannes and Paul for the good atmosphere at the chair and free-time activities. Thanks are due to Jonathan for his services concerning the cluster, and to Uli for joining forces on the blunt-end project.
I emphasize my gratitude to Asmar for being a reliable (office-)mate. The daily exchanges on science and plenty of other stuff were very instructive and entertaining alike. Moreover, I thank you for your soccer-initiative and the cooperation on the Caffeine-project.
Thank you, my dear T38-folks, I will look back to my time at this chair with great memories. Working here was not a burden, but a great enrichment. The same I wish also for you.
I express my thanks to the initiators of the SFB749 for funding, and my alma mater for the possibility to do my PhD here.
For the many good times outside of science, I thank my two brothers Quirin and Hansi, and my closest friends: Bene, David, Dole, Miche, Nik, Gabi, Vinc and Zize - Thank you!
To my parents: Thank you for all your doings for me. I am very grateful for your support in many ways, which allowed me to follow my passions and pursue my ambitions.

Finally, thank you, dear reader, for your interest in my lines.

Thank you all!