Comparison of DNA force fields at a multiscale level : towards bottom-up modelling of chromatin

Minhas, Vishal

2020

Minhas, V. (2020). Comparison of DNA force fields at a multiscale level : towards bottom-up modelling of chromatin. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/142499>

<https://doi.org/10.32657/10356/142499>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 17:56:18 SGT

Comparison of DNA force fields at a multiscale level: towards bottom-up modelling of chromatin

Vishal Minhas

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University in fulfillment of the requirement for the degree of Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Aug. 2020

.....

Date

.....

Vishal Minhas

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Aug. 2020

.....

Date

.....

Prof. Lars Nordenskiöld

Authorship Attribution Statement

This thesis contains material from [2] paper published in the following peer-reviewed journals in which I am listed as an author.

Section 3.1 and 3.2 of Chapter 3 is published as [V. Minhas, S. Tiedong, A. Mirzoev, N. Korolev, A.P. Lyubartsev and L. Nordenskiöld, Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels. J. Phys. Chem. B 124\(1\), 38–49 \(2020\). DOI: 10.1021/acs.jpcb.9b09106.](#)

The contributions of the co-authors are as follows:

- I prepared the initial manuscript draft. The manuscript was revised by Dr Korolev, Prof. Lyubartsev and Prof. Nordenskiöld.
- Dr. Mirzoev and Dr. Tiedong assisted in the collection and analysis of the trajectory data.
- Prof. Lyubartsev and Dr. Korolev assisted in the interpretation of the data.

Section 4.1.3, 4.1.4 and 4.2 of Chapter 4 is published as [S. Tiedong, A. Mirzoev, V. Minhas, N. Korolev, A.P. Lyubartsev and L. Nordenskiöld, A multiscale analysis of DNA phase separation: From atomistic to mesoscale level . Nucleic Acids Res. 47\(11\), 5550-5562 \(2019\). DOI: 10.1093/nar/gkz377.](#)

The contributions of the co-authors are as follows:

- Prof. Nordenskiöld provided the initial project direction and edited the manuscript drafts.
- I helped in preparing the manuscript drafts. The manuscript was revised by Dr Korolev, Prof. Lyubartsev and Prof. Nordenskiöld.
- Dr. Mirzoev assisted in the collection and analysis of the trajectory data.
- Prof. Lyubartsev and Prof. Nordenskiöld assisted in the interpretation of the data.

June. 2020

.....

Date

.....

Vishal Minhas

Acknowledgements

Firstly, I would like to thank Prof. Lars Nordenskiöld for the guidance and direction he gave me throughout the project. Also I would like to thank him for the generous help and encouragement he gave me to present my work at multiple conferences.

I am really grateful to Dr. Nikolay Korolev for productive discussions and constant guidance in the project. I thank Prof. Alexander Lyubartsev for his expertise on molecular modelling and computational theory. I am indebted to Dr. Sun Tiedong and Dr. Alexander Mirzoev for their constant assistance and inspiration during the project. This work wouldn't have been possible without their teamwork. Thank you to Assoc. Prof. Lu Lanyuan for giving constructive feedback on my research work.

I extend my gratitude to other members of the group including Chinmayi, Dr. Aghil Soman, Dr. Chen Qinming and Dr. Nikolay Berezhnoy for always being available for help and fruitful discussions on relevant topics. I thank my girlfriend Prachi, for her constant love and support.

Lastly I thank School of Biological Sciences, Nanyang Technological University for the great opportunity to pursue my graduate studies. The author is also heavily indebted to National Supercomputing Centre Singapore for providing computational resources.

To my dear family

Abstract

Computer simulations were performed to understand the dynamics of DNA molecules in presence of mono-valent cations. These simulations were run using different all-atom DNA force fields to study their behaviour at a multiscale level, from the base-pair to the mesoscopic level. Serious problems were found with the latest version of the CHARMM family of force fields, namely CHARMM36, as it distorted DNA structure and backbone at the microsecond timescale. Also, to achieve the long-term goal of modelling the folding of the chromatin fiber, we have built a multiscale model using systematic bottom-up approach based on structure based coarse-graining. At the first stage of bottom-up coarse-graining, all-atom and meso-atom representations of all principal chromatin components were bridged. To do this, the all-atom molecular dynamics trajectories obtained were bead-mapped to a coarse-grained representation and used for the calculation of the radial distribution functions, which were then used to derive the effective potentials using the Inverse Monte-Carlo method. These potentials were then validated in a coarse-grained molecular dynamics simulation of the original system. Based on these potentials, an excellent prediction of the dependence of DNA persistence length on salt concentration, in agreement with experiments was obtained, which validates the underlying CG-model. Once all the potentials are obtained for all the interactions, a large-scale CG simulation will be run using these potentials to model the NCP.

Contents

Acknowledgements	vii
Abstract	xi
List of Figures	xv
List of Tables	xxi
1 Introduction	1
1.1 DNA structure and compaction	1
1.1.1 Geometry of DNA	1
1.1.2 Higher levels of compaction	4
1.2 Factors governing DNA compaction	7
1.2.1 Role of counterions	7
1.2.2 Histone proteins and their tails	9
1.2.3 NCP-NCP stacking	10
1.2.4 Persistence length of DNA	11
1.3 Computational modelling of DNA compaction	12
1.3.1 All atom simulations	13
1.3.2 Selection of force fields	14
1.3.3 Coarse-grained modelling	15
1.4 Objective of the present work	22
2 Materials and methods	25
2.1 Introduction: the multiscale approach	25
2.1.1 Challenges	28
2.2 All-atom simulations	29
2.2.1 Single DNA system	29
2.2.2 Multiple DNA and histone tails systems	32
2.3 Coarse-grained models	36
2.3.1 DNA	36
2.3.1.1 Improvement of the model	38
2.3.2 Amino acids	38

2.4	RDF calculation for DNA systems	40
2.5	Inverse Monte Carlo procedure	41
2.5.1	Derivation of CG effective potentials	42
2.6	Coarse-grained simulations	44
2.6.1	Persistence length calculation	44
2.7	Data analysis	45
3	Results and Discussion: all atom simulations	47
3.1	DNA backbone and helical properties	48
3.1.1	A-DNA versus B-DNA	51
3.1.1.1	Parmbsc0 has a preference for A-DNA conformations	54
3.1.2	Torsional angles	56
3.1.3	Elastic properties at all-atom level	60
3.2	CHARMM36 distorts backbone DNA structure	61
3.3	Conclusions	66
4	Results and Discussion: coarse-grained simulations	69
4.1	Effective potentials using IMC	70
4.1.1	Reproduction of the reference RDFs	73
4.1.2	Coarse-grained MD simulations	74
4.1.2.1	Improvement of the CG model	75
4.1.3	Comparison of force fields	77
4.1.4	Comparison with other CG models	79
4.2	Conclusions	81
5	Ongoing and future work - large-scale NCP simulations	83
5.1	DNA - histone tail potentials	84
5.2	Large-scale CGMD simulations - preliminary results	85
6	Conclusions	91
	Appendices	93
A	Appendix	95

List of Figures

1.1	<i>Left:</i> DNA condensation depicted at different levels. <i>Right:</i> Chemical structure of the DNA nucleotides with the accompanying sugar bases.	2
1.2	Parameters describing DNA geometry	3
1.3	Illustrated figure showing the dimers involved in the histone core proteins and how they form the tetrameric structure. Also the handshake is shown for both the dimers. The final NCP structure is shown on the right. (Image courtesy: Zephyris (Richard Wheeler) and the Wikipedia)	6
1.4	Different models proposed for the folding of the chromatin fiber with different nucleosomal repeat lengths as signified by the number in white	7
1.5	Different DNA CG models developed. (A) 3SPN.2 model, (B) MARTINI CG FF, (C) Savalyev et. al. CG model and (D) Naome et. al. model.	19
2.1	Multiscale modelling of DNA compaction	26
2.2	Flowchart explaining multiscale modelling using IMC	27
2.3	Harmonic bonds added on terminal base pairs	31
2.4	Illustration of how the 48 fragments were derived from histone tails	34
2.5	A snapshot of the 8DNA with histone tail fragments simulation . .	35
2.6	Initial coarse-grained model used. A) shows the 5 beads that define the D bead and the four P beads connected to it. B) shows the bonds defined in the model with the equilibrium bonds and angles. C) A superposition of the all atom model and the coarse-grained model.	37
2.7	Illustration showing the updated DNA CG model used throughout this work. (A) Representation at the all-atom level, (B) superposition of all-atom and coarse-grained model. (C) Bonds and angles defined in the model, including the P-P bond added along the minor groove. Reproduced with permission from [162]	39
2.8	Schematic showing the bonds and the three different angles, ϕ , ψ and χ defined in the coarse-grained model for amino acids. P refers to Proline, G for Glycine and Others for all the other amino acids . . .	40

2.9	The 500bp DNA molecule used to calculate the persistence length for all the force fields and salt concentrations. The different coloured pieces show the 5 different fragments over which the calculation was done. Ions have not been shown for clarity.	46
3.1	Parameters characterizing base-pair geometry for simulations with 40-mer DNA with 130 mM KCl using four different force fields, (A) parmbsc0, (B) parmbsc1, (C) CHARMM27 and (D) CHARMM36. Y-axis shows the magnitude of each parameter in degrees. Colored bands show the corresponding range of values for crystal structures of B-DNA. Also the corresponding color-coded parameters are shown on the side.	48
3.2	Comparison of four different base-step and base-pair parameters. Variation of (A) shift, (B) tilt, (C) shear and (D) buckle. Black line and grey band in each of the panels show an average value and a standard deviation panel calculated from the DNA crystal structures. Simulation data for each base-step/base-pair was averaged over the entire productive run.	49
3.3	Comparison of four different base-step and base-pair parameters. Variation of (A) stretch, (B) propeller, (C) slide and (D) roll. Black line and grey band in each of the panels show an average value and a standard deviation panel calculated from the B-DNA crystal structures. Cyan line shows the average value of slide for A-DNA. Simulation data for each base-step were averaged over entire productive run.	50
3.4	Comparison of four different base-step and base-pair parameters. Variation of (A) stagger, (B) opening, (C) rise and (D) twist. Black line and grey band in each of the panels show an average value and a standard deviation panel calculated from the DNA crystal structures. Simulation data for each base-step were averaged over entire productive run.	51
3.5	(A) Differentiation between A-form and B-form DNA based on slide. A value of -0.8 (cyan line) defines the boundary between the A-form and B-form. Parmbsc0 (red) is more biased to A-form compared to other force fields. Black line and shaded area show respectively the average value and its standard deviation determined from the crystal structures of B-DNA. (B) The average distribution for the AG, (C) GG and (D) TA base-step for the four force fields.	52
3.6	Variation of different helical parameters over the first 1 ns. (A) shows the Inclination, (B) X-displacement and (C) shows the phase angle of pseudo rotation. In all the plots the A-DNA average value is shown in cyan and B-DNA value in grey. (D) shows the RMSD variations with respect to A-DNA starting structure (solid lines) and B-DNA starting structure (dotted lines).	53

3.7	(A) Average value of slide for two parmbsc0 trajectories for each individual base-step in DNA 1. Black line is the average value from crystal structures of B-DNA and cyan line the average for A-DNA crystal structures. (B) shows the slide values for DNA2, (C) DNA3 and (D) DNA4. Each figure shows the values for both the trajectories for parmbsc0, Traj1 and Traj2.	55
3.8	(A) and (B) show the average values of slide for three CHARMM27 trajectory for each individual base-step in DNA 1 and DNA three respectively. Black line is the average value from crystal structures of B-DNA and cyan line the average for A-DNA. (C) and (D) show the slide values for each CHARMM36 trajectory for DNA 1 and DNA 3 respectively.	56
3.9	Slide distributions for 10 unique base steps present in the four DNA sequences for the three force fields analyzed. For each base-step, the value of slide was averaged over all the trajectories for that particular force field.	57
3.10	Distribution of dihedral angles parameters calculated from the all-atom MD trajectories of all-atom MD simulations of the 40 bp DNA with four force fields (indicated in the graphs); (A) α ; (B) β ; (C) δ ; (D) ϵ ; (E) γ and (F) ζ . Four dihedral angles were analyzed for the four force fields and compared them with the average values determined from the DNA crystal structures (black vertical line in each graphs). Simulation data for each dihedral angle was averaged for each DNA nucleotide over the entire productive run.	58
3.11	The percentage of BI and BII conformations for each of the 15 base-steps in the DNA sequence for strand 1 (A) and strand 2 (B). The grey line and band show the experimental values for the corresponding base-steps.	60
3.12	Comparison of helical flexibility for the four force fields. (A) Twist rigidity modulus [pN.nm ²], (B) stretching modulus [pN] and (C) twist-stretch coupling [pN.nm]. Experimental values with their standard deviations are shown respectively as black line and shaded area.	61
3.13	Distribution of RMSD for the CHARMM36 based trajectory Traj1 for two fragments of the trajectory, 100 ns-1 μ s (orange curve) and 1 μ s - 2.5 μ s (blue curve). Other curves are also shown for which 2 (khaki) and three (brick-red) residues have been excluded for the RMSD calculation. The dark-red curve has reference structure at t=100 ns.	63
3.14	(A) and (C) show the distribution of two base-step parameters, namely rise and twist respectively for the four CHARMM36 trajectories. (B) and (D) show the variation of these parameters for the whole 40-bp DNA molecule averaged over the whole trajectory.	63

3.15	Comparison of helical flexibility for the four C36 trajectories. (A) Twist rigidity modulus [$\text{pN}\cdot\text{nm}^2$], (B) stretching modulus [pN] and (C) twist-stretch coupling [$\text{pN}\cdot\text{nm}$]. Experimental values with their standard deviations are shown respectively as black line and shaded area.	64
3.16	Base-pair opening for CHARMM36-based trajectory using 130 mM KCl and a 40 bp long DNA (red points), and DNA snapshots showed in ‘New Ribbon’ representation for (A) $2.5\mu\text{s}$ long Traj1 and (B) $1.2\mu\text{s}$ long Traj2. The transparent band shows the range of the crystal structures for base pair opening.	65
3.17	Variation of base-pair opening over $1.2\mu\text{s}$ trajectory of the all-atom MD simulation with the CHARMM36 FF (40 bp DNA, 130 mM NaCl, Traj3 (A) and Traj4 (B); see Table 2.1). Light-red shaded area shows the range of the values observed in the DNA crystal structures.	65
4.1	Comparison of intra-molecular DNA RDFs obtained from four different underlying all-atom force fields. These are based on the all-atom trajectories using single 40 bp DNA with KCl salt concentration of 130 mM for different force fields discussed in the previous chapter. .	70
4.2	Comparison of intra-molecular DNA potentials obtained by using the IMC procedure for the RDFs mentioned in previous Figure 4.1. These are the final converged potentials.	71
4.3	Comparison of DNA-ion and ion-ion RDFs obtained from four different underlying all-atom force fields.	71
4.4	Comparison of DNA-ion and ion-ion potentials obtained from RDFs (4.3) of four different underlying all-atom force fields.	72
4.5	Minor groove width variation averaged over the full $2\mu\text{s}$ trajectories (A), and (B) shows the distribution of the minor groove width for all all force fields averaged for all the base-steps over the full trajectory. .	72
4.6	A comparison of the different RDFs obtained from the underlying all-atom simulation (blue), reproduced RDF at coarse-grained level is shown in green and the IMC sampled RDF for the final iteration is also shown in red. A good agreement is seen in all the RDFs characterizing the 1DNA system, thus validating the IMC procedure for this system.	73
4.7	Contour length L_C versus $\ln(\langle \cos\alpha \rangle)$ for 100 bp fragments of DNA .	75
4.8	(A) Schematic showing the first generation CG model. Adapted from [44] (Figure 3B). (B) Dependence of DNA persistence length on the salt concentration, I (mM) of KCl obtained from CGMD simulations based on the first generation CG model. The effective potentials were obtained from corresponding all-atom simulation of a single 40 bp DNA molecule using 130 mM KCl.	76

4.9	(A) A schematic showing the improved CG model with the added P-P bond along the minor groove of DNA. For the corresponding RDF and the potentials refer to Figure 4.3. Reproduced with permission from [162] (Figure 1D) (B) Dependence of DNA persistence length on the salt concentration, I (mM) of KCl obtained from CGMD simulations based on the first generation model (black solid circles) and the improved generation CG model (black hollow circles. The effective potentials were obtained from corresponding all-atom simulation of a single 40 bp DNA molecule using 130 mM KCl. The grey band corresponds to the experimental range of values.	77
4.10	Dependence of DNA persistence length on monovalent salt concentration. Experimental studies are depicted in solid coloured bands and their smooth fitting shown in the grey band, while the results obtained using the updated CG model with IMC are shown in solid coloured circles.	79
4.11	Dependence of DNA persistence length on monovalent salt concentration. Experimental studies are only depicted using a smooth fitting in the grey band, while the results obtained using the updated CG model with IMC are shown in black and blue solid circles. Results from other CG models are shown as coloured circles with white fill.	80
4.12	DNA phase separation into toroids using a super CG DNA model, based on the coarse-grained model presented above. Reproduced with permission from [162]	82
5.1	The CG model used for the amino acids illustrated for one of the 48 tail fragments used in the system	84
5.2	Non-bonded potentials obtained between DNA-DNA beads (top row) and DNA amino acid bead types (bottom two rows)	85
5.3	Illustrative figure showing from which underlying MD sub-system each pair-potential is extracted from the CGMD simulation. The central yellow box is the large-scale CGMD system, (with Mg and CoHex) and the five MD systems are shown using the rectangular boxes. The arrows connect each pair-potential in the CGMD setup to the corresponding underlying MD system from which it is extracted.	86
5.4	Non-bonded potentials obtained between aa beads are shown in the first four rows. The five categories of aa beads are shown on the top of each plot, POS - Positive, NPL - Non-Polar, GLY - Glycine, POL - Polar and NEG - Negative. For eg. POS-NEG means the potentials between positive and negative aa beads. The bottom row shows the bonded potentials defined for the aa beads. Angle 1 refers to all angles with Glycine in the center, angle 2 with proline in the center and angle three for any other amino acid in the center. . . .	88
5.5	Non-bonded potentials obtained between aa beads and ions.	89

5.6	<i>Left:</i> CGMD simulation setup showing 20 NCPs in a 55 nm box. <i>Center:</i> Two insets of a single NCP structure in two orthogonal perspectives. <i>Right:</i> Snapshots showing the aggregated cluster of NCPs after 100 millions steps (≈ 500 ns) in two perspectives. The tails have not been shown to see the aggregation of NCPs clearly. .	89
A.1	All-atom MD structures averaged over the full trajectory for all the 4 force fields.	96
A.2	The percentage of BI and BII conformations shown for each of the individual bases in the DNA sequence for strand 1 (A) and strand 2 (B) for the four force fields. The grey line and band show the experimental values for the corresponding base-steps.	97
A.3	Variation of three base-pair step parameters shown for the CoHex 4DNA system for 3 different trajectories using the C36 FF. As can be seen the distortion of DNA structure is independent of the sequence as even DNA3 and DNA4 show unphysical values for these parameters. For the DNA sequences used see Section 3.1	98
A.4	Distribution of the 6 base-pair parameters for the 1DNA CHARMM36 trajectories	99
A.5	Distribution of the 6 base-step parameters for the 1DNA CHARMM36 trajectories	100
A.6	Distribution of the 6 base-step helical parameters for the 1DNA CHARMM36 trajectories	101
A.7	Contour length L_C versus $\ln(\langle \cos \alpha \rangle)$ for 4 fragments of CGMD trajectory	102
A.8	Block averaging of bonded RDFs obtained over 500 ns of trajectory for the DNA tails system	103
A.9	Block averaging of DNA-DNA and DNA-aa non-bonded RDFs obtained over 500 ns of trajectory for the DNA tails system	104
A.10	Block averaging of aa-aa non-bonded RDFs obtained over 500 ns of trajectory for the DNA tails system	105
A.11	Block averaging of aa - ions non-bonded RDFs obtained over 500 ns of trajectory for the DNA tails system	106
A.12	The classification of angle bonds of all other amino acids into three groups and the averaged plot is shown in the lower right corner. . .	107
A.13	Comparison of monovalent ionic potentials obtained from 5 different underlying all-atom MD systems.	108
A.14	Comparison of DNA bonded obtained from 2 underlying all-atom MD systems.	109

List of Tables

2.1	Summary of all-atom simulations	31
2.2	Parameters used for the harmonic constraints imposed on the terminal atoms	32
2.3	Amino acid concentration	34
3.1	Percentage of BI/BII conformations for strand 1 (strand 2)	59
3.2	Percentage of BII conformations for each base-step	59
4.1	Coarse-grained MD simulations, DNA length = 500 bp	75
4.2	L_P for 100 bp fragments	75
4.3	DNA L_p results based on four underlying all-atom force fields	79
A.1	48 fragments used in tail simulations	95
A.2	L_P for different fragments of CGMD trajectory	102

List of Abbreviations

aa	Amino acid
bsc1	parmbsc1
bsc0	parmbsc0
C36	CHARMM36
C27	CHARMM27
CG	Coarse-grained
CGMD	Coarse-grained molecular dynamics
dsDNA	Double strand DNA
FF	Force field
GLY	Glycine
GROMACS	GRoningen MACHine for Chemical Simulations
HO	Histone Octamer
HPC	High Performance Computing
IMC	Inverse Monte Carlo
MC	Monte Carlo
MD	Molecular Dynamics
NEG	Negative
NCP	Nucleosome core particle
NPL	Non-polar
NSCC	National Supercomputing Centre Singapore
PBC	Periodic Boundary Conditions
PME	Particle Mesh Ewald
PMF	Potential of mean force

POL	Polar
POS	Positive
PPPM	Particle-Particle Particle-Mesh
RDF	Radial Distribution Function
RMSD	Root Mean Square Deviation
SAXS	Small Angle X-ray Scattering
ssDNA	Single strand DNA
VMD	Visual Molecular Dynamics

Chapter 1

Introduction

DNA is the blue-print of life. It contains all the information required to keep the biological world moving and thus keep propagating the cycle of life. Every cell in a single organism contains the same DNA and by some estimates the total amount of DNA in a single human being is enough to go round the sun and the earth several hundred times [3]. So packing this tremendous amount of DNA in a single cell with a radius of the order of a micrometer is a complex puzzle nature presents to us. What is the level and mechanism of compaction? What are the forces governing the packing and unpacking of DNA during various biological processes? These are very poorly understood areas in Biophysics. So understanding the way how nature packs a negatively charged polymer like DNA into a tiny cell becomes a very pertinent and interesting question as it is central to most of the transcription and replication activities. A pictorial representation of DNA compaction in eukaryotes is shown in Figure 1.1.

1.1 DNA structure and compaction

1.1.1 Geometry of DNA

DNA is a double helical molecule. Each strand is made of a sequence of four nucleotides, adenine (A), guanine (G), cytosine (C) and thymine (T). The second strand is made up of the nucleotides complementary to the first strand as depicted in Figure 1.1 (Image courtesy: [Madeleine Price Ball](#) and the Wikipedia). This

FIGURE 1.1: *Left*: DNA condensation depicted at different levels. *Right*: Chemical structure of the DNA nucleotides with the accompanying sugar bases.

complementary pair is called a base-pair and can be either G-C or A-T [177]. These base pairs are bonded together with the help of hydrogen bonds between the adjacent nucleotides, 3 hydrogen bonds for the G-C pair and 2 hydrogen bonds for the A-T pair. In addition, the stacking interaction between two consecutive nucleotides on the same strand plays a significant role in stabilizing the DNA molecule as a whole [190]. The topology of these base-pairs heavily determines the flexibility of the DNA molecule as a whole, which in turn influences how compaction occurs at the nucleosome and chromatin level.

At a more detailed level, the geometry of DNA can be characterized by the arrangement of each complementary base-pair and how two consecutive base-pairs are arranged relative to each other. These parameters are very important in determining the overall flexibility and the helical structure of DNA [40]. All the 16 parameters are depicted in Figure 1.2 (Image reproduced with permission from [98]).

In nature, DNA is known to exist primarily in 3 conformations, B-DNA, A-DNA and Z-DNA, the first two being the most prominent. B-DNA differs from A-DNA in some structural features, mainly rise, twist, roll and slide. The A-DNA form is also wider as the base-pairs are stacked off-centre in the A-conformation, whereas in B-form they are stacked nearly on top of each other. This also makes the A-form more rigid than B-DNA [170]. The conformation change from B-DNA to A-DNA involves a lowering of twist, increase of roll and a decrease of slide. Similarly among

FIGURE 1.2: Parameters describing DNA geometry

the helical parameters, X-displacement plays an important role in differentiating A-DNA from B-DNA [131]. Another major difference in B-DNA and A-DNA is the conformation of the backbone deoxyribose sugar ring, which is C2' endo for B-DNA and C3' endo for A-DNA.

In the context of protein binding, for B-DNA as the minor groove and major groove are distinct, most sequence specific protein binding occurs in the major groove. Proteins that bind nonspecifically bind DNA in the minor groove, which is the same size as the major groove in A-DNA. Another study by Harvey group [66] postulates that A-B DNA transitions provide the driving force for genome packing in DNA bacteriophages.

To model DNA flexibility at a mesoscopic level using base-pair level data, Olson group [132] used the values of the inter-base pair parameters to derive elastic constants. They fitted the helical parameter distribution of twist, rise, roll, slide, shift

and tilt to simple Gaussian functions to derive their model. It has been observed that twist and rise influence the overall form of the nucleic acid, roll and tilt play a role in bending. Also, X-displacement is related to the groove depth.

In addition to the base-step and base-pair parameters, the dihedral angles defining the DNA backbone play an important role in protein-DNA interactions. DNA backbone conformations have been known to influence DNA-protein binding sites [41] and thus are an important feature of DNA structure and function. For example, it was observed from an analysis of high-resolution crystallographic data for B-DNA that non-canonical values of α , β and γ are observed mostly in protein-bound DNA complexes [41].

Two torsional angles, namely ϵ and ζ are primarily important as their difference defines the BI and BII conformations that are sampled over the simulations. BI conformation corresponds to the state when $\epsilon/\zeta : t/g^-$ and $\epsilon - \zeta$ is around -90° and BII is recognized with $\epsilon/\zeta : g^-/t$ and when $\epsilon - \zeta > 0$. The equilibrium between BI/BII states is seen to be sequence dependent and plays an important role in site-specific binding of proteins to DNA [144, 178]. Moreover, it has been observed that both BI and BII conformations are common in naked DNA as well as protein-DNA complexes and are highly correlated to the helical parameters, roll, twist and X-displacement [41, 166].

1.1.2 Higher levels of compaction

In most of the eukaryotic cells, DNA is compacted into several levels of hierarchically ordered structures. At the initial level of compaction, it exists in the form of a fundamental repeating unit called the nucleosome, which contain 147 base-pairs of DNA wrapped around the histone protein complex. This complex consists of an octamer of four proteins, called H2A (yellow), H2B (red), H3 (blue) and H4 (green) [145, 101] and together with DNA wrapped around, it is collectively known as the nucleosome core particle (NCP).

All the four core histone proteins are intrinsically disordered proteins and exist as dimers, possessing the highly conserved histone fold domain, which contains three alpha helices connected by two loops (See Figure 1.3). This enables heterodimeric interactions between core histones and is thus known as the handshake motif [4,

[116](#)]. The histone fold is known to directly influence nucleosome level organisation that occurs during several cellular processes. The different dimer surfaces are known to be different in their hydrophilicity and thus provide interaction sites with other proteins [\[5\]](#). The histone fold contains three well separated regions that serve as binding sites with the three consecutive phosphate backbones of the tightly wound DNA and thus form a thermodynamically stable nucleosome, with a specific position of the dyad with respect to the 8 histone folds [\[101\]](#). In addition the H2A-H2B dimers and the H3-H4 tetramer also show pseudodyad symmetry.

The histone core proteins also contain the highly flexible N-terminal tails which protrude out from the nucleosome core particle. The H2A, H2B, H3 and H4 tails are 23, 34, 37 and 24 residues long respectively and their sequences are highly conserved from yeast to human. They are highly basic in nature and thus interact with the nucleosomal DNA, linker DNA and the acidic patches of the nearby nucleosomes, thereby contributing in inter-nucleosome interactions. In addition they are also known to interact with several regulatory factors [\[123, 65\]](#).

The NCP carries a net total charge of -146e (-292e from the wound DNA, +58e from the histone core part and the +88e from the N-terminal histone tails). A large number of these NCPs are connected together with the help of small segments of DNA called linker DNA and this string of nucleosomes then reorganizes itself into a fiber structure called the chromatin fiber which is stabilized with the help of another linker histone protein called H1 [\[45\]](#). However many studies have also questioned the existence of the 30nm fiber in vivo and proposed other models of compaction at higher levels [\[111, 110, 49\]](#).

The exact structure of the 30 nm chromatin fiber is still not very clear and there are two models that exist to explain this, the solenoid model and the zig-zag model. In the solenoid model, the same linker DNA runs spirally along a helical path between two adjacent nucleosomes [\[46, 118, 183\]](#). Most of these studies are based on solution scattering, electron microscopy and x-ray diffraction. Since this bending of the linker DNA requires extra elastic energy there have also been other alternatives in which there is no bending of the linker DNA [\[179\]](#). One such alternative, based on cryoelectron microscopy and scanning force microscopy-measurements, is the zig-zag model, in which two opposite nucleosomes are connected by a straight linker DNA running between them [\[189, 18, 188\]](#). The zig-zag model has seen some support by other experimental studies [\[149\]](#) but it fails to reproduce properties like

FIGURE 1.3: Illustrated figure showing the dimers involved in the histone core proteins and how they form the tetrameric structure. Also the handshake is shown for both the dimers. The final NCP structure is shown on the right. (Image courtesy: [Zephyris](#) (Richard Wheeler) and the Wikipedia)

the linear mass density. Also the problem with these models as observed by Wedemann and Langowski [179] is their inherently static nature and their negligence for long range electrostatic interactions and inherent thermal fluctuations of the system. There have also been several experimental studies that have questioned the 30 nm fiber structure of the chromatin [129, 56].

Figure 1.4 illustrates various models proposed for the folding of the chromatin fiber (Image courtesy: [Julien Mozziconacci](#)) using a computational approach [187]. Under natural DNA twist, it was found that the phasing angle between the DNA at consecutive nucleosomes imposes a strong constraint on the arrangements of the NCPs. For instance for nucleosomal repeat length 177 bp only one structure satisfied all the constraints with all nucleosomes arranged in a 5-start right handed helix. Similarly for 187 and 197 bp models only 3 and 2-start structures were compatible and the 217 bp was compatible with both 4 and 5-start right handed helices.

The 30 nm chromatin fiber is again folded into several loops which go on to form the chromosome. Although it is well established that several NCPs come together to form a chromatin fiber, the exact arrangement and even number of the nucleosomes

FIGURE 1.4: Different models proposed for the folding of the chromatin fiber with different nucleosomal repeat lengths as signified by the number in white

and the mechanism by which this folding takes place is still not very clear [186]. Given the significance of this folding to various regulatory processes inside the nucleus it is necessary to understand this mechanism and the underlying forces and structures responsible for it.

1.2 Factors governing DNA compaction

1.2.1 Role of counterions

Mobile ions affect the flexibility of DNA and also the binding of DNA with the histone core. It has been observed in several studies that a high concentration of monovalent ions can dissociate DNA from the histone core in some ν -bodies from eukaryotic chromatin [181]. By screening the charge of the DNA polymer, they also affect the flexibility of DNA. In essence, monovalent ions on the one hand make DNA more flexible at low to moderate concentration and induce compaction and on the other hand, they compete with positively charged histones for DNA binding and at very high salt concentration (500 mM) they can induce histone dissociation and induce DNA decompaction.

Furthermore charge neutralization of the polyanionic DNA molecule by multivalent cations is seen to be the major mediator of condensation [17]. The cations neutralize the negative charge of the phosphate groups of DNA and thus screen the short-range electrostatic repulsion. The dynamic ion-ion correlations may induce an attractive

electrostatic force thus bringing the negatively charged DNA polyelectrolyte closer [57, 128, 58].

Overall repulsion of negative DNA turns to attraction when about 90% of the DNA charge is neutralized or screened by the counterions (multivalent or monovalent) as it was established experimentally [17] and theoretically [84]. Also a driving force for DNA condensation is the release of monovalent cations from DNA upon DNA interaction with multivalent cations. This purely entropic contribution is strongly dependent on monovalent salt concentration.

At the NCP level, there are many interactions governing its stability. The most important are DNA-DNA and DNA-ion interactions. There have been many studies that have focused on the solution state behavior of the nucleosome. A study by Yang and Hayes [191] reveals that there are specific binding sites for divalent cations like Ca^{2+} and Mg^{2+} on the NCP and even without histone tails, these interactions can further stabilize histone-DNA interactions within tailless nucleosomes and restrict its positional variance. NCP aggregation has been studied by the Livolant group focusing on the role of multivalent cations [143, 37]. Several other experimental studies have focused on the role of tails in governing nucleosome aggregation and interactions [14, 13, 95]

The general rules of DNA condensation discovered in experimental and theoretical studies of model DNA systems must also be valid in natural environment such as eukaryotic chromatin, with histones and other nuclear proteins performing not only as structure forming species but also as multivalent counterions, neutralizing about 50% of the DNA charge in the NCP. Similarly, 50% charge neutralization of linker DNA is observed in natural chromatin with respective supply of linker histones.

At the higher level, it has also been shown that chromatin folding depends not only on the salt concentrations but also on the valency of the counterions. Low salt is seen to encourage a beads-on-a-string arrangement while more compact forms are seen at higher salt concentrations [173]. Divalent cations like Mg^{2+} cause dramatic compaction of chromatin at physiological conditions which cannot be mimicked by monovalent salts by elevating salt concentrations. This has led to some speculation about Mg^{2+} decreasing the persistence length of DNA and thus facilitating bending of linker DNAs [148]. Additional neutralization of the DNA charge required for chromatin compaction comes from polyamine systems

(spermidine, spermine). In one of the earliest study by Wilson and Bloomfield [184], it was found that T7 DNA undergoes a change in conformation from an extended coil structure to a more compact sphere-like form with a critical concentration of 20-24 μM spermidine. Widom and Baldwin [182] found that $\text{Co}(\text{NH}_3)_6^{3+}$ is a five-fold more effective condensing agent as compared to spermidine with DNA collapsing at about 85% charge neutralization. This confirms the idea that DNA condensation by multivalent cations is achieved by the total charge neutralization than the binding activities of these cations.

A study by Clark et. al. [28] again emphasizes the fact that chromatin folding is mainly electrostatic in nature and can be explained using the polyelectrolyte theory of Manning [113]. Their conclusion is mainly based on the shielding of the electrostatic free energy of the linker DNA by the H1 linker histone and the prevailing ionic conditions, thus enabling them to come closer by overcoming electrostatic repulsion of the linker DNA. The electrostatic and polyelectrolyte properties of chromatin have been reviewed in [82].

In summary, DNA compaction is mediated mainly by electrostatic interactions. Being a molecule containing many charged species like the polyanionic DNA, positively charged histone octamer core and the positively charged N-terminal tails it sounds intuitive that DNA compaction must be a very electrostatic-driven process. This is indeed what happens with the histone proteins and their tails playing very important roles.

1.2.2 Histone proteins and their tails

The histones are the primary proteins involved in the compaction of DNA in most of the eukaryotic cells and help in compacting DNA into an orderly structure inside the nucleus. In addition to that, they also undergo interactions among themselves to influence various structural and functional properties of the NCP [102]. The core histone proteins are more responsible for stabilizing the nucleosome's foundational structure and play a major role in compacting DNA not only at the nucleosome level, but even at the chromatin level [116]. The inner DNA part (120 bp) interacts with the histones mainly through the histone fold domains while the remaining part (13 bp each) interacts with the N-termini.

Together with the H1 histone the core histone proteins are known to influence and facilitate chromatin folding [189] which is again mostly mediated by electrostatic interactions [18]. It is known that although chromatin folding still occurs without the H1 linker histones, proper folding is only possible in their presence and hence they play a role in chromatin compaction [172, 165].

On the other hand, the N-terminal tails of the histone proteins are known to play a major role in influencing the higher order chromatin structure [102] and gene regulation in particular [76, 119, 169, 167]. Various post-translational modifications enable them to control the transcription activity and they also serve as interaction sites for other proteins that influence higher-order chromatin structure [67, 43, 52]. Additionally with post-translational modifications they are known to influence DNA-histone interactions and in general even the chromosomal structure [62, 11]. Their removal has been shown to hamper DNA-histone binding in the nucleosome [139, 7] and the formation of fully condensed chromatin fibers [168, 47, 54]. Another study by [74] found that deletion of H2B or H3 tails affects histone-DNA interactions and also the stability of the nucleosome.

NCP-NCP

Also some studies of mono-nucleosomes [4] have elucidated the significance of linker histone tails on the folding of DNA around the histone core and thus the formation of the chromatin fiber. It has been observed that even without the tail domain, the linker histone encourages the folding of DNA around the histone core, however the wound DNA-ends start repelling each other. However with the tail, DNA winds properly around the core and forms a stem-like structure.

Given their important functions in DNA compaction, the NCP-core histones form the most conserved family of proteins among various eukaryotic [146] species. However, H1 is among the lesser conserved histone proteins as compared to them [63].

1.2.3 NCP-NCP stacking

In nucleosome interactions, NCP-NCP stacking with the two faces containing the histone octamer cores has been seen to play a major role. Many studies in vitro have focused on this effect showing at the chromatin level NCP-NCP stacking is observed along the dyad axis [154]. Various other studies have established close

NCP-NCP stacking in NCP crystals [36], in the tetranucleosome crystal [153] and in the liquid crystalline phases [91].

Single molecule techniques have been employed by Cui and Bustamante [31] using magnetic tweezers to measure the attractive forces among the nucleosomes in a single chromatin fiber. The John van Noort group has elucidated chromatin folding and force mechanics using magnetic tweezers [88] and single molecule FRET [81].

The dynamic nature of DNA packing at higher order with constantly moving tail domains in the presence of mobile multivalent cations makes it difficult to study this system using structural techniques like single crystal X-ray crystallography and one needs to resort to computational techniques to study the nucleosome interactions and how it affects the higher order packing of DNA at the chromatin level.

1.2.4 Persistence length of DNA

To form the nucleosome, DNA has to turn 1.7 times around the histone core. This naturally compels us to ask the question concerning the chain flexibility of DNA and the role it plays in the packing of DNA into the small volume of the nucleus. Not only to form the nucleosome, but even in other forms of compaction, flexibility plays an important role in packing DNA in highly compact biological organisms like the viruses, microorganisms and other eukaryotic structures. Also, since DNA is a highly charged polymer, the flexibility is in turn affected by the presence of charged ions. Since DNA is frequently accessed by the molecular machinery, this compaction also has to be very dynamic so as to be easily and orderly retrievable for various physiological purposes. Hence to study DNA compaction and its functional importance, a proper understanding of the flexibility of the DNA is warranted.

An important parameter to quantify DNA stiffness is the persistence length of DNA. It is a measure of the decay of the correlation in the direction cosines along the tangent of DNA and is defined as:

$$L_P = -\frac{L_C}{\ln(\langle \cos \alpha \rangle)} \quad (1.1)$$

here L_C is the contour length and $\cos\alpha$ is the average of the cosine between two adjacent segments over the length of the simulation for the corresponding contour length.

Various experimental techniques have been used to quantify and measure the persistence length. These include flow birefringence [59, 22], transient electric birefringence [59, 100], magnetic birefringence [93, 115], optical techniques like light scattering [78, 158] and linear dichroism [147, 140], laser tweezers [9] among others. To derive the persistence length from the aforementioned experimental data several analytical techniques have also been developed [114, 141, 61]. For a more comprehensive review on the experimental efforts devoted to derive the persistence length of DNA, see these [60, 150].

Computationally, there have been various efforts devoted to understand the flexibility of DNA. For persistence length, given the relatively long length of DNA required to simulate, most of the studies have been devoted to study the persistence length at a coarse-grained level of DNA. Some of these models are discussed in more details in the upcoming sections.

1.3 Computational modelling of DNA compaction

Growth of computational resources has heralded an era of unprecedented growth in the simulation timescale of bio-molecules with hundreds of thousands of atoms. Microsecond-range simulations are now the norm when a several nanosecond simulation was considered a big leap not many years ago. Over the years, molecular simulations have been used extensively to understand the organization of genomic information over different scales ranging from a single DNA molecule represented by detailed all-atom models to the complex chromosome-level structures represented by coarse-grained (CG) models of different complexity and resolution. The importance of computer simulations follows from the fact that most of the experimental studies including X-ray diffraction and NMR provide ensemble averaged properties, which are not always representative for instant single-molecule structures. Additionally, experimental measurements rarely provide information on the dynamics of the DNA at atomistic level and over small timescales. Computer simulations can fill this gap by providing time-resolved evolution of 3D atomistic structures where

the motion of each atom and each molecule can be followed. In summary, computational techniques can play an important role in giving a complete theoretical description of the structure and motion of DNA.

1.3.1 All atom simulations

At the all-atom level there have been several efforts devoted to study the structure and dynamics of the DNA molecule since the first DNA simulation by Levitt [92] in the 1980s. DNA has been extensively studied since then and our knowledge of the geometry of DNA at the base-pair level has improved by a great deal. Owing to the growth in available computational resources, we have reached the stage where microsecond simulations are routine. Nonetheless, simulations at the ns timescale have an important contribution as most of the force fields were parameterized based on them. For a detailed review see [27, 136, 25, 24, 112]. However given the large number of atoms for even a relatively short DNA (10-20 base-pairs), most of the studies have been focused on very specific sequences much shorter than the persistence length of DNA (≈ 150 bp). These include mostly the Dickerson dodecamer sequence [40] and the JunFos oligomer [64, 195] among others. The length of the DNA sequence is also limited to tens of base-pairs to decrease the total atom count.

There have been a few studies that involve studying DNA-histone interactions at the all-atom level. For example a computational study by Biswas et. al. [16] sheds light on the role on histone tails on the nucleosome stability. They used different all-atom simulations totaling 800 ns of the nucleosome core particle with and without the tails. In another all-atom study [185] involving the 601L nucleosome [174], reversible and spontaneous detaching of outer stretches of DNA was observed and was found to be highly correlated with the CG content of the sequence. Another study [29] found that the H3 and H2A tail domains tend to form α -helices when present in the NCP. Also H3 and H2A tails have the most pronounced effect on the structural stability of the histone core domains while the other two tails had no effect. Shaytan et. al. [157] used microsecond long all-atom simulations of mononucleosome with DNA linkers and observed twist-defects and DNA bulging at the entry-exit sites of the nucleosomes, indicative of nucleosome sliding. Such studies are rare, especially at the nucleosome level, as it involves humongous computational

resources to study large biomolecules at physiologically relevant timescales. This motivates the development of coarse-grained models with the ‘redundant’ degrees of freedom removed. These are discussed in the future sections.

1.3.2 Selection of force fields

One important consideration when running DNA all-atom simulations is the choice of the force field used to model the interactions and the energy function. Hence the quality of information obtained from molecular simulations depends critically on the force field used to describe interactions in the simulated system. For nucleic acids, the initial versions of the atomistic force fields AMBER [180] and CHARMM [127] were developed in the 1980s on the basis of computations for simple molecules, and were tested on small DNA oligomers [126, 155]. Further development of the AMBER family of force fields resulted in the parm94 parameter set [30]. More recent modifications, [176, 26, 137] particularly parmbsc0[137] became widely used in simulations of nucleic acids. Recently, when microsecond-long simulations became feasible, a new refinement of AMBER FF, the parmbsc1 parameter set [73], was released. The CHARMM family of force fields [109] was also developing, with the milestone versions CHARMM27 [48, 108] and CHARMM36 [64]. Further information of force field development for nucleic acids can be found in the reviews [144, 50, 34].

Most force fields development is focused on parameter optimization of some the force constants and the bond length [48]. C27 involved parameterization of all the force constants and the equilibrium bond lengths. For C36, to reproduce correct BI/BII proportion characterising the torsion angles, the dihedral parameters associated with ϵ and η torsion angles were optimized. For the AMBER family of force fields, parmbsc0 involved the parameterization of the α/γ torsion angles. Similarly parmbsc1, the latest force field was parameterized based on QM data [73]. For more details please refer to the referenced works.

Given the increasing variety of force fields, it becomes imperative to test their capability to predict various properties of DNA. A number of works in the past have been devoted to validation and comparison of existing versions of force fields [73, 64, 24, 135, 90, 33, 192]. In most of these studies validation has been focused

on RMSD variations for relatively short DNA oligonucleotides such as the Drew-Dickerson Dodecamer (DDD) [42] with simulations at timescales less than 200 ns [24, 15, 135, 90, 41]. Experimental structures used as a reference in these studies were obtained typically from X-ray crystallography of samples crystallized using non-physiological solvents, which raises the question of the effect of crystallization conditions for physiologically relevant DNA structures. In a recent paper by Orozco and coworkers, [33] several recent versions of AMBER, as well as the CHARMM36 force field have been analyzed in multi-microsecond simulations of the DDD sequence and two other 12-mer oligonucleotides with NMR-resolved structures, from which conclusions about the suitability of the parmbsc0 and parmbsc1 versions of the AMBER force field have been made.

Noteworthy, all aforementioned papers provide comparative analysis of the force fields for short DNA sequences, which leaves open the question on the predictive power of different force fields for simulations of long DNA and higher order aggregated DNA structures. This question is becoming increasingly important since atomistic DNA simulations are nowadays used in the bottom-up multiscale modelling schemes as a reference to derive effective potentials for coarse-grained models suitable for mesoscale simulations [125, 86, 152]. Good quality of the atomistic force fields, and particularly their ability to reproduce large-scale properties of DNA are of primary importance for CG bottom-up models whose predictive power relies on accuracy of the underlying all-atom force fields.

Since water molecules form the majority of the computational load of the all-atom MD simulations in explicit aqueous solutions, the right choice of the water model is important that is computationally efficient. For the water models the most efficient being the 3-point TIP3P is preferred mostly for all-atom MD simulations. It is computationally efficient and also contains a polarization term.

In addition the ion parameters used are the ones

1.3.3 Coarse-grained modelling

Given that we want to study large biomolecules like a double-stranded DNA with 100 base-pairs containing hundreds of thousands of atoms, or like the chromatin fiber of eukaryotic cells, it still seems too impractical and computationally expensive

to simulate at the all-atom level, given the large number of redundant degrees of freedom of the system. Furthermore, for the description of many mesoscale properties of DNA, such as persistence length and other elastic properties as well as aggregation of DNA to various compact and ordered structures, formation of nucleosome core particles and their further organization into chromatin structures, all-atom simulations are unfeasible. To address this problem, building a coarse-grained model with the unnecessary details removed, is one of the approaches used to study large biomolecular systems.

In most CG models, many atoms of macromolecules are represented by single CG site and also the solvent is not considered explicitly, thus reducing the degrees of freedom considerably. A CG model which achieves an optimum balance between the computational efficiency and also the elucidation of important properties of the system, is generally used. A coarse-grained method has the added advantage of reducing the time required to study the system and one can thus achieve larger time-scales mimicking timescales of *in vivo* processes.

One central problem in developing of any coarse-grained model is the choice of the effective potentials to be used among the representative CG sites. In the top-down approach, one generally starts with a simplistic model defined by experimental data and then fitting the data to reproduce some physical properties [156]. These can be radius of gyration, osmotic pressure, etc. On the other hand for a bottom-up approach we start from detailed all-atomistic simulations. The effective potentials for the coarser model are built and parametrized from these more detailed atomistic force fields and mostly they also include the effect of solvent. So in effect one can afford modelling of large biomolecular systems and at the same time use effective potentials implicitly including the atomistic details. One example, the 3SPN.2 model, which is essentially a top-down model, the potentials have been parameterized using experimental data (see [71]) and other adjustable parameters. A drawback of the model is that there are no explicit ions used to model the counter ion condensation, but done using effective charges on the phosphates. Since electrostatics is treated using the Debye-Hückel approximation, it limits its applicability to high ionic strength.

Several approaches exist to derive these effective potentials in the bottom up approach. One such approach using experimental data to parameterise effective potentials is the MARTINI CG force field [117] which uses thermodynamic properties

as well as structural data for the underlying system. Another notable example is the molecular renormalization group coarse graining (MRG-CG) technique by Papoian group [151]. Other methods include the inverse Monte Carlo method [105, 120] which is discussed in this work, the Iterative Boltzmann Inversion (IBI) [175], force-matching [75, 96, 97] and relative entropy minimization method [8].

Modelling the interactions governing the nucleosome stability and chromatin folding

As we already know that the nucleosome is comprised of various components including DNA, histone core and the histone tails, to build a coarse-grained model, one has to first understand how all these coarse-grained components interact with each other. Plus, to account for the salt concentration effects, implicit methods like the Debye-Huckel approximation or explicit mobile ions have to be used. Also to make a fully functional biologically relevant model, the effect of solvation, again either explicitly [39] or implicitly using e.g. Langevin dynamics, as in most of the CG models, has to be considered.

In short one needs to know the effective potentials that need to be used to describe the interactions governing the system. However, it needs to be mentioned a major approximation in the current multiscale approach is that the various effective potentials will be obtained from different all-atom simulations but will be used in a unified way in the final large-scale CG simulation and it is assumed that these potentials are transferable. This approximation is well justified [162]. To model the NCP and chromatin in a CG approach, all the interactions below are required and will be obtained from different all-atom simulations:

- DNA-DNA interactions
- DNA-histone octamer (HO) core interactions
- DNA-histone tail interactions
- DNA-ion interactions
- Histone tail-HO core interactions
- Histone tail-ion interactions

- HO core-HO core interactions
- Ion-ion interactions
- Solvent effects

DNA coarse-grained models

A wide variety of CG DNA models [142] with different levels of complexity and resolution have appeared recently (see reviews [38, 130] and references cited therein). As mentioned above, there are two basic strategies in development of CG models, top-down and bottom-up. Within the top-down approach the CG model parameters are fitted to reproduce some available experimental data. An example of the top-down approach is the 3SPN CG DNA model developed by the de Pablo group [80, 71, 70]. This CG model contains 33 beads per nucleotide, representing the phosphate, sugar and the base-pairs and includes sequence specificity. Debye-Huckel approximation was used to account for the ionic environment and a separate term was included to take into account the effect of hydrated layers. This model is able to reproduce some experimental results like the persistence length dependence on salt, sequence dependent DNA melting and the thermodynamics of double helix - single strand transition. Other top-down CG models that have appeared in the literature include the SIRAH force field [107, 106] which gives a CG force field based framework for studying DNA interactions with proteins.

The MARTINI CG FF [117] combines the top-down and bottom-up approaches for parametrizations. An updated MARTINI CG model [171] uses three beads per nucleotides Cytosine and Thymine and four for Guanine and Adenine. For the backbone, it uses one bead for the phosphate and two for the sugars. Special bead types to describe the hydrogen bonding were incorporated into the model and they used an explicit solvation technique with one bead representing four water molecules. Their model was able to reproduce the radius of gyration of ssDNA and the ion distribution curves for dsDNA for Na^+ . However its treatment of the elastic network and its inability to use directional hydrogen bonds make it not very suitable for structural studies but a good choice to study large-scale DNA packing and other applications.

FIGURE 1.5: Different DNA CG models developed. (A) 3SPN.2 model, (B) MARTINI CG FF, (C) Savelyev et. al. CG model and (D) Naome et. al. model.

Within a systematic bottom-up multiscale modelling scheme, the interaction potentials between CG sites are derived from the information obtained in underlying atomistic simulations [105, 125, 86, 152]. While different CG strategies have their respective strengths and weaknesses [163], it is evident that for bottom-up approaches the quality of the underlying atomistic force field is of primary importance. Savelyev and Papoian [152] have developed a CG model with two beads per base-pair using specific potential terms for the ion-ion and DNA-ion interactions derived from the underlying all-atom MD simulations [151]. In this model the DNA is represented as two chains of phosphate groups where each site was connected by 10 bonds with the sites of the opposite chain. In total there were 39 DNA interactions and 8 ion-DNA interactions. The ions were represented explicitly. The model was able to reproduce experimental results for DNA persistence length and also predict structural phase transitions in torsionally stressed DNA nanocircles which gives an insight into the contribution of electrostatic and mechanical effects on nucleosome stability.

In the bottom-up approach, Naome et. al. [125] used a 5 bead per nucleotide model and used Iterative Boltzmann Inversion combined with Newton Inversion to fit all-atom MD parameters and also emphasized the importance of long ranged electrostatics in rendering a proper flexibility to DNA polyelectrolyte. This model was able to correctly reproduce ring closure probabilities and the DNA mini-circle topologies. For a more detailed review of all the coarse-grained models discussed please see [34]. Korolev et. al. [86] employed a 5 bead model to derive the salt dependence of persistence length DNA using the Inverse Monte Carlo approach which uses explicit electrostatics and with charges assigned to each bead. Figure 1.5 depicts the different DNA CG models and their bead mapping schemes. (Images reproduced with permission from [71], [171], [152] and [125])

Coarse-grained models of NCP and chromatin

There have been several computational studies focused at the nucleosome and chromatin level of DNA compaction. Given the complexity and the size of the systems involved most of the computer modelling studies have been done at the coarse-grained level as an all-atom model with just a few nucleosomes would easily contain around 1 million atoms and are computationally expensive.

To study the effect of salt concentration and other multivalent cations as mentioned above, Korolev et. al. [83] used a combination of experimental and computational approaches. They found that a threshold concentration is needed for the folding of the nucleosome array. Using a dielectric medium and employing explicit mobile ions with charged histone tails their results were in better agreement with experiments as compared to the Debye-Huckel based approach for the folding of chromatin based on multivalent cations. The argument is based on the fact that by including explicit mobile ions one can include ion correlation and ion competition effects. In another work Lyubartsev et. al. [104] developed a super coarse-grained model for the NCP and were able to observe NCP precipitation into a large array for a large system containing around 5000 NCPs, in agreement with experimental SAXS data.

Beard and Schlick [10] have simulated a model based on disc-shaped nucleosomes connected by elastic DNA segments. Their model uses charged particles distributed on ideal rigid disks which are connected by linker DNA beads. However, they included no explicit treatment of the H1 histone and only parts of the histone tails

present in the crystal structure were considered. The NCP core was treated as completely rigid and DNA unwinding and breathing were not considered. Brownian Dynamics simulations for both a dinucleosome and a trinucleosome were done over a course of 100 ns with a 2 ps time-step in both high (0.05 M) and low ($C \leq 0.02$ M) salt concentrations. They observed that at low concentrations the linker DNA bends and assumes a disordered configuration and assumes a more condensed form at higher salt concentrations in agreement with experimental results. For the bigger system with 48 nucleosomes, a similar behavior is observed with a unfolding of the fiber at low salt concentration (0.01 M) and stability under high salt conditions (0.05 M) correctly reproducing small-angle X-ray scattering data.

Wedemann and Langowski [179] have also modeled the chromatin based on a coarse-grained approach. They approximate the nucleosome cores as ellipsoids interacting via the Gay-Berne potentials [53] while using harmonic potentials for the coupling between the chromatosome and the linker DNA. The repulsion between linker DNA is modeled by line charges interacting via a Debye-Huckel potential, with all the potentials parameterised using experimental data. Using this model they have been able to reproduce experimental data at physiological salt conditions like mass density, diameter, persistence length of bending and the tilt angle of the nucleosome and DNA to the fiber axis.

Schlick et. al. [161] subsequently improved the model by employing a more comprehensive DiSCO model [197] for charge distribution on irregular surfaces like the smooth molecular surface of the nucleosome. The electrostatics for the nucleosome was obtained by solving the non-linear Poisson-Boltzmann equation. Another difference was an explicit treatment of the histones as compared to the previous model in [10]. It was observed that different underlying forces lead to folding and unfolding of the fiber, with linker DNA repulsion leading to unfolding at low salt concentrations while internucleosomal attraction leading to the folding of the fiber at higher concentrations.

Arya and Schlick [6] further improved on the model in [10] by incorporating linker histones bound to nucleosomes and divalent cations and also improving DNA-nucleosome geometry and mechanics. Using a much more advanced MC sampling technique they comprehensively studied the role of linker DNA and histone tails on chromatin fiber's stability. They also uncovered different roles for different histone tails: H3 and H4 mediate more closer internucleosomal attractions. H3

also encourages packing by screening the linker DNA repulsion, while H2A and H2B mediate internucleosomal interactions across distant regions of the fiber thus encouraging fiber-fiber aggregation.

Using a two angle CG-model similar to the one used in [179], Wedemmann et. al. [159] modeled chromatin folding with different shapes of nucleosomes, spherical described a Lennard-Jones potential, oblate ellipsoids described by a modified Gay-Berne potential [53], and spherocylinders described by the Zewdie potential [160, 194]. DNA was modeled using harmonic potentials for stretching, bending, and torsion. The Debye Huckel approximation described DNA-DNA interactions and the excluded volume effect was taken care of by a hard core potential. Simulating with 100 nucleosomes connected by the two-state model described in [197], this study mainly evaluated the effect of local nucleosome configuration and the nucleosome repeat length on the folding of the chromatin fiber.

Another interesting work by Wedemmann et. al. [79] was a virtual force spectrometric study using coarse-grained Monte-Carlo simulations on chromatin fibers and connected the force spectrometry curves from experimental studies. The CG model used is similar to the one used in [179]. They considered various factors including the interactions, nucleosome geometry and also the unwinding of DNA from the histone core on the force-extension curves. However they did not include the effect of multivalent salts (eg. Mg^{2+}) on the chromatin fiber.

One serious problem with Poisson-Boltzmann based models is that they have limitations with regards to the fact that they do not treat mobile ions explicitly and thus fail to adequately include their effects on the compaction behavior. Also most of the studies have not included the entropic effects and the ion-correlation effects that result from the replacement of monovalent cations by the physiologically important multivalent cations like spermine, spermidine and Co-hexamine [85].

1.4 Objective of the present work

Computer simulations play an important role not only in complementing experimental techniques but also in advancing our knowledge about various biological systems. Here we studied the performance of various all-atom force fields for the

DNA. Testing these force fields is important as they are used heavily by the computational research community in solving various problems involving DNA. Also, since many coarse-grained models are dependent on underlying atomistic simulations their performance needs to be assessed as they heavily influence the CG model used to study large-scale problems. So we also go further and test the ability of all-atom DNA force fields to parameterize coarse-grained models. To this end, we carried out microsecond long molecular dynamics simulations of a 40 base-pairs DNA fragment using two recent versions of AMBER (parmbsc0 and parmbsc1) and CHARMM (C27 and C36) force fields. Besides detailed analyses of the atomistic DNA structure in terms of various DNA helix parameters and comparison with average data for B-form of DNA, on the basis of these simulations we derived parameters for a CG DNA model using the Inverse Monte Carlo (IMC) methodology [120, 121]. We used the CG DNA model with parameters obtained from each of the underlying atomistic force fields to simulate a 500 bp long DNA and compute DNA persistence length as a function of ionic strength. Such an approach represents, to the best of our knowledge, the first attempt to compare the ability of different atomistic force fields to predict mesoscale DNA properties such as persistence length and mechanical rigidity.

The long term goal is to perform bottom-up CG multiscale simulations of NCP and chromatin. To do this, it is highly important to have an accurate model of DNA with potentials that are reliable and describe well the flexibility of DNA over various salt concentrations. The underlying atomistic simulations and the chosen force field need to be validated and tested for the model and how it performs in CG simulations. This is important because DNA flexibility and bending is highly important for describing properties of NCP and chromatin. The following step, already in progress is to obtain bottom-up CG potentials for the amino acids and interactions between components of NCP to model NCP and chromatin compaction.

Chapter 2

Materials and methods

2.1 Introduction: the multiscale approach

Our final goal is to model the chromatin fiber using a systematic bottom-up multiscale approach. This includes using computer simulations to achieve compaction and self-assembly of DNA with histone proteins at varying levels of complexity. As described above all-atom studies are too expensive computationally for large biomolecules such as chromatin. One needs to simulate large DNA molecules with the histone proteins in multivalent salt surrounding the solvent for several microseconds. As this is not computationally feasible with high resolution all-atom models, one needs to resort to lower resolution coarse-grained techniques which require less computational time. All this has to be done while preserving an accurate treatment of the forces and potentials at which various biomolecules interact. A general overview of the multiscale approach to model DNA compaction is shown in Figure [2.1](#).

The multiscale approach consists of various steps to attain the final goal of a large-scale simulation of the nucleosome and subsequently the chromatin fiber.

- **All atom MD:** Using the high-resolution structures of the bio-molecules under interest, mainly DNA and histone proteins, all-atom simulations are performed using a physiological salt concentration. These simulations are run using state of the art MD simulation engines and all-atom force fields. To obtain a good sampling of the configuration space, these are run for at least

FIGURE 2.1: Multiscale modelling of DNA compaction

a microsecond which also ensures convergence of the structural properties of the system. A point to be noted is that the choice of the force field is important as all the properties obtained are affected by the underlying force field.

- **Bead-Mapping:** Once we obtain an all-atom trajectory of our system it is bead-mapped to a coarse-grained representation. The CG representation is chosen intuitively in such a way that it preserves most of the properties of the high resolution all-atom model and also reduces the redundant degrees of freedom for the system under study.
- **RDF Calculation:** The coarse-grained trajectory obtained is then used to calculate principal structural properties (RDF) of the system. These properties are purely based on the all-atom trajectory and the underlying force field. These structural distributions will define the topology of the CG model and bead-types defining the unique interactions in the coarse-grained model.
- **Deriving CG effective potentials:** Using these RDFs as a footprint of the all-atom system, we go on to extract effective potentials using an iterative procedure called the IMC method [105]. For each underlying structural distribution, a unique potential is obtained. For example for the single-DNA

FIGURE 2.2: Flowchart explaining multiscale modelling using IMC

system, there are 14 such potentials describing the interactions between the coarse-grained beads.

- **Reproduction of RDFs:** Once sufficiently converged potentials are obtained these can be tested by reproducing the underlying all-atom reference RDFs at the coarse-grained level. A good agreement between the all-atom model RDFs and the RDFs from the coarse-grained simulations ensure a good convergence of the IMC procedure.
- **Large-scale CG simulation:** Using the potentials obtained from the IMC, one can finally run large-scale coarse-grained simulations to understand the organization of DNA and histones and the role of counterions in chromatin compaction. A good test for the coarse-grained model can be obtained by reproducing some experimentally measured quantity in the simulations. Some examples include persistence length of DNA, toroid formation inside DNA capsids, etc.

The model is purely based on results of the all-atom simulation, without having any ad-hoc parameters. A flowchart showing all the steps involved is described in Figure 2.1.

2.1.1 Challenges

This summarizes the basic procedure one follows in deriving the effective potentials for a coarse-grained model in a multiscale approach. There are many challenges one can encounter along the way that need to be overcome for an accurate and realistic description of the underlying biological system, corresponding to physiological conditions.

Firstly, the chosen force field for the all-atom model. There can be various artifacts and anomalies in the force field and one needs to carefully scrutinize the system behavior at the all-atom level so that any such artifacts can be taken care of.

One problem we faced was base-pair fraying and base-flipping for one of the force fields. As described in more detail in the subsequent sections, the DNA molecule starts melting from one end. This may have two interpretations: the force field is so exact that it may be similar to physiological conditions and the simulation is actually imitating nature; the second possibility is of some artifacts in the force field that fail to properly account for the hydrogen bonding in the DNA double helix. We faced this problem for one of the force fields used and this motivated us to study in more detail the underlying force field and thus contribute in parameterizing the all-atom force field by comparing with experiments.

Secondly at the coarse-grained model, mostly the interactions are approximations of the all-atom MD situations as the system is complex and thus solving a N-body potential mean force (PMF) problem is not possible for such large systems. So rather than finding the exact solution that includes integrating the respective partition function over all the missing degrees of freedom, one tries to reduce the N-body PMF problem by approximating it using certain pair potentials within a certain distance range. The best solution is obtained, whether by force matching technique, entropy minimization or other structure based approaches like IBI or IMC and is the one that best approximates the N-body PMF.

Within the IMC procedure one tries to obtain the potential in an iterative procedure by starting from a trial potential. The corrections to the potential are obtained by inverting the Jacobian of the differences relating the reference RDFs and the iterated potential. There may be certain convergence issues but as IMC also considers the correlations between different types of RDFs the convergence is better as compared to the Iterative Boltzmann Inversion approach [87].

2.2 All-atom simulations

2.2.1 Single DNA system

To understand the dynamics and structure of the DNA at a single molecule level, simulations of a 40 bp long double helical DNA molecule with 42.5% GC content (typically found in eukaryotic cells [94]) were carried out in physiological salt (Na^+/K^+) conditions and water. The full specifications of the system are given below:

- Sequence (42.5-% GC):
5'-GGATTAATGGAACGTAGCATATTCTTCAAGTTGTCACGCC-3'
- Salt concentration: 130 mM
- 116586 TIP3P [77] water molecules
- 15 nm cubic box
- Forcefields tested: parmbsc0 [137], parmbsc1 [73], CHARMM27 [48, 108], CHARMM36 [64]

For CHARMM27, the potential energy function can be written as:

$$\begin{aligned}
 U(\mathbf{R}) = & \sum_{bonds} K_b(b - b_0)^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 \\
 & + \sum_{dihedrals} K_\chi(1 + \cos(n\chi - \delta)) \\
 & + \sum_{UB} K_{UB}(S - S_0)^2 + \sum_{impropers} K_{imp}(\phi - \phi_0)^2 \\
 & + \sum_{nonbonded} \epsilon_{ij} \left[\left(\frac{Rmin_{ij}}{r_{ij}} \right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}} \right)^6 \right] \\
 & + \frac{q_i q_j}{er_{ij}}
 \end{aligned} \tag{2.1}$$

where b is the bond length, θ is the valence angle, χ is the torsion/dihedral angle, ϕ is the improper angle, S is the distance between atoms separated by two covalent

bonds and r_{ij} is the distance between atoms i and j . The corresponding force constants are denoted by a K , subscripted by the name of the interaction.

All simulations were performed using GROMACS 5.1.2 [69, 1]. DNA initial structure was built using the [SCFBio server](#). All the simulations were started from a linear B-DNA molecule in a cubic box with periodic boundary conditions. Ions and water molecules were added to the simulation box using the GROMACS utility *genbox*. For the simulations with parmbsc0 and parmbsc1, the improved ion parameters CUFIX [193] were used. For CHARMM27 and CHARMM36, the default ion parameters included in the force field package were used.

Equilibration was done in three steps. In the first step DNA and ions were restrained while the solvent molecules were allowed to uniformly attain the temperature of 300 K. In the second step only the ions were released and the pressure coupling was switched on to attain the pressure of 1.023 bar. The temperature was held constant by velocity-rescaling [20] and the pressure coupling by the Parrinello-Rahman thermostat [133]. This was followed by unrestrained motion of all the molecules for 10 ns. Bonds to hydrogen atoms were constrained using the LINCS algorithm.

For production simulations, the leap frog integrator was used. The time-step used was 2 fs. The Verlet cut-off criteria was used for neighbor searching with short range cutoff defined by the default buffer tolerance. The particle mesh Ewald (PME) [35] method was used for electrostatics with the real-space cut-off set to be 1 nm and also van der Waals interactions truncated at 1 nm. A summary of all the simulations done for the 1DNA system is given in Table 2.1. Simulations were performed on the NSCC cluster and some on the NTU's HPC cluster. Visualization of the trajectories was done using VMD [72].

As noted above, the choice of the DNA force field used is very crucial. Hence to study the effect of different force fields on the flexibility and dynamics of the DNA at the all-atom level, we performed all-atom simulations for four different force fields as noted in Table 2.1.

Harmonic constraints To prevent base-pair fraying at the termini, the terminal base pairs were constrained by additional harmonic bonds (two for the AT pair and three for the GC pair). The harmonic bonds were chosen from the distance

TABLE 2.1: Summary of all-atom simulations

Force field	System	Trajectory Length (μ s)
<i>Parmbsc0</i>	1DNA + KCl	2.0
<i>Parmbsc1</i>	1DNA + KCl	2.0
<i>CHARMM27</i>	1DNA + KCl	2.0
	1DNA + NaCl	2.0
<i>CHARMM36</i>	1DNA + KCl	2.5 (Traj1)
		1.2 (Traj2)
	1DNA + NaCl	1.6 (Traj3)
		1.2 (Traj4)

FIGURE 2.3: Harmonic bonds added on terminal base pairs

restraints function in GROMACS and the equilibrium distance was chosen to represent such as in the B-DNA structure. Figure 2.3 describes the imposed restraints for the corresponding atoms. This ensured a stable DNA structure and minimized the contamination of the RDF data as observed in [86]. Also the parameters and type of constraints used are described in Table 2.2. The first three atom pairs belong to the base-pair 1-40 and the last three atom pairs belong to the base-pair 40-41.

TABLE 2.2: Parameters used for the harmonic constraints imposed on the terminal atoms

Atom1	Atom2	R (Å)	Force constant (kJ mol ⁻¹ nm ⁻²)
O6	N4	0.27	25000
N1	N3	0.29	25000
N2	O2	0.30	25000
O6	N4	0.27	25000
N1	N3	0.29	25000
N2	O2	0.3	25000

2.2.2 Multiple DNA and histone tails systems

The system studied in the first and main part of this work, is a simple single DNA system and gave mainly information regarding the interactions of DNA with the monovalent ions and the flexibility of the DNA molecule and its structure. However to obtain effective potentials for the full system like the nucleosome, as mentioned in the previous section, we need to study DNA-DNA interactions, DNA-histone tail and histone tail-ion interactions. In line with this approach two more all-atom simulations were done to derive the corresponding CG effective potentials. The force field used was C27, given its excellent performance in the comparison of all the four force fields tested at the single DNA level and the coarse-grained level (see results section). The simulation protocol was similar to the single-DNA case, unless otherwise mentioned.

1. *Histone tails with ions*

This simulation was used to derive effective potentials for histone tail-ion interactions. Rather than using the tail as it is for deriving the potentials, each tail was divided into fragments of eight amino acids each. 48 fragments containing 8 amino acid each were built using the sequences of the histone tails. This procedure eased the computational cost as individual tails are quite long and this would make it difficult to sample the configuration space efficiently. This is in part because long tails would get stuck more often in metastable states. The rationale for introducing additional fragments was mostly based on:

- Proline being at the center of the fragments
- Increasing presence of some under-represented amino-acids
- Better representation of Val, Leu, His, Gln
- Not to deviate from tail sequences

The fragments were chosen using the sequences of the histone tails in such a manner that in the end we get a well-represented amino-acid concentration corresponding to the histone tails that compensates about half of the negative charge of DNA molecules. Additionally several permutations and combinations were done in the sequences to obtain a good sample. Figure 2.4 illustrates this methodology and Table 2.3 shows the relative concentration of the amino acids used in the system. Also the full list of fragments with their respective charges is given in the Appendix (A.1)

- 48 peptide fragments derived from histone tails
- Charge balanced with 120 Cl^- ions
- 338676 TIP3P water molecules
- 15 nm cubic box
- Salt concentration:
 - (a) System 1: 130 mM NaCl/KCl
 - (b) System 2: 22 mM Mg^{2+} , 45 mM NaCl/KCl
 - (c) System 3: 15 mM $\text{Co}(\text{NH}_3)_6^{3+}$, 45 mM KCl/NaCl
- Simulation length: $2\mu\text{s}$

The fragments were built with Ambertools [32]. As the resulting fragments were linear they were subjected to 400 ps MD to attain a more curved shape. Also the ends of the fragments were capped with an acetyl group on the N-terminal and a amide group on the C-terminal to make the ends neutral. Later these groups were removed from the analysis.

An all-atom simulation with other multivalent cations like Mg^{2+} , Cobalt(III) Hexamine ($\text{Co}(\text{NH}_3)_6^{3+}$) and four DNA molecules has already been done by other members of the group. Also the effective potentials have been obtained for some of these systems, thus validating the IMC technique for more complex systems with both intermolecular and intramolecular DNA interactions.

FIGURE 2.4: Illustration of how the 48 fragments were derived from histone tails

TABLE 2.3: Amino acid concentration

Category	Amino-acid	Number
<i>Negative</i>	Aspartic Acid (D)	12 (3.1%)
	Glutamic Acid (E)	9 (2.3%)
<i>Positive</i>	Lysine (K)	99 (25.8%)
	Arginine (R)	42 (10.7%)
<i>Polar</i>	Serine (S)	29 (7.6%)
	Threonine (T)	24 (6.2%)
	Glutamine (Q)	14 (3.6%)
	Histidine (H)	14 (3.6%)
<i>Non-polar</i>	Alanine (A)	41 (10.7%)
	Leucine (L)	13 (3.4%)
	Valine (V)	15 (3.9%)
	Proline (P)	20 (5.2%)
<i>Special</i>	Glycine (G)	52 (13.5%)

2. *DNA with histone tails in monovalent ions* To specifically obtain DNA-histone tail effective potentials for DNA-tail part, a system containing four DNA molecules with fragments of the histone tails and other monovalent counterions including K^+ and Na^+ was simulated. However, in some preliminary simulations there were initially some convergence as well as sampling issues with the longer DNA molecules, probably due to more crowding and the tendency of longer DNA to form mesh like structures. To avoid this the length of DNA molecules was reduced to 16 base-pairs each. This is expected to give a good sampling as shorter DNA molecules will also have a

low probability of interacting at the boundaries due to PBC.

The setup will use 8 DNA molecules with the same sequence but each DNA molecule was divided into two shorter DNA molecules. Additionally, since the system is complex and the DNA molecules tend to aggregate, ten different simulation setups with the same system composition were run (using different starting velocities for the different simulations) to ensure proper sampling of the configuration space and convergence of the system.

- 8 DNA molecules with 16 base-pairs each, sequences:
 - (a) DNA1: TAATGGAACGTAGCAT
 - (b) DNA2: ATTCTTCAAGTTGTCA
 - (c) DNA3: AAACCTGATGCACACT
 - (d) DNA4: GTAACATGAGATCCCG
 - (e) DNA5: GGCTTATAGAGGGCCA

FIGURE 2.5: A snapshot of the 8DNA with histone tail fragments simulation

(f) DNA6: GCTCGTATCGACGGAC

(g) DNA7: TAGTACCCCAACCAATT

(h) DNA8: TAGGCGAAAGGAGTCT

- 48 tail fragments
- Salt: 130 mM NaCl/KCl
- 330168 TIP3P water molecules
- 15 nm cubic box
- Simulation length: 2 μ s (10 copies)

A snapshot of the 8DNA system is shown in Figure 2.5. To keep the hydrogen bonds intact, we introduced harmonic bonds at all the terminal base-pairs similar to the single DNA case. These distance restraints have been calculated using the equilibrium distances from the all-atom trajectories.

2.3 Coarse-grained models

2.3.1 DNA

Once the all-atom MD trajectories were obtained they were bead-mapped to a coarse-grained representation. The coarse-grained model used here was essentially based on a systematic bottom-up multiscale approach. In this approach a low resolution model is obtained based on properties of a high resolution model. The coarse-graining was done based on the previous model (Figure 2.6) developed by Korolev et. al. [86]. It contains two types of beads with a central D bead representing two consecutive sets of complementary nucleosides. The P bead represents the phosphate groups. So the DNA is modelled as a repeating units of 5 beads each, with each D bead bonded to four P beads. Each unit represents two base-pairs at the all-atom level. To keep terminal beads distinct, they were named DT and PT respectively. This helped later in excluding them if need be, to eliminate end-end contacts.

The essence of the model (Figure 2.6B) is in the bonds defined:

- *Bonded Potentials*

FIGURE 2.6: Initial coarse-grained model used. A) shows the 5 beads that define the D bead and the four P beads connected to it. B) shows the bonds defined in the model with the equilibrium bonds and angles. C) A superposition of the all atom model and the coarse-grained model.

1. D-D bond connecting two neighboring D beads
2. D-P bonds connecting D and the P beads
3. P-P bonds connecting the P beads along the backbone
4. P-P bond across the minor groove (improved model)

Each D bead is bonded to four P beads in the core and two beads at the termini

- *Angular Potentials*

1. D-D-D across the central DNA core
2. P-D-P across the central D bead with two opposite P beads
3. P-P-P connecting the P beads along a single strand

The model includes explicit mobile ions and charges on the DNA, and thus emphasizes the importance of electrostatic interactions in determining interactions of DNA with the ions. Each P bead has a charge of $-1e$ and the D bead is kept neutral. The CG DNA potentials were approximated by a quadratic function. The solvent is considered implicitly, which in essence is taken into account by the effective short

range potentials and screening of electrostatic interactions according to the solvent dielectric permittivity. The model does not consider sequence effects and all the potentials derived are independent of the sequence.

As mentioned before there was terminal base-pair fraying observed in the all atom simulations. So to take into account this behavior and exclude this effect from the CG model, the beads corresponding to these base pairs were treated separately.

2.3.1.1 Improvement of the model

The aforementioned CG model was used to run large-scale CGMD simulations. However, it was observed in these simulations that the model was too flexible and underestimated the persistence length of the DNA. This was based on coarse-grained simulations under various salt concentrations. In line with this, it was decided to add another bond to the model between the P-P beads along the minor-groove. This affords greater stiffness to the DNA and also produces excellent agreement with experimental results on the salt dependence of persistence length of DNA.

In addition to this, unlike the previous model [86], in the new model all the CG potentials were tabulated and calculated using IMC and not fit to any functional form. Also the work in [86] was based on very short simulations (50 ns) and for a system of several DNA that shows compaction and not an ideal all-atom model to test and validate persistence length behavior.

2.3.2 Amino acids

Following the common biochemical classification of amino acids into different types (charged, polar and nonpolar [Table 2.3]) and using some preliminary analysis on the obtained MD trajectories for DNA and histone tail fragments simulations the following classification and five bead types was assigned to the amino acids (Table-1). Also the model in general is one bead type per amino acid.

1. Negative: ASP, GLU
2. Positive: LYS, ARG

3. Polar: SER, THR, GLN, HIS
4. Non-Polar: ALA, LEU, VAL, PRO
5. Special treatment: GLY

For the CG model for the histone tails, it was observed that for bonded amino acids, they generally tend to have a similar bond length, so the same bond length would be a good approximation for all the amino acids. For the angular bonds no such similarity was seen except in case of Proline and Glycine which gave single well-defined peaks as compared to others with multiple and broad peaks. So it was decided to have three types of angular bonds with centers on Proline, Glycine and all others classified as one group.

For the histone core, a similar model to the histone tails was used, having one bead per aminoacid resolution and five bead types for aminoacid-beads and Ca-Ca pairwise bond along the chain. In contrast with histone-tails model, no angle-bonds along the chain were used. Instead multiple individual bonds were placed, providing certain rigidity to the core. These bonds were placed using following rules:

FIGURE 2.7: Illustration showing the updated DNA CG model used throughout this work. (A) Representation at the all-atom level, (B) superposition of all-atom and coarse-grained model. (C) Bonds and angles defined in the model, including the P-P bond added along the minor groove. Reproduced with permission from [162]

FIGURE 2.8: Schematic showing the bonds and the three different angles, o , g and p defined in the coarse-grained model for amino acids. P refers to Proline, G for Glycine and $Others$ for all the other amino acids

- Each bead of the core was connected to two nearest core beads (preferably belonging to another chain), to make all core-beads have four bonds with other core beads (except terminal beads in the chain).
- Each D-bead of DNA was bonded to the nearest core-bead.
- The distances used for the bond placing were taken from the 1kx5.pdb structure.

Tails were bonded to the core by Ca-Ca-bond.

2.4 RDF calculation for DNA systems

Once the trajectory was bead-mapped according to the model described above in Figure 2.7 the radial distribution functions were calculated. These radial distribution functions serve as the connecting point between the all-atom and the coarse-grained model. These RDFs were then used as inputs for the IMC procedure to extract the effective potentials for the coarse-grained model.

- **Non-bonded**

D-D, D-P, D-Na/K, D-Cl, P-P, P-Cl, P-Na/K, Na/K-Cl

With respect to atom i , at a distance r , consider another atom j and a shell of width dr , the non-bonded RDF is defined as:

$$f_{ij}^{NB}(r) = \frac{N(r)}{4\pi dr \left(r^2 + \frac{dr^2}{12} \right)} \times \frac{V}{N_{total}} \quad (2.2)$$

where $N(r)$ is the number of atoms of type j in the spherical volume element at distance r

V is the system's volume and

N_{total} is the total number of j -atoms

For the non-bonded RDF calculations the pairs of beads connected by bonds were excluded.

For bonds and angles we used bond length distributions and bond angle distributions for each bead connected by bonds. The pairwise bonds calculated are:

- **Bonded**

D-D, P-P, D-P

- **Angular**

D-D-D, P-D-P, P-P-P

For the non-bonded and bonded interactions the cut-off used was 25 Å.

2.5 Inverse Monte Carlo procedure

The IMC method has been used here for deriving the effective potentials from the RDFs which are purely based on the all-atom MD and the topology of the CG model. The IMC procedure uses a sampling technique based on the Metropolis algorithm and has been shown to be successful for many systems [103, 105, 120]. The IMC procedure has some notable advantages in terms of speed and its spanning of the configuration space.

As compared to the IBI approach it also considers cross-correlation terms in the RDFs, and so is more suitable for complex systems. Also it is more robust as compared to a normal MD engine as it only visits the configuration space sampled in the all-atom simulations owing to its option of keeping infinite repulsive potentials at the boundaries. For the shape of the trial potential as it does not require finite derivatives at each point, it is much easier to impose stiff potential walls for regions of the phase space not visited in the all-atom simulation. Finally as it uses both local and global transformations it is relatively faster in sampling this space.

Since the IMC method also takes into account the correlation between different RDF, it has good convergence for multi-component systems like the NCP [87, 104].

2.5.1 Derivation of CG effective potentials

There are several differences in the way the CG potentials were defined in the previous CG model by Korolev et. al. [86]. The bonded potentials were fitted to a harmonic function and the non-bonded parameters were derived using a term consisting of electrostatic and truncated Lenard-Jones potentials. In the updated CG model all the effective potentials are tabulated, derived using IMC in a rigorous way. For each interaction term in the coarse-grained term, a corresponding RDF is calculated from the underlying all-atom trajectory and then used to derive the corresponding effective potential.

The IMC calculation was carried out using the MagiC software [120]. For the non-bonded potentials for single DNA systems, the DNA-DNA potentials were zeroed as there was just a single DNA molecule with no inter-DNA interactions. All other potentials were started from a trial value of zero. The bonded potentials were initialized using the potential of mean force. Electrostatic interactions were present between all charged non-bonded beads, i.e. excluded only for those beads which are directly bonded by pair or angle bond.

$$U_{PMF} = -k_B T \ln(g(r)) \quad (2.3)$$

The IMC procedure uses an iterative approach, so starting with the trial potentials after about 20 iterations, convergence in potentials is achieved. Each iteration

consists of 100 parallel Monte Carlo sampling simulations, and each simulation runs for 250 million MC steps. The cutoff for the RDF and non-electrostatic part of the effective potentials is 25 Å. Long range interactions were treated using Ewald summation method, with 40 Å being the real-space cutoff. Dielectric permittivity constant used was 78.

The IMC method relies on the fact that the Hamiltonian of a system can be described as

$$H(q_i) = \sum_a K_a S_a(q_i) \quad (2.4)$$

here K_a is a set of constants spanning the space of Hamiltonians given by $S_a(q_i)$. This is the general equation. Narrowing down to a more specific system where these potentials are pair potentials, we have the continuous case defined by the delta functions:

$$H = \sum_{i,k} \Psi(|q_i - q_k|) = \sum_{i,k} \int_0^\infty \Psi(r) \delta(r - |q_i - q_k|) dr \quad (2.5)$$

where, for radial distribution functions we have $r - |q_i - q_k|$ representing the distance range and the delta function gives the number of particle pairs having the potential $\Psi(r)$ at that distance range. For the radial distribution function $g(r)$ we have

$$\langle S_a \rangle = 4\pi r_a^2 \text{ dr } g_a(r) \quad (2.6)$$

If for all α these canonical averages are equal, the Hamiltonians are equivalent. Now these averages can be obtained from either structural data or reference all-atom MD simulations. In equation (2.3), as $\langle S_a \rangle$ are functions of K_α , if we solve the reverse problem we obtain K_α

$$\Delta \langle S_\alpha \rangle = \sum_\gamma \frac{\partial \langle S_\alpha \rangle}{\partial K_\gamma} \Delta K_\gamma + O(\Delta K^2) \quad (2.7)$$

where

$$\frac{\partial \langle S_\alpha \rangle}{\partial K_\gamma} = \frac{\partial}{\partial K_\gamma} \left[\frac{1}{Q} \int dq S_\alpha e^{-1/k_B T \sum_i K_i S_i} \right] \quad (2.8)$$

$$= \frac{1}{k_B T} (\langle S_\alpha \rangle \langle S_\gamma \rangle - \langle S_\alpha S_\gamma \rangle)$$

Using an iterative process by starting from a trial potential one can calculate the averages $\langle S_\alpha \rangle$ and also their product $\langle S_\alpha S_\gamma \rangle$ and using the S_α^{ref} calculated from the all-atom MD we get the difference for every iteration

$$\Delta \langle S_\gamma \rangle = \langle S_\gamma \rangle - S_\alpha^{ref} \quad (2.9)$$

and substituting to Eq. (2.7) we get ΔK_α . The higher order terms are ignored. The process is repeated until the reference RDFs S_α^{ref} are reproduced.

2.6 Coarse-grained simulations

The effective potentials obtained above are then tested in large-scale simulations with a much better computational efficiency which wouldn't have been feasible at the all-atom level.

2.6.1 Persistence length calculation

As a first test of the system, the persistence length of the DNA was calculated at different salt concentrations. Since explicit ions were used in this CG model, the underlying potentials can be used for a system with any salt concentration using the same set of potentials assuming their transferability. In this case we calculated the persistence length over a range of monovalent salt concentrations from 0.1 mM to 100 mM and then assess its dependence on salt concentration.

Since potentials were obtained for the four underlying all-atom force fields, simulations were done using these set of potentials over the entire salt concentration range. The DNA length used was 500 bp. All CG simulations were run using the LAMMPS software [138] in the NVT ensemble maintained by the Langevin thermostat with a relaxation time of 10 ps. Simulations were started with a linear DNA molecule. Equilibration was done for 1 ns with a time-step of 2 fs. The production run was done at at time-step of 20 fs. Electrostatic interactions were computed

using the PPPM method with a 40 Å real space cut-off and relative force error of 10^{-5} .

Using Eq. (1.1), $\ln(\langle \cos\alpha \rangle)$ is plotted for different values of L_C and the corresponding slope gives the persistence length. However it must be mentioned that there is an approximation in assuming the correlation function is exponential and the linear fitting is fully valid. For a different approach to estimate the persistence length please see the work of Mitchell et. al. [122].

To calculate the persistence length, the 500 bp DNA is divided into 5 equal parts, 100 bp each and then the persistence length is calculated over that contour length ($\approx 330\text{\AA}$), thereby giving 5 values. The average of these 5 points is then taken to be the persistence length for that concentration and the standard deviation is then calculated to estimate the error range. An illustration showing the 500bp DNA is shown in Figure 2.9

2.7 Data analysis

For all the all-atom simulations, to study the effect of various force fields, the DNA base-pair geometry and helical parameters (see section 3.1) were studied. This was done using the 3DNA program [98, 99]. Also for data-analysis on the trajectory files and plot generation the do_x3dna program [89] for GROMACS was used. The R program (version 3.4.2) [164] was also used to generate most of the plots and data analysis. Experimental data for the experimental B-DNA structures was taken from the NDB database [12]. The structures with following ndb ids were used:

- *bd0001, bd0016, bdj019, bdj037, bdl001, bd0005, bd0018, bdj025, bdj051, bdl005, bd0006, bd0019, bdj031, bdj052, bdl020, bd0014, bdj017, bdj036, bdj060, bdl084*

FIGURE 2.9: The 500bp DNA molecule used to calculate the persistence length for all the force fields and salt concentrations. The different coloured pieces show the 5 different fragments over which the calculation was done. Ions have not been shown for clarity.

Chapter 3

Results and Discussion: all atom simulations

This chapter describes the results obtained from the all-atom simulations, their convergence and how much of the sample space has been explored in the all-atom simulations. Additionally the behaviour of the different force fields in studying the base-pair geometry and backbone properties of DNA will be discussed. A detailed study of the behaviour of the base-pair geometry is very important as it can play important roles not only in modeling the DNA flexibility but also protein DNA interactions during various cellular processes.

Since the all-atom simulations form the basis of the bottom-up CG approach, it is important to verify the quality and behaviour of the underlying all-atom force fields and their ability in properly treating the helical and backbone properties of the DNA. To verify that the all-atom simulations are well converged at the all-atom level and they properly preserve the helical structure of DNA a study of the base-pair parameters and other helical parameters is first done. Performance of the force fields is then accessed by a comparison with the experimental values of these parameters from the crystal structures. This solves two purposes, first it tells us about the convergence of these all-atom simulations and also forms the basis for the choice of the force field to be used for modelling of DNA structure and other properties.

3.1 DNA backbone and helical properties

Figure 3.1 describes the evolution of three of the 12 base pair and base-step parameters averaged for each base-pair over the whole time range of the simulations for all the force fields with a similar system composition. Note that all the simulations were started from the canonical form of B-DNA. The schematic for each of the three parameters is shown on the side pane of the figure color coded according to its plotting color. A comparison is also made with the corresponding experimental values averaged over B-DNA crystal structures for those parameters. In the first few hundred nanoseconds of all four simulations, the three chosen parameters were stable and lie within the range corresponding to their experimental values for B-form of DNA.

FIGURE 3.1: Parameters characterizing base-pair geometry for simulations with 40-mer DNA with 130 mM KCl using four different force fields, (A) parmb0, (B) parmb1, (C) CHARMM27 and (D) CHARMM36. Y-axis shows the magnitude of each parameter in degrees. Colored bands show the corresponding range of values for crystal structures of B-DNA. Also the corresponding color-coded parameters are shown on the side.

From Figure 3.1A and 3.1B we can infer that AMBER force fields behave quite well, with minor fluctuations in opening, propeller and twist. Most of these fluctuations were reversible and well within the standard deviation described by the experimental structures (colored bands in Figure 3.1) of B-DNA. Upon visual analysis no base-pair fraying was observed, thus the base-pair geometry was conserved,

except for some minor opening events that happened around 800 ns. Parmbsc1 shows some fluctuations outside the experimental range for some parameters like opening but these were reversible. Both AMBER force fields maintain the canonical DNA structure throughout the simulation with the exception of C36, as seen in the average structures shown in the Appendix (Figure A.1).

For C27 there were no serious fluctuations observed for the DNA structure and most of them are within the experimental values for B-DNA as seen in Figure 3.1C. The most ill-behaved force field at least in terms of these three parameters is C36 as the fluctuations after the first microsecond became much larger and outside the range that is expected for physically viable double helical structures. These fluctuations were not reversible and persisted with a tendency to increase towards the end of the simulations at $2.5\mu\text{s}$. In summary, C27 FF did not show such behavior and the

FIGURE 3.2: Comparison of four different base-step and base-pair parameters. Variation of (A) shift, (B) tilt, (C) shear and (D) buckle. Black line and grey band in each of the panels show an average value and a standard deviation panel calculated from the DNA crystal structures. Simulation data for each base-step/base-pair was averaged over the entire productive run.

DNA geometry was perfectly maintained throughout the $2\mu\text{s}$ run, and is the most stable among all the four tested force fields as far as the geometry of the base-pair is concerned.

To look deeper into the behavior of these force fields, a few other DNA parameters were analyzed for each base-step/base-pair averaged over the whole production

FIGURE 3.3: Comparison of four different base-step and base-pair parameters. Variation of (A) stretch, (B) propeller, (C) slide and (D) roll. Black line and grey band in each of the panels show an average value and a standard deviation panel calculated from the B-DNA crystal structures. Cyan line shows the average value of slide for A-DNA. Simulation data for each base-step were averaged over entire productive run.

part of the trajectory as shown in Figures 3.2, 3.3 and 3.4. For the parmbsc0 FF, a few of the base-step parameters are described very well like shift, tilt, rise and twist, however slide and roll are well outside the experimental range. The base-pair parameters on the other hand give good values in comparison with B-DNA crystal structure averages except for some fluctuations for stretch. For parmbsc1, which is the improved and latest version of AMBER family of force fields, some fluctuations are quite large, especially for shift, stretch and slide and outside the experimental range for B-DNA structures. For the CHARMM family of force fields, C27 performs fairly well for most of the parameters, except for roll and stagger, for which both C27 and C36 give values outside the experimental range. A general outlier among the four force fields is C36 (yellow curve) as some of the base-step and base-pair parameters along one end of the DNA give large unphysical values. In all the figures it can be seen that for the first to the ninth base-pairs, C36 (yellow) reports the

FIGURE 3.4: Comparison of four different base-step and base-pair parameters. Variation of (A) stagger, (B) opening, (C) rise and (D) twist. Black line and grey band in each of the panels show an average value and a standard deviation panel calculated from the DNA crystal structures. Simulation data for each base-step were averaged over entire productive run.

values incompatible with double helical structure. This is different from end-effects as has been observed in previous studies [33, 196] as these kinds of distortions are observed even towards the middle of the DNA molecule for base-pairs like 9-10 steps from one end.

3.1.1 A-DNA versus B-DNA

It has been noted previously that the slide parameter can be used to distinguish the A-form and the B-form of DNA [131]. Values below $-0.8 \text{ \AA}$ are characteristic for A-form DNA and values above $-0.8 \text{ \AA}$ are characteristic for B-form DNA. Data presented in Figure 3.5 shows that parmbsc0 and parmbsc1 (to a lesser extent) demonstrate more affinity to A-form DNA than the CHARMM force fields. Figure 3.5B shows an example of slide distribution for the GG base pair determined in the simulations with clear indication of the parmbsc0 FF bias towards the A-form. Unlike other force fields, parmbsc0 (red curve) clearly falls below the range

determined for the B-DNA from X-ray structures (grey shaded area). Similar behaviour was observed in other base-steps (C) GG, and (D) GC.

FIGURE 3.5: (A) Differentiation between A-form and B-form DNA based on slide. A value of -0.8 (cyan line) defines the boundary between the A-form and B-form. Parmbsc0 (red) is more biased to A-form compared to other force fields. Black line and shaded area show respectively the average value and its standard deviation determined from the crystal structures of B-DNA. (B) The average distribution for the AG, (C) GG and (D) TA base-step for the four force fields.

However it has been recently observed that slide alone can be insufficient to determine the form of DNA. In addition to slide other parameters were analyzed like X-displacement, inclination and pseudo rotation angle. Where pseudorotation is defined in terms of the five angles describing the sugar ring angles as:

The conformation of the sugar ring is described by the five endocyclic torsion angles: O(1')-C(1'): θ_3 , C(1')-C(2'): θ_4 , C(2')-C(3'): θ_0 , C(3')-C(4'): θ_1 and C(4')-O(1'): θ_2 .

and the pseudorotation angle according to [2] is defined in terms of these angles as:

$$\tan P = \frac{(\theta_2 + \theta_4) - (\theta_1 + \theta_3)}{2\theta_0(\sin 36 + \sin 72)} \quad (3.1)$$

In addition the RMSD with respect to starting A-DNA and B-DNA structures was also calculated to quantify this issue. The variations of these parameters are shown in Figure 3.6. Since the transition mostly occurs in the 100-1000 ps time-frame, only the first 1 ns data is shown. Two parameters, namely x-displacement and the RMSD variation suggest a state intermediate between A-DNA and B-DNA. However for inclination angle and pseudo rotation angle the values signify a more A-DNA state.

FIGURE 3.6: Variation of different helical parameters over the first 1 ns. (A) shows the Inclination, (B) X-displacement and (C) shows the phase angle of pseudo rotation. In all the plots the A-DNA average value is shown in cyan and B-DNA value in grey. (D) shows the RMSD variations with respect to A-DNA starting structure (solid lines) and B-DNA starting structure (dotted lines).

3.1.1.1 Parmbsc0 has a preference for A-DNA conformations

To check if the A-DNA behaviour was just a feature of one of the trajectories or indeed a general feature of the force field, the above analysis was done with other available trajectories with the same force fields. The forthcoming analysis is based on trajectories run by other members of the group. A brief description of the system configuration follows.

- Four 36 base-pair DNA molecules
 1. 5'-ATTAATGGAACGTAGCATATTCTTCAAGTTGTCACG-3'
 2. 5'-CAAAACCTGATGCACACTGTAACATGAGATCCCGCG-3'
 3. 5'-TCGGCTTATAGAGGGCCAGCTCGTATCGACGGACCG-3'
 4. 5'-GCTAGTACCCACCAATTTAGGCGAAAGGAGTCTGC-3'
- Salt concentration: CoHex³⁺ 50 mM, Mg⁺ 50 mM, 35 mM Na⁺
- Box size: 16nm
- Trajectory length:
 1. Parmbsc0: 2 x 1 μ s
 2. CHARMM27: 3 x 1 μ s
 3. CHARMM36: 3 x 1 μ s

Since there were multiple trajectories containing the same four DNA molecules and similar conditions, the base-pair wise variation of slide was calculated for each DNA sequence and each trajectory for all the three force fields used. All the trajectories were started with similar starting conditions from B-DNA. Figure 3.7 shows the variation of slide over the two parmbsc0 1 μ s trajectories for each base-step for the four DNA molecules present. As we can see similar to the single DNA system, slide values are more towards the A-DNA experimental average of -0.8 for all the four sequences tested for both the trajectories started with the same B-DNA initial structure.

For the CHARMM force fields (Figure 3.8), we can see that C27 produces slide values very close to B-DNA average for all the four DNA sequences (data shown

FIGURE 3.7: (A) Average value of slide for two parmbosc0 trajectories for each individual base-step in DNA 1. Black line is the average value from crystal structures of B-DNA and cyan line the average for A-DNA crystal structures. (B) shows the slide values for DNA2, (C) DNA3 and (D) DNA4. Each figure shows the values for both the trajectories for parmbosc0, Traj1 and Traj2.

for DNA1 and DNA3) for all the three trajectories. Also it can be seen that there is little variation among individual trajectories. However all the three C36 trajectories vary in the values and sample both A-DNA and B-DNA forms for all the four DNA sequences (data shown for DNA1 and DNA3).

It should be noted that the four DNA sequences tested contain all the 10 unique base-steps possible with the four nucleobases A, G, C and T. To get a more general overview of how these force fields treat slide, it was averaged over all the 4DNA trajectories for the three force fields for each of the 10 unique base-step belonging to each DNA molecule. The resulting distribution curves are shown in Figure 3.9. As we can see, the red curve in all the figures is always to the left of the average B-DNA value from crystal structures, demonstrating that parmbosc0 samples more of A-DNA structures. Also for C36, most of the base steps show a binormal distribution with more than one peak for the slide parameter. C27 on the other hand produces slide values in near agreement with the experimental values for B-DNA as can be inferred from the blue curves in Figure 3.9. Yet again, in terms of base-pair/base-step parameters and helical structure C27 is the most stable force field for DNA.

FIGURE 3.8: (A) and (B) show the average values of slide for three CHARMM27 trajectory for each individual base-step in DNA 1 and DNA three respectively. Black line is the average value from crystal structures of B-DNA and cyan line the average for A-DNA. (C) and (D) show the slide values for each CHARMM36 trajectory for DNA 1 and DNA 3 respectively.

3.1.2 Torsional angles

The torsional angles have been used primarily in tuning most of the DNA force fields available and thus are an important factor in force field parameterization as they define stability and flexibility of the DNA backbone. To quantify the performance of these force fields it is necessary to study their treatment of DNA backbone torsional angles. To this effect, torsional angles characterizing the DNA backbone sugars were analyzed and compared with experimental values averaged over naked DNA crystal structures. The average of these angles is shown in Figure 3.10. The experimental averages of B-DNA crystal structures are also shown for each angle.

The most recent version of the CHARMM family, C36 was specifically parameterized to produce experimentally viable proportions of BI and BII states as compared to C27 which underestimated the BII population. To calculate the ratio of these conformations, the difference $\epsilon - \zeta$ is calculated for each frame of the trajectory

FIGURE 3.9: Slide distributions for 10 unique base steps present in the four DNA sequences for the three force fields analyzed. For each base-step, the value of slide was averaged over all the trajectories for that particular force field.

FIGURE 3.10: Distribution of dihedral angles parameters calculated from the all-atom MD trajectories of all-atom MD simulations of the 40 bp DNA with four force fields (indicated in the graphs); (A) α ; (B) β ; (C) δ ; (D) ϵ ; (E) γ and (F) ζ . Four dihedral angles were analyzed for the four force fields and compared them with the average values determined from the DNA crystal structures (black vertical line in each graphs). Simulation data for each dihedral angle was averaged for each DNA nucleotide over the entire productive run.

and then the proportion of each state (BI or BII) over the entire simulations is calculated. The results are shown in Table 3.1. The experimental values in Table 3.1 correspond to the average of the B-DNA solution structures (Experiments 1) and crystal structures (Experiments 2) as in [68]. As observed in previous studies [64], C27 underestimates the population of BII by about 7-8% as compared to experimental structures but these simulations were based on limited sampling (<2 ns). But as we can see even at longer timescales the BII population is underestimated by both parmbsc0 and C27, if we average results for the whole sequence.

To understand the sequence effects in the BI-BII conformation space, the analysis was repeated for the four force fields, but this time the averaging was done not over the whole DNA but similar base-steps. The calculation was done for all the 16 unique base-steps found in the DNA sequence used. If a base-step occurred more

TABLE 3.1: Percentage of BI/BII conformations for strand 1 (strand 2)

Force field	BI (%)	BII (%)
bsc0	99.55 (99.51)	0.45 (0.49)
bsc1	77.06 (80.57)	22.96 (19.43)
C27	94.98 (95.46)	5.02 (4.54)
C36	80.69 (82.54)	19.31 (17.46)
Experiments 1	80.31	19.69
Experiments 2	79.5	20.5

TABLE 3.2: Percentage of BII conformations for each base-step

Base Step	Strand-1				Strand-2			
	bsc0	bsc1	C27	C36	bsc0	bsc1	C27	C36
CA	0.25	34.52	3.95	16.30	0.17	35.78	18.41	43.94
CG	0.85	30.65	21.53	41.26	0.73	34.57	13.73	30.48
TA	0.24	29.27	10.55	32.75	0.21	25.48	5.61	18.56
TG	0.14	14.61	2.61	17.81	0.20	10.17	5.07	13.54
AC	0.25	11.14	0.80	4.84	0.22	12.23	0.47	3.49
AT	0.07	2.25	0.28	13.51	0.08	3.03	0.27	15.43
GC	0.24	16.44	1.49	5.47	0.51	26.56	1.74	8.40
GT	0.20	14.71	1.39	7.01	0.08	15.14	1.15	7.07
AA	1.27	41.45	3.61	18.64	1.00	27.94	1.87	12.70
AG	1.08	32.72	0.72	4.18	0.64	23.64	1.06	3.49
GA	1.02	26.72	7.55	37.85	3.63	62.36	13.68	33.00
GG/CC	0.96	56.20	4.60	38.04	0.02	6.06	1.64	27.44
CT	0.00	2.68	0.73	3.46	0.00	0.16	0.99	3.71
TC	0.03	5.66	2.77	11.16	0.04	8.14	2.03	22.41
TT	0.00	0.47	0.36	9.28	0.00	0.41	0.45	9.44

than once, averaging was done and calculation was done independently for both the strands. Results are presented in Figure 3.11 and we can see that certain base-steps have a higher propensity for BII particularly, GG and CG in most of the force fields except parmbsc0 and CHARMM27. Both of these force fields underestimates BII for most of the base-steps. In terms of BII%, parmbsc1 and CHARMM36 perform well and sample conformations in line with experimental structures for most of the base-steps. For clarity the results are also tabulated in Table 3.2. Please note that Strand 1 did not have any *CC* base-step and there was no *GG* step for Strand 2, so in total there are 15 base-steps for each strand.

To summarize, parmbsc0 underestimates the BII population the most (Table 3),

FIGURE 3.11: The percentage of BI and BII conformations for each of the 15 base-steps in the DNA sequence for strand 1 (A) and strand 2 (B). The grey line and band show the experimental values for the corresponding base-steps.

which has not been observed before [50, 134]. CHARMM36 populates more BII states as compared to the CHARMM27 version but in doing so it becomes too flexible and unstable and thus distorts the DNA-backbone and other helical properties and gives rise to unphysical structures (see below). The analysis was also done over the whole sequence without any averaging and the results are shown in (Figure A.2). However, the trend is similar for the ones observed in Figure 3.11.

3.1.3 Elastic properties at all-atom level

An important aspect of a force field is how it treats the elastic properties of the double-helix. Elastic properties of DNA influence many biological processes since they affect the behavior of a long DNA molecule. Bending can be characterized using the bending angles which are calculated between the tangents of the global helical axis after repeat of a certain DNA length. Stretching and twisting motions are calculated from the sum of the helical rise and the helical twist respectively of the DNA molecule. These quantities can be calculated from the simulation trajectory, as well as the covariance matrix which in turn gives the elastic matrix showing coupling between these parameters.

We have computed these properties using the *do_x3dna* module. At the all-atom

FIGURE 3.12: Comparison of helical flexibility for the four force fields. (A) Twist rigidity modulus [pN.nm²], (B) stretching modulus [pN] and (C) twist-stretch coupling [pN.nm]. Experimental values with their standard deviations are shown respectively as black line and shaded area.

level we compared the twist rigidity modulus, stretching modulus and the twist-stretch coupling to get a picture of the elastic properties of the double helix simulated by each of the four force fields (Figure 3.12). These quantities have also been measured using experimental techniques like AFM and optical tweezers. Experimental values for twist rigidity, stretching modulus and twist-stretch coupling are 460 ± 20 pN.nm² [55, 19], 1100 ± 200 pN [23, 21] and -90 ± 20 pN.nm [55] respectively. Comparison of the atomistic simulation data on DNA elasticity shows that CHARMM36 produces values close to the experimental ones during the initial few hundred ns of the simulation, but it strongly deviates from the experiment from about the same time of the simulations when strong fluctuations of DNA base pair parameters were observed (see Figure 3.1).

Except the CHARMM36 data, most of the trajectories show normal behavior for all the three parameters. Parmbsc0 provides stretching and torsional modulus values most close to the experimental data. CHARMM27 and to a lesser extent parmbsc1 overestimate the elasticity modulus. Note that the few abrupt changes could be due to the low number of frames used to display the data, for each trajectory, data points were taken after an interval of 500 frames.

3.2 CHARMM36 distorts backbone DNA structure

In the previous section we made a general comparison of the force fields at the base-pair level and how different force fields affect the backbone structure and the helical features of the DNA. From the yellow curves corresponding to CHARMM36

in Figures 3.2, 3.3 and 3.4, a general conclusion can be made that CHARMM36 samples base-steps and base-pair parameters well outside the experimental range of values for many of the base-steps. This aberrant behavior compels a more deeper examination of the force field and how it affects the DNA structure and backbone.

Previously [51] it has been observed that base-pair fraying and backbone distortions occur in CHARMM36-based 1 μ s long simulations of DNA. It has been observed that CHARMM36 suffers from minor fluctuations and some disturbances in base-pair geometry mostly at the termini of the DNA. The simulations done in that work were 1 μ s in length and the DNA length was way shorter than the one considered in the present work. Still the fluctuations were within the experimental range.

Data from our simulations of a single 40 bp DNA in the presence of KCl shows that the DNA structure was generally stable during the initial part of the trajectory with minor violations happening at 300-400 ns. However, after 1 μ s, the structure was severely deteriorated with all parameters going out of the acceptable value range from experiments. To get an overview of this difference in behavior after 1 μ s, we analyzed our longest CHARMM36 trajectory (Traj1) over the first 1 μ s and the remaining 1.5 μ s with respect to their fluctuations from the starting structure. The RMSD was computed over all the frames, taking the structure at $t=0$ as reference (see Figure 3.13) and a distribution curve of the RMSD was plotted. The curves are plotted using the *stat_density* function in R, which is the smoothed version of a histogram of RMSD of all the frames using the structure at $t = 0$ as reference. The blue curve is the distribution for all the frames from 1-2.5 μ s and orange curve is for the trajectory fragment 100-1000 ns. Clearly the blue curve has a peak around 8 Å much higher than for the orange curve, indicating that the structure changes by a high degree after about 1 μ s. This in general proves that C36 behavior changes after 1 μ s and longer simulations are needed to see this effect.

In order to ensure that this observed behavior was not an occasional isolated event for just one trajectory, we analyzed three additional DNA trajectories simulated with CHARMM36 force field, referred as Traj2, Traj3 and Traj4 (the trajectory analyzed in the previous sections referred as Traj1, see Table 1).

Two of the base-step parameters are shown in Figure 3.14 and we can see indeed the fluctuations are well outside for all the four trajectories. In Figure 3.14A and B the distribution and bp-wise variation respectively of rise is shown for the four

FIGURE 3.13: Distribution of RMSD for the CHARMM36 based trajectory Traj1 for two fragments of the trajectory, 100 ns-1 μ s (orange curve) and 1 μ s - 2.5 μ s (blue curve). Other curves are also shown for which 2 (khaki) and three (brick-red) residues have been excluded for the RMSD calculation. The dark-red curve has reference structure at t=100 ns.

FIGURE 3.14: (A) and (C) show the distribution of two base-step parameters, namely rise and twist respectively for the four CHARMM36 trajectories. (B) and (D) show the variation of these parameters for the whole 40-bp DNA molecule averaged over the whole trajectory.

C36 trajectories. It is clear that the values sampled deviate from experimental averages for many base-steps. This behavior is indeed observed for all the four

C36 trajectories. Similarly in Figure 3.14, panels C and D show distribution and bp-wise variation respectively of twist for the four C36 trajectories, which again deviate well outside the experimental range of values. In addition, two other all-atom trajectories using a shorter DNA with no harmonic constraints imposed on the terminal base-pairs were also used to study the behavior of C36 and the results are shown in the Appendix (Figure A.4, A.5 and A.6). Traj5 refers to a trajectory with a 38-bp DNA molecule with a single base-pair removed from each side while Traj6 refers to a 25 bp DNA trajectory with 7 base-pairs removed from the 5' end and 8 base-pairs removed from the 3' end. The results show that CHARMM36 distorts DNA structure irrespective of DNA length and sequence.

To examine how these distortions change the elasticity of the DNA at the all-atom level, the bending and twisting moduli were again computed for these four trajectories. Figure 3.15 shows variation of the DNA elasticity parameters during the whole simulation time, and further demonstrates that the unphysical effects

FIGURE 3.15: Comparison of helical flexibility for the four C36 trajectories. (A) Twist rigidity modulus [pN.nm^2], (B) stretching modulus [pN] and (C) twist-stretch coupling [pN.nm]. Experimental values with their standard deviations are shown respectively as black line and shaded area.

caused by the use of CHARMM36 were not just an instance of one trajectory; similar effects were also observed for the other three trajectories. After a few hundred ns, the moduli follow an irregular trend for each trajectory and after 1 μs attain values which are far apart from the experimentally observed ones. Lower stretching and twisting moduli can be understood as lower resistance of DNA molecule to undergo these changes when the base-pair geometry has been severely disturbed and hence stacking and hydrogen bonding interactions have been seriously disrupted. For the stretch twist coupling, DNA overtwists when extended and conversely extends when overtwisted, thus giving a negative constant. This is the case for all the other three force fields except CHARMM36 (3.12) where it tends to undertwist when extended. This again can be due to the disturbances at the base-pair level and DNA backbone distortion.

FIGURE 3.16: Base-pair opening for CHARMM36-based trajectory using 130 mM KCl and a 40 bp long DNA (red points), and DNA snapshots showed in ‘New Ribbon’ representation for (A) 2.5 μ s long Traj1 and (B) 1.2 μ s long Traj2. The transparent band shows the range of the crystal structures for base pair opening.

FIGURE 3.17: Variation of base-pair opening over 1.2 μ s trajectory of the all-atom MD simulation with the CHARMM36 FF (40 bp DNA, 130 mM NaCl, Traj3 (A) and Traj4 (B); see Table 2.1). Light-red shaded area shows the range of the values observed in the DNA crystal structures.

To see how this behavior at the base-pair level affects the structure of DNA, visual examination of the trajectories was done for all the four trajectories. For both the KCl trajectories with CHARMM36, the structures were examined at different time-points. Figure 3.16A shows such a time-series analysis of base-pair opening

with the corresponding structures produced at that timepoint. The representation used is ‘New Ribbons’ in VMD. It can be seen for the longest trajectory namely Traj1, even at around 400 ns one of the base-pairs is flipping out of the DNA backbone. This improves with time for a short time (750 ns) but again severe distortions are seen in the later part of the trajectory for the later part of the trajectory. It should be noted that these distortions are irreversible and continue to show up in the DNA structures around $2.5\mu\text{s}$. This further corroborates the data seen in Figure 3.1D where the disturbances in the other helical and base-step parameters are observed which are manifested in the structures seen in Figure 3.16. Similar behavior, albeit not as bad as Traj1 is seen for the second trajectory, Traj2 in Figure 3.16B. Deformation of the double helical backbone structure as well as base pair opening for the later stages of both the KCl trajectories can be clearly seen in Figure 3.16. This analysis was also repeated for the two trajectories with NaCl using CHARMM36 and results are shown in Figure 3.17. It can be observed that the fluctuations and the corresponding structures produced are still not corresponding to physically viable structures in Traj2, Traj3 and Traj4, but better than Traj1. This can be due to the shorter trajectory lengths for these simulations.

Additionally in some preliminary analysis of other C36 trajectories it was seen that these distortions also occur for other sequences, thus eliminating any sequence bias in the current analysis. For some preliminary data please see the Appendix (Figure A.3).

3.3 Conclusions

Given the excellent agreement of CHARMM27 with experiments for various base-pair and base-step parameters and also its stability in preserving the backbone of the DNA, it was chosen as the force field of choice for further model development at the multiscale level.

This conclusion was also strengthened given its performance at the coarse-grained level and how accurately it predicts the salt concentration dependence of persistence length of DNA (see next chapter). Still to make a comparison with other force fields at the coarse-grained level, effective CG potentials were calculated for other

force fields. But unless otherwise mentioned CHARMM27 is used for most of the simulations with DNA for model development.

It was also observed that the latest version of CHARMM, CHARMM36 mistreats DNA structure and distorts the backbone. This is reflected in a general disagreement with experiments for almost all the base-pair parameters and even DNA flexibility. It does a good job in reproducing the BI-BII conformation balance in agreement with experiments, but in doing so, the DNA becomes too flexible and thus suffers in its stability.

Chapter 4

Results and Discussion: coarse-grained simulations

As mentioned previously the main aim of this study is to derive effective coarse-grained potentials to facilitate the long term aim of the simulation of biological systems at physiological conditions. This section describes the overall process of modelling the system at a coarse-grained level and the derivation of effective potentials to run the coarse-grained simulations.

The converged all-atom trajectories obtained in the previous chapter were bead-mapped to the coarse-grained model and then the RDF calculation was done using these bead-mapped trajectories. The IMC procedure uses underlying RDFs derived from all-atom simulations to derive the effective potentials.

The obtained effective potentials are then used to run coarse-grained MD simulations to derive the persistence length of DNA. A comparison is then made with experimental persistence length experiments and other coarse-grained models to compare the accuracy of the coarse-grained model. In addition, multiple CGMD simulations were done with the four different sets of potentials obtained for the four FF, namely parmbse0, parmbse1, C27 and C36. This included doing four set of CGMD simulations for 7 different concentrations to compare the performance of the underlying all-atom force fields at a coarse-grained level.

FIGURE 4.1: Comparison of intra-molecular DNA RDFs obtained from four different underlying all-atom force fields. These are based on the all-atom trajectories using single 40 bp DNA with KCl salt concentration of 130 mM for different force fields discussed in the previous chapter.

4.1 Effective potentials using IMC

There is an inherent assumption for most bottom-up based CG models that the all-atom simulations are well-converged and properly describe the properties of the system. Given the results in the previous chapter, this assumption is mostly valid for the AMBER set of force fields and also CHARMM27 force field. For CHARMM36 simulations, the RDFs were calculated based only on the "clean" parts of the trajectories and the part showing distorted DNA structures was omitted. This corresponds to two independent trajectories, 1100 ns and 600 ns long. The bonded RDFs for the 4 force fields are shown in Figure 4.1. Most of them are similar except for some differences in the angle RDFs. This can be attributed to the different elastic and backbone properties of the force fields at the all-atom level.

The corresponding converged bonded potentials obtained after several IMC iterations are shown in Figure 4.2. It must be noted that the potentials shown here are the final converged potentials obtained by doing over 20-25 set of iterations in a parallel way. The initial set of bonded potentials were derived using PMF (see Methods) and the non-bonded potentials were initialised from zero. Each iteration was run for around 25 million Monte-Carlo steps including 10 million steps for equilibration on a total of 200 parallel cores. The RDFs corresponding to the final

FIGURE 4.2: Comparison of intra-molecular DNA potentials obtained by using the IMC procedure for the RDFs mentioned in previous Figure 4.1. These are the final converged potentials.

FIGURE 4.3: Comparison of DNA-ion and ion-ion RDFs obtained from four different underlying all-atom force fields.

converged potential was then compared to the reference RDF to ensure that a well converged potential has been obtained.

The non-bonded RDFs and potentials are shown in Figure 4.3 and 4.4. Most of the RDFs are similar for the four force fields except a few differences. One is the shortened P-P minor groove bond length with parmbsc1 force field. In the all-atom representation this can generally mean that the minor groove width is affected by the underlying force field. This is in fact seen in the all-atom simulations. To substantiate this fact minor groove width calculation was done for the all-atom trajectories. The same distribution curve is seen also for the all-atom representation. Figure 4.5 clearly shows that parmbsc1 exhibits the shortest minor groove width

FIGURE 4.4: Comparison of DNA-ion and ion-ion potentials obtained from RDFs (4.3) of four different underlying all-atom force fields.

among all the force fields. Parmbsc0 gives a bit higher average as seen in the red curve and CHARMM27 and CHARMM36 have similar distribution curves for the minor groove widths. This in turn also validates the CG-model in its preservation of the groove structure of DNA.

The CG potentials are constructed to reproduce selected average structural properties of the fine-grained system such as RDF between non-bonded CG sites and bond length and angle distributions for bonded sites. For the CG DNA model, these are obtained by mapping the all-atom MD trajectory from the four-DNA system to a corresponding CG site representation. Thus, no empirical or adjustable parameters are used in the CG DNA model and it rests only on the parameters from all-atom DNA force fields as well as on the ion parameters for NaCl/KCl. The IMC-derived potentials are tabulated and not prescribed by any specific functional form.

FIGURE 4.5: Minor groove width variation averaged over the full $2\mu\text{s}$ trajectories (A), and (B) shows the distribution of the minor groove width for all all force fields averaged for all the base-steps over the full trajectory.

4.1.1 Reproduction of the reference RDFs

To verify that the CG potentials obtained are well representative of the underlying all-atom system, a coarse-grained simulation was performed with a similar length of DNA and the same salt concentration. The resulting RDFs at CG level are then compared with the corresponding RDFs at the all-atom level and also the IMC sampled RDFs. A good agreement between the RDFs gives a validation of the IMC procedure in its ability to preserve the selected structural properties of the system. Such a comparison is shown in Figure 4.6.

FIGURE 4.6: A comparison of the different RDFs obtained from the underlying all-atom simulation (blue), reproduced RDF at coarse-grained level is shown in green and the IMC sampled RDF for the final iteration is also shown in red. A good agreement is seen in all the RDFs characterizing the 1DNA system, thus validating the IMC procedure for this system.

The MagiC software supports two engines. One can either use GROMACS as in the all-atom model or use another software called LAMMPS [138] as it is more flexible in terms of its choice of parameters. The present system was tested with both the softwares and better agreement of the RDFs was seen with LAMMPS as compared to GROMACS suggesting that GROMACS electrostatics is more suited for all-atom simulation. However this can also be due to the fact that the way long range electrostatics is treated in MagiC and GROMACS is different.

The potentials obtained from the IMC procedure are exported to the native file format of the corresponding software to test the model. These potentials are then used in a CGMD simulation whereby each potential describes the interactions between each of the CG-sites.

In the present case a CG-simulation was run to test the model. As seen in Figure 4.6, selected RDFs have been shown. Similar agreement was found for other RDFs. There were various parameters that were tested to be used in the the CG-simulation. These include: the time-step, the coupling constant τ_T , the electrostatic cutoff and the fourier grid spacing.

4.1.2 Coarse-grained MD simulations

A good reproduction of the RDFs at the coarse-grained level ensures that the potentials are a good representative of the underlying all-atom system and can be used in CGMD simulations at a much larger scale and longer timescales with lesser computational power as compared to the all-atom simulations.

One such application in which these RDFs can be used is the calculation of the persistence length of DNA and its dependence on salt concentration. This results also lends validity to the coarse-grained model as the results can be directly compared with experiments. Since we use explicit ions in the model, the derived potentials can be used for multiple salt concentrations using the same DNA-ion potentials derived from the all-atom simulation at just one salt concentration. To this end, CGMD simulations were done using a relatively longer DNA of 500 bp (40 bp at all-atom level) with multiple salt concentrations and even larger box sizes (see Table 4.1). The results from these simulations are then used to calculate the persistence length of DNA at various salt concentrations. Fitting Equation 1.1 with a straight line gives the persistence length.

One such example of the fitting is given in Figure 4.7 for the 100 mM NaCl system. To verify that the persistence length is similar over the full length of the DNA, calculation was done over different pieces of DNA molecule, each piece being 100bp each. A table showing the L_p for each 100 bp fragment and also the full 500 bp DNA is also shown in Table 4.2. The errors are obtained from the linear fitting Equation 1.1.

TABLE 4.1: Coarse-grained MD simulations, DNA length = 500 bp

Salt conc. (mM)	Box size (nm)	CG beads	Trajectory Length (μ s)
0.1	180	2946	2
0.5	180	5756	2
1	180	9270	2
5	180	37370	3
10	180	72495	2.5
50	125	119880	2.5
100	125	237518	2.5

In addition, to check that the results have converged over the full length of the trajectory, the calculation was repeated over different fragments of the CGMD trajectory. Similar results (within the standard deviation) were obtained for all the fragments of the trajectory, thus ensuring the convergence of the CGMD simulations. The corresponding L_P values and the plot is shown in the Appendix (Figure A.7).

DNA fragment	L_P (Å)	Error (Å)
1- 100 bp	353.43	2.31
101 - 200 bp	408.42	0.85
201 - 300 bp	428.35	3.05
301 - 400 bp	346.11	1.69
401 - 500 bp	391.59	0.61
1 - 500 bp	377.54	0.91

TABLE 4.2: L_P for 100 bp fragments FIGURE 4.7: Contour length L_C versus $\ln(\langle \cos \alpha \rangle)$ for 100 bp fragments of DNA

4.1.2.1 Improvement of the CG model

As mentioned in the methods section, there was a first generation of the coarse-grained model developed in [86], similar to the one used here but the present model is significantly more advanced as all the interactions were derived using IMC, unlike in the previous work, where only the internal DNA parameters were extracted using IMC and the bonded potentials were fitted to a known functional form (harmonic potentials). Also ionic interactions were treated using a Coulombic dielectric

medium model in the previous model. Another difference in the underlying atomistic simulation is that they were too short and included multiple DNA molecules that exhibited condensation, which may not be ideal to study persistence length behaviour.

For this old version of the CG model, when the calculation was done for the persistence length, it was observed that the DNA is too soft and thus underestimated the persistence length as compared to experiments. One such comparison is made in Figure 4.8. It can be clearly seen that the L_p values obtained using the model without the P-P minor groove bonds are quite low as compared to the experiments.

FIGURE 4.8: (A) Schematic showing the first generation CG model. Adapted from [44] (Figure 3B). (B) Dependence of DNA persistence length on the salt concentration, I (mM) of KCl obtained from CGMD simulations based on the first generation CG model. The effective potentials were obtained from corresponding all-atom simulation of a single 40 bp DNA molecule using 130 mM KCl.

In line with this result and other results obtained using the 4DNA system by other members of the group, the CG model was updated by adding the P-P bond along the minor groove. This will lend more rigidity to the DNA and mimic correct behavior at the nucleosome level. A schematic depicting the improved CG model with the P-P bond is shown in Figure 4.9A.

To do this, the same all-atom trajectories were used but now with a different bead-mapping scheme including the additional P-P bond. The RDFs were calculated again and finally the IMC procedure was run again for the system with the new CG system. Using the new set of converged potentials, the coarse-grained simulations were run again to derive the persistence length for the new model. When the results

are compared with experimental results they are closer than the model without P-P bonds. Since the range of the experimental values is too wide for L_p , the range was estimated using a smoothing function (*stat_smooth*) from the R package. The confidence interval used is 0.99995. For *parmbsec0*, not enough data was available for the older model without P-P bonds, so it is excluded from this comparison.

FIGURE 4.9: (A) A schematic showing the improved CG model with the added P-P bond along the minor groove of DNA. For the corresponding RDF and the potentials refer to Figure 4.3. Reproduced with permission from [162] (Figure 1D) (B) Dependence of DNA persistence length on the salt concentration, I (mM) of KCl obtained from CGMD simulations based on the first generation model (black solid circles) and the improved generation CG model (black hollow circles). The effective potentials were obtained from corresponding all-atom simulation of a single 40 bp DNA molecule using 130 mM KCl. The grey band corresponds to the experimental range of values.

It must be noted that the model has not been improved based only on the persistence length results but also because the DNA was too flexible in simulations of multiple DNA molecules (other members of the group). The addition of the P-P bond effectively excluded more negatively charged P beads undergoing electrostatic repulsion and thus lent additional rigidity to the DNA molecule.

4.1.3 Comparison of force fields

To advance further the comparison of force fields beyond the all-atom level, calculation of effective CG potentials was done based on the all-atom trajectories for each of the four underlying force fields. These are then used in CGMD simulations

to calculate the persistence length of DNA and thus assess the performance of these force fields in predicting mesoscale-level properties of DNA.

For a summary of the CGMD simulations done for all the four force fields please refer to Table 4.1. An exception is the low salt cases for CHARMM36 (0.1 mM and 0.5 mM), for which the simulations were not run.

The results obtained are tabulated in Table 4.3 and depicted in Figure 4.10. The experimental values are shown as solid coloured squares and the corresponding range is shown as the grey band. Please note that most of the experimental studies are based on NaCl whereas our simulations were done with KCl. Hence to compare, an additional set of potentials for CHARMM27 but with NaCl were calculated and CGMD simulations were also run for this set of potentials. As can be seen in Figure 4.11 the results for NaCl and KCl are similar within the standard deviation.

Comparing the results for different atomistic force fields one can see that they are close to each other, especially at low salt. This can be understood by the fact that at low salt concentrations the higher persistence length is due to the long range electrostatic interactions which are essentially the same in all the models. The atomistic force field mainly affects the internal rigidity of DNA, and one can see similar trends in the elastic properties and stability of atomistic models and L_P , where CHARMM27 demonstrated higher rigidity and also stability (see Figure 3.1 and 3.12). This translates to higher persistence length as compared to other force fields. Consequently, CHARMM27 again gives the most satisfactory results for L_P as far as comparison with experiments is concerned.

Experimental results show a rather wide spread, but generally lie somewhat higher than the computed curves. Comparing data on persistence length obtained from different sources one should keep in mind that persistence length is strictly defined only for a worm-like polymer model. However, for real DNA excluded-volume effects are not negligible and make larger contribution at low ionic strengths due to the increased importance of long-range interactions. Most of the experimental data are based on methods determining the total size of the DNA coil, which is then translated to persistence length within some models. Omission or too approximate a description of the electrostatic interactions in such models may affect the results, which is likely the main reason behind the observed spread of experimental data. Still, our results are consistent with the major trend of the experimental results:

TABLE 4.3: DNA Lp results based on four underlying all-atom force fields

Salt Conc. [mM]	Force field			
	bsc0	bsc1	C27	C36
0.1	994 ± 121	1019 ± 89	1066 ± 102	-
0.5	782 ± 84	765 ± 77	777 ± 56	-
1	549 ± 50	594 ± 37	751 ± 83	578 ± 87
5	611 ± 189	648 ± 189	670 ± 219	573 ± 274
10	573 ± 56	373 ± 75	533 ± 115	530 ± 74
50	294 ± 24	467 ± 58	387 ± 83	329 ± 68
100	342 ± 33	347 ± 42	493 ± 69	352 ± 77

FIGURE 4.10: Dependence of DNA persistence length on monovalent salt concentration. Experimental studies are depicted in solid coloured bands and their smooth fitting shown in the grey band, while the results obtained using the updated CG model with IMC are shown in solid coloured circles.

the persistence length computed in the CG model based on the CHARMM27 force field is between 400 and 500 Å for salt concentrations above 10 mM, and rapidly increases at lower salt concentrations.

4.1.4 Comparison with other CG models

Figure 4.11 compares the CHARMM27 force field results of the present work to persistence length data (with KCl) obtained from other bottom-up CG DNA models

with explicit ions. One can see good agreement with results of our previous work, [86] where the intramolecular CG DNA potentials were derived from CHARMM27 force field simulations and with short-range intramolecular interactions described by Lennard-Jones potentials with empirical parameters. Our results are also close to the results obtained by Naome et. al. [124] based on the parmbsc0 AMBER force field and to those of Hinckley and de Pablo. [70]

FIGURE 4.11: Dependence of DNA persistence length on monovalent salt concentration. Experimental studies are only depicted using a smooth fitting in the grey band, while the results obtained using the updated CG model with IMC are shown in black and blue solid circles. Results from other CG models are shown as coloured circles with white fill.

In the latter work, bottom-up derived CG potentials based on the parm94 version of the AMBER force field were used only for the intermolecular interactions, while internal DNA interactions were obtained from the empirically parameterized 3SPN model. The persistence length calculations by Savelyev and Papoian [152] resulted in considerably higher L_p values. However, uniformly rescaling all CG DNA structural parameters by a factor of 0.7, leaving the rest of the CG force field unchanged resulted in a considerable improvement in the agreement with experimental data. This rescaling was done in order to soften the underlying AMBER parmbsc0 force field.

Among the bottom-up scheme of CG modeling there have been very few models that predicted the persistence length dependence on the ionic concentration accurately without using any adjustable parameters. The current CG model includes explicit ions and there are no adjustable parameters to model specific DNA properties. Given this fact, accurate persistence length behavior is a natural prediction of the model and agrees well with experiments.

4.2 Conclusions

In addition to the comparison of force fields at the atomistic level, using an improved coarse-grained model we calculated effective CG potentials for all the four force fields using the IMC procedure. Based on these potentials, we performed large-scale MD simulations to further assess the ability of these force fields to predict the persistence length of DNA for varying salt concentrations. All the force fields predicted correct behavior of this important physiological property of DNA, as compared to experimental results in the physiological range. This validates two points. Firstly, it proves the ability of computational techniques to mimic physiologically relevant behavior of complex systems like the DNA. Secondly, it validates the IMC procedure and the coarse-grained model used to study large-scale biologically relevant systems.

These results validate the CG DNA model and the bottom-up approach for studying DNA condensation. Based on this model, Sun et. al. [162] have used a hierarchical approach for systematic multiscale CGMD simulations, predicting DNA phase separation in the presence of three-valent cobalt(III)-hexammine (CoHex^{3+}). Solvent mediated effective potentials for a CG model of DNA were extracted from all-atom simulations. Using these potentials CGMD simulations of several 100 bp long DNA molecules demonstrated aggregation to a liquid crystalline hexagonally ordered phase. Further coarse-graining of this CG model enabled mesoscale modelling, and the resulting potentials were used in a super-coarse-grained MD simulation of a 10.2-kbp DNA molecule. This simulation showed phase separation to a toroid or a fibre with hexagonal packing, in agreement with electron microscopy observations. A picture illustrating this phase separation is shown in Figure 4.12.

FIGURE 4.12: DNA phase separation into toroids using a super CG DNA model, based on the coarse-grained model presented above. Reproduced with permission from [162]

Chapter 5

Ongoing and future work - large-scale NCP simulations

To model the compaction of DNA at the nucleosome and chromatin level, one needs to derive the potentials between DNA, the histone tails and the ions. One also needs the interaction potentials including the HO core, HO core and the ions. However, due to the complexity of the system, it seems not possible to extract effective potentials for all interactions in the NCP in a single large all-atom simulation. The approach used here is to construct a composite model, where the effective potentials between components of the NCP are independently derived in somewhat simpler environment and then gathered together in the resulting model. This approach raises the question of transferability of potentials which is addressed by comparing potentials obtained from different subsystems.

The amino acids were modeled based on the 48 tail fragments as discussed in Section 2.2.2. Given the assumption that these 48 fragments form a representative sample containing all the tail amino-acids present in the tails, one can derive effective potentials for the tails using this system. For the CG model each amino acid is represented by one bead each and in total there are 5 different kinds of beads, non-polar, polar, negative, positive and glycine. The bead-mapping for the amino-acids is illustrated in [Figure 5.1](#).

FIGURE 5.1: The CG model used for the amino acids illustrated for one of the 48 tail fragments used in the system

5.1 DNA - histone tail potentials

For the DNA-tail fragments simulation, DNA-amino acid (aa) and aa-ions potentials were derived using the underlying RDFs. The convergence of the system was assessed using block averaging of the RDFs over a 500 ns time window (See Figure A.8 - A.11). Also for the DNA-tails system, ten different all-atom simulations were run to ensure proper sampling of the system. In total this system involved 78 non-bonded RDFs, 336 bonded RDFs and 288 RDFs for the angular pairwise interactions. To decrease the complexity of such a large number of interactions it was decided to use just one single bond type for all the amino acids. This was also validated by the fact that all the bonded interactions had similar distributions for each amino acid bead type.

For the angle distributions, it was decided to define three groups of angles (see methods chapter), one including glycine in the center, another group with proline in the center and the third one for any other amino acid type in the center (See Appendix A.12 for the variation). IMC was then done for this system using this reduced set of RDFs including, 78 non-bonded potentials, 5 bonded and 6 angular pairwise interactions. Also it must be noted that since multiple DNA molecules were present, the inter-DNA interactions (D-D D-P and P-P) were not zeroed in this case (different from the single DNA case). The RDFs taken were averaged

FIGURE 5.2: Non-bonded potentials obtained between DNA-DNA beads (top row) and DNA amino acid bead types (bottom two rows)

over the 10 underlying all-atom trajectories. The resultant RDFs and potentials are shown in Figure 5.2.

To get the potentials between multivalent ions and the tails, additional simulations were run including the 48 tail fragments and the ions, namely K^+ , Na^+ , Mg^{2+} and $Co(NH_3)_6^{3+}$ (see methods). As in the DNA-tail fragment simulation, potentials were also derived between the aa beads and the ions. Also the inter-aa interactions are shown in Figure 5.4. Also the RDFs are plotted in the same plot for brevity. The potentials between amino acid beads and ions are shown in Figure 5.5.

5.2 Large-scale CGMD simulations - preliminary results

As mentioned earlier several sub-systems were used to derive the CG potentials for the large-scale CGMD simulations involving multiple NCP molecules. The underlying all-atom MD subsystems used are mentioned below and the potentials extracted from each one is also mentioned. Since each pair-potential can be extracted from multiple subsystems, a comparison of the monovalent ion potentials

FIGURE 5.3: Illustrative figure showing from which underlying MD sub-system each pair-potential is extracted from the CGMD simulation. The central yellow box is the large-scale CGMD system, (with Mg and CoHex) and the five MD systems are shown using the rectangular boxes. The arrows connect each pair-potential in the CGMD setup to the corresponding underlying MD system from which it is extracted.

is also shown in the Appendix (See Figure A.13 and A.14). Figure 5.3 illustrates from which underlying sub-system each pair potential is obtained.

A summary of how each pair potential is obtained is given below:

- 4DNA + Mg^{2+} simulation
 1. Mg^{2+} - Mg^{2+}
 2. Mg^{2+} -monovalent ions
 3. Monovalent ions- monovalent ions
 4. Mg^{2+} -DNA
- 4DNA + CoHex^{3+} simulation
 1. CO^{3+} - CO^{3+}
 2. CO^{3+} -monovalent ions
 3. CO^{3+} -DNA

- Tail fragments + Mg^{2+} simulation
 1. Mg^{2+} -histone tails
 2. Mg^{2+} -histone core
- Tail fragments + CoHex^{3+} simulation
 1. CO^{3+} -histone tails
 2. CO^{3+} -histone core
- 8DNA + tail fragments simulation
 1. DNA-histone tails
 2. DNA-histone core
 3. DNA-DNA
 4. DNA-monovalent ions
 5. Histone tail-monovalent ions
 6. Histone core-monovalent ions
 7. Histone tail-histone tail
 8. Histone tail-histone core
- Single NCP simulation: A system containing one single NCP simulation of 500 ns long was performed. The bond-length distributions for all core-core and core-DNA bonds as were derived from this simulation. The potentials for these bonds are derived by running iterative Boltzmann inversion (IBI).
 For the short-range non-bonded interactions between the core-core part and the DNA-core part, they are represented with the repulsive part of Lenard-Jones (excluded volume) and all other potentials are used from different systems simulated in the group.

These potentials obtained above (partly done by former member of the group, Alexander Mirzoev) have been used in a simulation with multiple NCPs. These simulations mostly involve 20-28 NCPs in presence of multivalent salt (CoHex^{3+}). A typical system setup is shown in Figure 5.6.

Preliminary analysis shows a hexagonal arrangement of the NCPs in a single plane. There are also some stacked pairs observed but most of the interactions involve side

FIGURE 5.4: Non-bonded potentials obtained between aa beads are shown in the first four rows. The five categories of aa beads are shown on the top of each plot, POS - Positive, NPL - Non-Polar, GLY - Glycine, POL - Polar and NEG - Negative. For eg. POS-NEG means the potentials between positive and negative aa beads. The bottom row shows the bonded potentials defined for the aa beads. Angle 1 refers to all angles with Glycine in the center, angle 2 with proline in the center and angle three for any other amino acid in the center.

to side DNA contact. Since the system size is big and there are many interactions involved, much longer simulations are required to obtain more orderly structures. Another simulation with monovalent ions was also done to see if aggregation is caused by CoHex ions. No aggregation was observed in the simulation with monovalent ions (data not shown).

FIGURE 5.5: Non-bonded potentials obtained between aa beads and ions.

FIGURE 5.6: *Left*: CGMD simulation setup showing 20 NCPs in a 55 nm box. *Center*: Two insets of a single NCP structure in two orthogonal perspectives. *Right*: Snapshots showing the aggregated cluster of NCPs after 100 millions steps (≈ 500 ns) in two perspectives. The tails have not been shown to see the aggregation of NCPs clearly.

Chapter 6

Conclusions

In this work, we used molecular dynamics simulations to assess the ability of DNA atomistic force fields at a multiscale level. Using multi-microsecond atomistic simulations we compared the force fields in their prediction of DNA stability and geometry at the base-pair level. It was found that the most recent version of the CHARMM family of force fields, CHARMM36 mistreats DNA structure at the base-pair level and severely distorts the DNA backbone and is not suitable to carry out atomistic simulations or study DNA dynamics. The older version, CHARMM27 behaves much better when it comes to the base-pair geometry of DNA. Among the AMBER family of force fields, parmbsc0 and parmbsc1 are mostly stable at the all-atom level, but parmbsc1 is better suited for DNA simulations as it conserves DNA geometry and backbone. Also the dihedral angles that characterise the DNA backbone and which play an important role in DNA-protein interactions are much better reproduced with parmbsc1, in agreement with experiments. Among all the force fields analysed, we uncovered that CHARMM27 is the most stable force field at the base-pair level as it treats the DNA backbone structure and overall elasticity of DNA fairly well. The only problem it has is its underestimation of the BII conformations as compared to experimental crystal structures.

The underlying all-atom MD simulations were then used in a coarse-grained model to predict the persistence length of DNA and then compare with other CG models. The model predicts the persistence length dependence of DNA on monovalent salt concentration fairly well in excellent agreement with experiments.

To build the multiscale model of NCP, effective CG potentials were then derived for the DNA - amino acids, amino acid - ions and amino acid- amino acid interactions. Using different underlying all-atom simulations, these potentials were derived from diverse set of systems. Combining these potentials large scale CGMD simulations involving multiple NCPs to study the behavior of nucleosomes are being run currently and forms the future part of the work.

Appendices

Appendix A

Appendix

TABLE A.1: 48 fragments used in tail simulations

Sequence	Charge	Sequence	Charge
SGRGKQGG	2	KTQKKDGG	3
GGKTRAKA	3	KTQEKDGG	1
KAKTRSSR	4	VLRDSGRG	1
PKKTESSK	2	TKTQKKDG	2
SSKSAKSK	3	HRKVLRDQ	2
AKSAPAPK	2	SAPATGRV	1
PKKGSKKA	4	SRHGKQGG	2
KKAVTKTQ	3	KAVTKTQK	3
KKDGGKKRR	5	GKGLGKGG	2
ARTKQTAR	3	AKRHRKVL	4
RKSTGGKA	3	GAKRHRKV	4
PRKQLATK	3	GAKRHRDV	2
KAARKSAP	3	VKKPATGG	2
ATGGVKKP	2	RKSAPKAA	3
SGRGKGGK	3	QLATKPRK	3
KGGKGLGK	3	ESSKPKKT	2
GGAKRHRK	4	SKKAPKKG	4
RHRKVLRD	3	RHEKVLRD	1
SAPAPKKG	2	PRQQLATH	1
APKKGSKK	4	RHRKGHAK	4
GGKAPRKQ	3	SKDAPKHG	1
KSAPATGR	2	VLRDRHEK	1
TESSKSAK	1	KGGEGLHK	1
TEVVKSAL	0	KKDGGKERR	3

FIGURE A.1: All-atom MD structures averaged over the full trajectory for all the 4 force fields.

FIGURE A.2: The percentage of BI and BII conformations shown for each of the individual bases in the DNA sequence for strand 1 (A) and strand 2 (B) for the four force fields. The grey line and band show the experimental values for the corresponding base-steps.

FIGURE A.3: Variation of three base-pair step parameters shown for the CoHex 4DNA system for 3 different trajectories using the C36 FF. As can be seen the distortion of DNA structure is independent of the sequence as even DNA3 and DNA4 show unphysical values for these parameters. For the DNA sequences used see Section 3.1

FIGURE A.4: Distribution of the 6 base-pair parameters for the 1D DNA CHARMM36 trajectories

FIGURE A.5: Distribution of the 6 base-step parameters for the 1DNA CHARMM36 trajectories

FIGURE A.6: Distribution of the 6 base-step helical parameters for the 1DNA CHARMM36 trajectories

CGMD fragment	L_P (Å)	Error (Å)
Fragment 1	537.15	7.43
Fragment 2	490.61	4.28
Fragment 3	324.28	2.41
Fragment 4	242.05	4.67
Full trajectory	377.54	0.91

TABLE A.2: L_P for different fragments of CGMD trajectory

FIGURE A.7: Contour length L_C versus $\ln(\langle \cos \alpha \rangle)$ for 4 fragments of CGMD trajectory

FIGURE A.8: Block averaging of bonded RDFs obtained over 500 ns of trajectory for the DNA tails system

FIGURE A.9: Block averaging of DNA-DNA and DNA-aa non-bonded RDFs obtained over 500 ns of trajectory for the DNA tails system

FIGURE A.10: Block averaging of aa-aa non-bonded RDFs obtained over 500 ns of trajectory for the DNA tails system

FIGURE A.11: Block averaging of aa - ions non-bonded RDFs obtained over 500 ns of trajectory for the DNA tails system

FIGURE A.12: The classification of angle bonds of all other amino acids into three groups and the averaged plot is shown in the lower right corner.

FIGURE A.13: Comparison of monovalent ionic potentials obtained from 5 different underlying all-atom MMD systems.

FIGURE A.14: Comparison of DNA bonded obtained from 2 underlying all-atom MD systems.

Bibliography

- [1] M. T. S. R. P. S. S. J. C. H. B. L. E. Abraham, M. J. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015. [30](#)
- [2] C. t. Altona and M. Sundaralingam. Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation. *Journal of the American Chemical Society*, 94(23):8205–8212, 1972. [53](#)
- [3] A. Annunziato. DNA packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26, 2008. [1](#)
- [4] G. Arents, R. W. Burlingame, B.-C. Wang, W. E. Love, and E. N. Moudrianakis. The nucleosomal core histone octamer at 3.1 a resolution: a tripartite protein assembly and a left-handed superhelix. *Proceedings of the National Academy of Sciences*, 88(22):10148–10152, 1991. [5](#), [10](#)
- [5] G. Arents and E. N. Moudrianakis. The histone fold: a ubiquitous architectural motif utilized in dna compaction and protein dimerization. *Proceedings of the National Academy of Sciences*, 92(24):11170–11174, 1995. [5](#)
- [6] G. Arya and T. Schlick. A tale of tails: how histone tails mediate chromatin compaction in different salt and linker histone environments. *The Journal of Physical Chemistry A*, 113(16):4045–4059, 2009. [21](#)
- [7] J. Ausio, F. Dong, and K. Van Holde. Use of selectively trypsinized nucleosome core particles to analyze the role of the histone “tails” in the stabilization of the nucleosome. *Journal of Molecular Biology*, 206(3):451–463, 1989. [10](#)

- [8] G. S. Ayton, W. G. Noid, and G. A. Voth. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current Opinion in Structural Biology*, 17(2):192–198, 2007. [17](#)
- [9] C. G. Baumann, S. B. Smith, V. A. Bloomfield, and C. Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proceedings of the National Academy of Sciences*, 94(12):6185–6190, 1997. [12](#)
- [10] D. A. Beard and T. Schlick. Computational modeling predicts the structure and dynamics of chromatin fiber. *Structure*, 9(2):105–114, 2001. [20](#), [21](#)
- [11] A. S. Belmont and K. Bruce. Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *The Journal of Cell Biology*, 127(2):287–302, 1994. [10](#)
- [12] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. Srinivasan, and B. Schneider. The nucleic acid database. a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical journal*, 63(3):751, 1992. [45](#)
- [13] A. Bertin, A. Leforestier, D. Durand, and F. Livolant. Role of histone tails in the conformation and interactions of nucleosome core particles. *Biochemistry*, 43(16):4773–4780, 2004. [8](#)
- [14] A. Bertin, M. Renouard, J. S. Pedersen, F. Livolant, and D. Durand. H3 and H4 histone tails play a central role in the interactions of recombinant NCPs. *Biophysical journal*, 92(7):2633–2645, 2007. [8](#)
- [15] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham III, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, et al. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. research design and results on d(CpG) steps. *Biophysical journal*, 87(6):3799–3813, 2004. [15](#)
- [16] M. Biswas, K. Voltz, J. C. Smith, and J. Langowski. Role of histone tails in structural stability of the nucleosome. *PLoS computational biology*, 7(12):e1002279, 2011. [13](#)
- [17] V. A. Bloomfield. DNA condensation by multivalent cations. *Biopolymers: Original Research on Biomolecules*, 44(3):269–282, 1997. [7](#), [8](#)

- [18] J. Bordas, L. Perez-Grau, M. Koch, M. Vega, and C. Nave. The superstructure of chromatin and its condensation mechanism. II. theoretical analysis of the X-ray scattering patterns and model calculations. *European biophysics journal: EBJ*, 13(3):175–185, 1986. [5](#), [10](#)
- [19] Z. Bryant, M. D. Stone, J. Gore, S. B. Smith, N. R. Cozzarelli, and C. Bustamante. Structural transitions and elasticity from torque measurements on DNA. *Nature*, 424(6946):338–341, 2003. [61](#)
- [20] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007. [30](#)
- [21] C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith. Single-molecule studies of DNA mechanics. *Current Opinion in Structural Biology*, 10(3):279–285, 2000. [61](#)
- [22] K. L. Cairney and R. E. Harrington. Flow birefringence of t7 phage dna: dependence on salt concentration. *Biopolymers: Original Research on Biomolecules*, 21(5):923–934, 1982. [12](#)
- [23] C. P. Calderon, W. H. Chen, K. J. Lin, N. C. Harris, and C. H. Kiang. Quantifying DNA melting transitions using single-molecule force spectroscopy. *Journal of Physics: Condensed Matter*, 21(3):34114, 2009. [61](#)
- [24] T. E. Cheatham. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Current Opinion in Structural Biology*, 14(3):360–367, 2004. [13](#), [14](#), [15](#)
- [25] T. E. Cheatham III and D. A. Case. Twenty-five years of nucleic acid simulations. *Biopolymers*, 99(12):969–977, 2013. [13](#)
- [26] T. E. Cheatham III, P. Cieplak, and P. A. Kollman. A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 16(4):845–862, 1999. [14](#)
- [27] T. E. Cheatham III and P. A. Kollman. Molecular dynamics simulation of nucleic acids. *Annual review of physical chemistry*, 51(1):435–471, 2000. [13](#)
- [28] D. J. Clark and T. Kimura. Electrostatic mechanism of chromatin folding. *Journal of Molecular Biology*, 211(4):883–896, 1990. [9](#)

- [29] R. Collepardo-Guevara, G. Portella, M. Vendruscolo, D. Frenkel, T. Schlick, and M. Orozco. Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: a multiscale computational study. *Journal of the American Chemical Society*, 137(32):10205–10215, 2015. [13](#)
- [30] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995. [14](#)
- [31] Y. Cui and C. Bustamante. Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proceedings of the National Academy of Sciences*, 97(1):127–132, 2000. [11](#)
- [32] R. B. D. C. T. C. I. T. D. R. D. T. G. H. G. A. G. N. H. S. I. P. J. J. K. A. K. T. L. S. L. P. L. T. L. R. L. B. M. K. M. G. M. P. N. H. N. H. N. I. O. A. O. D. R. A. R. R. S.-F. C. S. W. S. J. S. R. W. J. W. R. W. X. W. D. Y. D.A. Case, J.T. Berryman and P. Kollman. Ambertools 15. *Amber 15*, 2015. [33](#)
- [33] P. D. Dans, I. Ivani, A. Hospital, G. Portella, C. Gonzalez, and M. Orozco. How accurate are accurate force-fields for B-DNA? *Nucleic Acids Research*, 45(7):4217–4230, 2017. [14](#), [15](#), [51](#)
- [34] P. D. Dans, J. Walther, H. Gomez, and M. Orozco. Multiscale simulation of DNA. *Current Opinion in Structural Biology*, 37:29–45, 2016. [14](#), [20](#)
- [35] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993. [30](#)
- [36] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of Molecular Biology*, 319(5):1097–1113, 2002. [11](#)

- [37] M. De Frutos, E. Raspaud, A. Leforestier, and F. Livolant. Aggregation of nucleosomes by divalent cations. *Biophysical journal*, 81(2):1127–1132, 2001. [8](#)
- [38] J. J. de Pablo. Coarse-grained simulations of macromolecules: from DNA to nanocomposites. *Annual Review of Physical Chemistry*, 62:555–74, 2011. [18](#)
- [39] R. C. DeMille, T. E. Cheatham III, and V. Molinero. A coarse-grained model of DNA with explicit solvation by water and ions. *The Journal of Physical Chemistry B*, 115(1):132–142, 2010. [17](#)
- [40] R. E. Dickerson. Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Research*, 17(5):1797–1803, 1989. [2](#), [13](#)
- [41] 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. [4](#), [15](#)
- [42] 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. *Proceedings of the National Academy of Sciences*, 78(4):2179–2183, 1981. [15](#)
- [43] D. G. Edmondson, M. M. Smith, and S. Y. Roth. Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes & development*, 10(10):1247–1259, 1996. [10](#)
- [44] Y. Fan, N. Korolev, A. P. Lyubartsev, and L. Nordenskiöld. An advanced coarse-grained nucleosome core particle model for computer simulations of nucleosome-nucleosome interactions under varying ionic conditions. *PLoS One*, 8(2):e54228, 2013. [xviii](#), [76](#)
- [45] G. Felsenfeld and M. Groudine. Controlling the double helix. *Nature*, 421(6921):448, 2003. [5](#)
- [46] J. Finch and A. Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences*, 73(6):1897–1901, 1976. [5](#)

- [47] T. M. Fletcher and J. C. Hansen. Core histone tail domains mediate oligonucleosome folding and nucleosomal DNA organization through distinct molecular mechanisms. *Journal of Biological Chemistry*, 270(43):25359–25362, 1995. [10](#)
- [48] N. Foloppe and A. D. MacKerell. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000. [14](#), [29](#)
- [49] E. Fussner, R. W. Ching, and D. P. Bazett-Jones. Living without 30 nm chromatin fibers. *Trends in biochemical sciences*, 36(1):1–6, 2011. [5](#)
- [50] R. Galindo-Murillo, J. C. Robertson, M. Zgarbova, J. Sponer, M. Otyepka, P. Jurec̃ka, and T. E. Cheatham III. Assessing the current state of amber force field modifications for dna. *Journal of Chemical Theory and Computation*, 12(8):4114–4127, 2016. [14](#), [60](#)
- [51] R. Galindo-Murillo, D. R. Roe, and T. E. Cheatham III. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):1041–1058, 2015. [62](#)
- [52] M. Garcia-Ramirez, C. Rocchini, and J. Ausio. Modulation of chromatin folding by histone acetylation. *Journal of Biological Chemistry*, 270(30):17923–17928, 1995. [10](#)
- [53] J. Gay and B. Berne. Modification of the overlap potential to mimic a linear site–site potential. *The Journal of Chemical Physics*, 74(6):3316–3319, 1981. [21](#), [22](#)
- [54] F. Gordon, K. Luger, and J. C. Hansen. The core histone n-terminal tail domains function independently and additively during salt-dependent oligomerization of nucleosomal arrays. *Journal of Biological Chemistry*, 280(40):33701–33706, 2005. [10](#)
- [55] J. Gore, Z. Bryant, M. Nollmann, M. U. Le, N. R. Cozzarelli, and C. Bustamante. DNA overwinds when stretched. *Nature*, 442(7104):836–839, 2006. [61](#)

- [56] S. A. Grigoryev and C. L. Woodcock. Chromatin organization - the 30 nm fiber. *Experimental cell research*, 318(12):1448–1455, 2012. [6](#)
- [57] N. Grønbech-Jensen, R. J. Mashl, R. F. Bruinsma, and W. M. Gelbart. Counterion-induced attraction between rigid polyelectrolytes. *Physical Review Letters*, 78(12):2477, 1997. [8](#)
- [58] L. Guldbrand, L. G. Nilsson, and L. Nordenskiöld. A Monte Carlo simulation study of electrostatic forces between hexagonally packed DNA double helices. *The Journal of Chemical Physics*, 85(11):6686–6698, 1986. [8](#)
- [59] P. J. Hagerman. Investigation of the flexibility of DNA using transient electric birefringence. *Biopolymers: Original Research on Biomolecules*, 20(7):1503–1535, 1981. [12](#)
- [60] P. J. Hagerman. Flexibility of DNA. *Annual review of biophysics and biophysical chemistry*, 17(1):265–286, 1988. [12](#)
- [61] P. J. Hagerman and B. H. Zimm. Monte Carlo approach to the analysis of the rotational diffusion of wormlike chains. *Biopolymers: Original Research on Biomolecules*, 20(7):1481–1502, 1981. [12](#)
- [62] J. C. Hansen. Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. *Annual review of biophysics and biomolecular structure*, 31(1):361–392, 2002. [10](#)
- [63] S. W. Harshman, N. L. Young, M. R. Parthun, and M. A. Freitas. H1 histones: current perspectives and challenges. *Nucleic Acids Research*, 41(21):9593–9609, 2013. [10](#)
- [64] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, and A. D. MacKerell Jr. Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *Journal of Chemical Theory and Computation*, 8(1):348–362, 2011. [13](#), [14](#), [29](#), [58](#)
- [65] G. A. Hartzog and F. Winston. Nucleosomes and transcription: recent lessons from genetics. *Current opinion in genetics & development*, 7(2):192–198, 1997. [5](#)

- [66] S. C. Harvey. The scrunchworm hypothesis: transitions between a-dna and b-dna provide the driving force for genome packaging in double-stranded dna bacteriophages. *Journal of structural biology*, 189(1):1–8, 2015. [3](#)
- [67] A. Hecht, T. Laroche, S. Strahl-Bolsinger, S. M. Gasser, and M. Grunstein. Histone H3 and H4 N-termini interact with SIR3 and SIR4 proteins: a molecular model for the formation of heterochromatin in yeast. *Cell*, 80(4):583–592, 1995. [10](#)
- [68] B. Heddi, N. Foloppe, N. Bouchemal, E. Hantz, and B. Hartmann. Quantification of DNA BI/BII backbone states in solution. implications for DNA overall structure and recognition. *Journal of the American Chemical Society*, 128(28):9170–9177, 2006. [58](#)
- [69] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008. [30](#)
- [70] D. M. Hinckley and J. J. de Pablo. Coarse-grained ions for nucleic acid modeling. *Journal of Chemical Theory and Computation*, 11(11):5436–5446, 2015. [18](#), [80](#)
- [71] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization. *J Chem Phys*, 139(14):144903, 2013. [16](#), [18](#), [20](#)
- [72] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics and Modelling*, 14(1):33–38, 1996. [30](#)
- [73] I. Ivani, P. D. Dans, A. Noy, A. Perez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goni, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpi, C. Gonzalez, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case, and M. Orozco. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55–58, 2016. [14](#), [29](#)
- [74] W. Iwasaki, Y. Miya, N. Horikoshi, A. Osakabe, H. Taguchi, H. Tachiwana, T. Shibata, W. Kagawa, and H. Kurumizaka. Contribution of histone n-terminal tails to the structure and stability of nucleosomes. *FEBS open bio*, 3(1):363–369, 2013. [10](#)

- [75] S. Izvekov and G. A. Voth. Multiscale coarse graining of liquid-state systems. *The Journal of Chemical Physics*, 123(13):134105, 2005. [17](#)
- [76] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, 2001. [10](#)
- [77] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983. [29](#)
- [78] Z. Kam, N. Borochoy, and H. Eisenberg. Dependence of laser light scattering of DNA on NaCl concentration. *Biopolymers: Original Research on Biomolecules*, 20(12):2671–2690, 1981. [12](#)
- [79] N. Kepper, R. Ettig, R. Stehr, S. Marnach, G. Wedemann, and K. Rippe. Force spectroscopy of chromatin fibers: extracting energetics and structural information from monte carlo simulations. *Biopolymers*, 95(7):435–447, 2011. [22](#)
- [80] T. A. Knotts IV, N. Rathore, D. C. Schwartz, and J. J. de Pablo. A coarse grain model for DNA. *The Journal of Chemical Physics*, 126(8):02B611, 2007. [18](#)
- [81] W. Koopmans, R. Buning, T. v. Schmidt, and J. Van Noort. spFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes. *Biophysical journal*, 97(1):195–204, 2009. [11](#)
- [82] N. Korolev, A. Allahverdi, A. P. Lyubartsev, and L. Nordenskiöld. The polyelectrolyte properties of chromatin. *Soft Matter*, 8(36):9322–9333, 2012. [9](#)
- [83] N. Korolev, A. Allahverdi, Y. Yang, Y. Fan, A. P. Lyubartsev, and L. Nordenskiöld. Electrostatic origin of salt-induced nucleosome array compaction. *Biophysical journal*, 99(6):1896–1905, 2010. [20](#)
- [84] N. Korolev, N. V. Berezhnoy, K. D. Eom, J. P. Tam, and L. Nordenskiöld. A universal description for the experimental behavior of salt-(in) dependent oligocation-induced DNA condensation. *Nucleic Acids Research*, 40(6):2807, 2012. [8](#)

- [85] N. Korolev, Y. Fan, A. P. Lyubartsev, and L. Nordenskiöld. Modelling chromatin structure and dynamics: status and prospects. *Current Opinion in Structural Biology*, 22(2):151–159, 2012. [22](#)
- [86] N. Korolev, D. Luo, A. Lyubartsev, and L. Nordenskiöld. A coarse-grained DNA model parameterized from atomistic simulations by inverse Monte Carlo. *Polymers*, 6(6):1655–1675, 2014. [15](#), [19](#), [20](#), [31](#), [36](#), [38](#), [42](#), [75](#), [80](#)
- [87] N. Korolev, L. Nordenskiöld, and A. P. Lyubartsev. Multiscale coarse-grained modelling of chromatin components: DNA and the nucleosome. *Advances in Colloid and Interface Science*, 232:36–48, 2016. [28](#), [42](#)
- [88] M. Kruithof, F.-T. Chien, A. Routh, C. Logie, D. Rhodes, and J. Van Noort. Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30 nm chromatin fiber. *Nature Structural & Molecular Biology*, 16(5):534, 2009. [11](#)
- [89] R. Kumar and H. Grubmüller. do_x3dna: a tool to analyze structural fluctuations of dsDNA or dsRNA from molecular dynamics simulations. *Bioinformatics*, 31(15):2583–2585, 2015. [45](#)
- [90] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Research*, 38(1):299–313, 2010. [14](#), [15](#)
- [91] A. Leforestier and F. Livolant. Liquid crystalline ordering of nucleosome core particles under macromolecular crowding conditions: evidence for a discotic columnar hexagonal phase. *Biophysical journal*, 73(4):1771–1776, 1997. [11](#)
- [92] M. Levitt. Computer simulation of DNA double-helix dynamics. In *Cold Spring Harbor symposia on quantitative biology*, volume 47, pages 251–262. Cold Spring Harbor Laboratory Press, 1983. [13](#)
- [93] R. J. Lewis, R. Pecora, and D. Eden. Transient electric birefringence measurements of the rotational and internal bending modes in monodisperse DNA fragments. *Macromolecules*, 19(1):134–139, 1986. [12](#)

- [94] X.-Q. Li and D. Du. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PLoS One*, 9(2):e88339, 2014. [29](#)
- [95] Y. Liu, C. Lu, Y. Yang, Y. Fan, R. Yang, C.-F. Liu, N. Korolev, and L. Nordenskiöld. Influence of histone tails and H4 tail acetylations on nucleosome–nucleosome interactions. *Journal of Molecular Biology*, 414(5):749–764, 2011. [8](#)
- [96] L. Lu, S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *Journal of Chemical Theory and Computation*, 6(3):954–965, 2010. [17](#)
- [97] L. Lu and G. A. Voth. The multiscale coarse-graining method. vii. free energy decomposition of coarse-grained effective potentials. *The Journal of Chemical Physics*, 134(22):06B607, 2011. [17](#)
- [98] X.-J. Lu and W. K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003. [2](#), [45](#)
- [99] X.-J. Lu and W. K. Olson. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature protocols*, 3(7):1213, 2008. [45](#)
- [100] Y. Lu, B. Weers, and N. C. Stellwagen. DNA persistence length revisited. *Biopolymers: Original Research on Biomolecules*, 61(4):261–275, 2002. [12](#)
- [101] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251, 1997. [4](#), [5](#)
- [102] K. Luger and T. J. Richmond. DNA binding within the nucleosome core. *Current Opinion in Structural Biology*, 8(1):33–40, 1998. [9](#), [10](#)
- [103] A. Lyubartsev and A. Laaksonen. Effective potentials for ion-DNA interactions. *The Journal of Chemical Physics*, 111(24):11207–11215, 1999. [41](#)
- [104] A. P. Lyubartsev, N. Korolev, Y. Fan, and L. Nordenskiöld. Multiscale modelling of nucleosome core particle aggregation. *Journal of Physics: Condensed Matter*, 27(6):064111, 2015. [20](#), [42](#)

- [105] A. P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach. *Physical Review E*, 52(4):3730, 1995. [17](#), [19](#), [26](#), [41](#)
- [106] M. R. Machado, E. E. Barrera, F. Klein, M. Sónora, S. Silva, and S. Pantano. The SIRAH force field 2.0: Altius, fortius, citius. *Journal of Chemical Theory and Computation*, 2019. [18](#)
- [107] M. R. Machado and S. Pantano. Exploring LacI-DNA dynamics by multiscale simulations using the SIRAH force field. *Journal of Chemical Theory and Computation*, 11(10):5012–5023, 2015. [18](#)
- [108] A. D. MacKerell and N. K. Banavali. All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000. [14](#), [29](#)
- [109] A. D. Mackerell, J. Wiorkiewicz-Kuczera, and M. Karplus. An all-atom empirical energy function for the simulation of nucleic acids. *Journal of the American Chemical Society*, 117(48):11946–11975, 1995. [14](#)
- [110] K. Maeshima, S. Hihara, and M. Eltsov. Chromatin structure: does the 30 nm fibre exist in vivo? *Current Opinion in Cell Biology*, 22(3):291–297, 2010. [5](#)
- [111] K. Maeshima, S. Ide, and M. Babokhov. Dynamic chromatin organization without the 30 nm fiber. *Current Opinion in Cell Biology*, 58:95–104, 2019. [5](#)
- [112] C. Maffeo, J. Yoo, J. Comer, D. B. Wells, B. Luan, and A. Aksimentiev. Close encounters with DNA. *Nature Structural & Molecular Biology*, 26(41):413101, 2014. [13](#)
- [113] G. S. Manning. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quarterly Reviews of Biophysics*, 11(2):179–246, 1978. [9](#)
- [114] G. S. Manning. A procedure for extracting persistence lengths from light-scattering data on intermediate molecular weight DNA. *Biopolymers: Original Research on Biomolecules*, 20(8):1751–1755, 1981. [12](#)

- [115] G. Maret and G. Weill. Magnetic birefringence study of the electrostatic and intrinsic persistence length of DNA. *Biopolymers: Original Research on Biomolecules*, 22(12):2727–2744, 1983. [12](#)
- [116] L. Mariño-Ramírez, M. G. Kann, B. A. Shoemaker, and D. Landsman. Histone structure and nucleosome stability. *Expert review of proteomics*, 2(5):719–729, 2005. [5](#), [9](#)
- [117] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries. The Martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007. [16](#), [18](#)
- [118] J. D. McGhee, J. M. Nickol, G. Felsenfeld, and D. C. Rau. Higher order structure of chromatin: orientation of nucleosomes within the 30 nm chromatin solenoid is independent of species and spacer length. *Cell*, 33(3):831–841, 1983. [5](#)
- [119] E. L. Mersfelder and M. R. Parthun. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Research*, 34(9):2653–2662, 2006. [10](#)
- [120] A. Mirzoev and A. P. Lyubartsev. MagiC: software package for multiscale modeling. *Journal of Chemical Theory and Computation*, 9(3):1512–1520, 2013. [17](#), [23](#), [41](#), [42](#)
- [121] A. Mirzoev, L. Nordenskiöld, and A. P. Lyubartsev. Magic v.3: An integrated software package for systematic structure-based coarse-graining. *Comput. Phys. Commun.*, 237:263–273, 2019. [23](#)
- [122] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks. Sequence-dependent persistence lengths of DNA. *Journal of Chemical Theory and Computation*, 13(4):1539–1555, 2017. [45](#)
- [123] V. Morales and H. Richard-Foy. Role of histone n-terminal tails and their acetylation in nucleosome dynamics. *Molecular and cellular biology*, 20(19):7230–7237, 2000. [5](#)
- [124] A. Naômé, A. Laaksonen, and D. P. Vercauteren. A solvent-mediated coarse-grained model of DNA derived with the systematic newton inversion method. *Journal of Chemical Theory and Computation*, 10(8):3541–3549, 2014. [80](#)

- [125] A. Naômé, A. Laaksonen, and D. P. Vercauteren. A coarse-grained simulation study of the structures, energetics, and dynamics of linear and circular DNA with its ions. *Journal of Chemical Theory and Computation*, 11(6):2813–2826, 2015. [15](#), [19](#), [20](#)
- [126] L. Nilsson, G. M. Clore, A. M. Gronenborn, A. T. Brunger, and M. Karplus. Structure refinement of oligonucleotides by molecular dynamics with nuclear overhauser effect interproton distance restraints: application to 5' d(C-G-T-A-C-G)₂. *Journal of Molecular Biology*, 188(3):455–475, 1986. [14](#)
- [127] L. Nilsson and M. Karplus. Empirical energy functions for energy minimization and dynamics of nucleic acids. *Journal of Computational Chemistry*, 7(5):591–616, 1986. [14](#)
- [128] L. G. Nilsson, L. Guldbrand, and L. Nordenskiöld. Evaluation of the electrostatic osmotic pressure in an infinite system of hexagonally oriented DNA molecules: A Monte Carlo simulation study. *Molecular Physics*, 72(1):177–192, 1991. [8](#)
- [129] Y. Nishino, M. Eltsov, Y. Joti, K. Ito, H. Takata, Y. Takahashi, S. Hihara, A. S. Frangakis, N. Imamoto, T. Ishikawa, et al. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30 nm chromatin structure. *The EMBO journal*, 31(7):1644–1653, 2012. [6](#)
- [130] L. Nordenskiöld, A. P. Lyubartsev, and N. Korolev. Coarse-grained modeling of nucleosomes and chromatin. In G. A. Papoian, editor, *Coarse-Grained Modeling of Biomolecules*, Series in Computational Biophysics, pages 291–334. CRC Press, 2017. [18](#)
- [131] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X. J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C. S. Tung, E. Westhof, C. Wolberger, and H. M. Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology*, 313(1):229–37, 2001. [3](#), [51](#)
- [132] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19):11163–11168, 1998. [3](#)

- [133] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981. [30](#)
- [134] M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton, et al. μ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Research*, 42(19):12272–12283, 2014. [60](#)
- [135] A. Perez, F. Lankas, F. J. Luque, and M. Orozco. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Research*, 36(7):2379–2394, 2008. [14](#), [15](#)
- [136] A. Perez, F. Luque, and M. Orozco. Frontiers in molecular dynamics simulations of DNA. *Accounts of Chemical Research*, 45(2):196–205, 2011. [13](#)
- [137] A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophysical Journal*, 92(11):3817–3829, 2007. [14](#), [29](#)
- [138] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995. [44](#), [73](#)
- [139] K. Polach, P. Lowary, and J. Widom. Effects of core histone tail domains on the equilibrium constants for dynamic DNA site accessibility in nucleosomes. *Journal of Molecular Biology*, 298(2):211–223, 2000. [10](#)
- [140] D. Porschke. Persistence length and bending dynamics of DNA from electrooptical measurements at high salt concentrations. *Biophysical chemistry*, 40(2):169–179, 1991. [12](#)
- [141] C. B. Post. Excluded volume of an intermediate-molecular-weight DNA. A Monte Carlo analysis. *Biopolymers: Original Research on Biomolecules*, 22(4):1087–1096, 1983. [12](#)
- [142] D. A. Potoyan, A. Savelyev, and G. A. Papoian. Recent successes in coarse-grained modeling of DNA. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(1):69–83, 2013. [18](#)

- [143] E. Raspaud, I. Chaperon, A. Leforestier, and F. Livolant. Spermine-induced aggregation of DNA, nucleosome, and chromatin. *Biophysical journal*, 77(3):1547–1555, 1999. [8](#)
- [144] S. Y. Reddy, F. Leclerc, and M. Karplus. DNA polymorphism: a comparison of force fields for nucleic acids. *Biophysical journal*, 84(3):1421–1449, 2003. [4](#), [14](#)
- [145] T. Richmond, J. Finch, B. Rushton, D. Rhodes, and A. Klug. Structure of the nucleosome core particle at 7 Å resolution. *Nature*, 311(5986):532, 1984. [4](#)
- [146] P. J. Rizzo. Those amazing dinoflagellate chromosomes. *Cell research*, 13(4):215, 2003. [10](#)
- [147] V. Rizzo and J. Schellman. Flow dichroism of T7 DNA as a function of salt concentration. *Biopolymers: Original Research on Biomolecules*, 20(10):2143–2163, 1981. [12](#)
- [148] I. Rouzina and V. A. Bloomfield. DNA bending by small, mobile multivalent cations. *Biophysical journal*, 74(6):3152–3164, 1998. [8](#)
- [149] B. Rydberg, W. R. Holley, I. S. Mian, and A. Chatterjee. Chromatin conformation in living cells: support for a zig-zag model of the 30 nm chromatin fiber. *Journal of Molecular Biology*, 284(1):71–84, 1998. [5](#)
- [150] A. Savelyev. Do monovalent mobile ions affect dna’s flexibility at high salt content? *Physical Chemistry Chemical Physics*, 14(7):2250–2254, 2012. [12](#)
- [151] A. Savelyev and G. A. Papoian. Molecular renormalization group coarse-graining of polymer chains: application to double-stranded dna. *Biophysical journal*, 96(10):4044–4052, 2009. [17](#), [19](#)
- [152] A. Savelyev and G. A. Papoian. Chemically accurate coarse graining of double-stranded DNA. *Proceedings of the National Academy of Sciences*, 107(47):20340–20345, 2010. [15](#), [19](#), [20](#), [80](#)
- [153] T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436(7047):138, 2005. [11](#)

- [154] M. P. Scheffer, M. Eltsov, J. Bednar, and A. S. Frangakis. Nucleosomes stacked with aligned dyad axes are found in native compact chromatin in vitro. *Journal of Structural Biology*, 178(2):207–214, 2012. [10](#)
- [155] G. Seibel, U. Singh, and P. Kollman. A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proceedings of the National Academy of Sciences*, 82(19):6537, 1985. [14](#)
- [156] K. Shahzad and J. J. Loor. Application of top-down and bottom-up systems approaches in ruminant physiology and metabolism. *Current Genomics*, 13(5):379–394, 2012. [16](#)
- [157] A. K. Shaytan, G. A. Armeev, A. Goncarencu, V. B. Zhurkin, D. Landsman, and A. R. Panchenko. Coupling between histone conformations and DNA geometry in nucleosomes on a microsecond timescale: atomistic insights into nucleosome functions. *Journal of Molecular Biology*, 428(1):221–237, 2016. [13](#)
- [158] E. Sobel and J. Harpst. Effects of Na⁺ on the persistence length and excluded volume of T7 bacteriophage DNA. *Biopolymers: Original Research on Biomolecules*, 31(13):1559–1564, 1991. [12](#)
- [159] R. Stehr, N. Kepper, K. Rippe, and G. Wedemann. The effect of internucleosomal interaction on folding of the chromatin fiber. *Biophysical journal*, 95(8):3677–3691, 2008. [22](#)
- [160] A. Stone. The description of bimolecular potentials, forces and torques: the S and V function expansions. *Molecular Physics*, 36(1):241–256, 1978. [22](#)
- [161] J. Sun, Q. Zhang, and T. Schlick. Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *Proceedings of the National Academy of Sciences*, 102(23):8180–8185, 2005. [21](#)
- [162] T. Sun, A. Mirzoev, V. Minhas, N. Korolev, A. P. Lyubartsev, and L. Nordenskiöld. A multiscale analysis of DNA phase separation: from atomistic to mesoscale level. *Nucleic Acids Research*, 47(11):5550–5562, 2019. [xv](#), [xix](#), [17](#), [39](#), [77](#), [81](#), [82](#)
- [163] S. Takada. Coarse-grained molecular simulations of large biomolecules. *Current Opinion in Structural Biology*, 22(2):130–137, 2012. [19](#)

- [164] R. C. Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, 2013. [45](#)
- [165] F. Thoma, T. Koller, and A. Klug. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *The Journal of Cell Biology*, 83(2):403–427, 1979. [10](#)
- [166] C. Tisné, M. Delepierre, and B. Hartmann. How NF- κ B can be attracted by its cognate DNA. *Journal of Molecular Biology*, 293(1):139–150, 1999. [4](#)
- [167] P. Tropberger and R. Schneider. Scratching the (lateral) surface of chromatin regulation by histone modifications. *Nature Structural & Molecular Biology*, 20(6):657, 2013. [10](#)
- [168] C. Tse and J. C. Hansen. Hybrid trypsinized nucleosomal arrays: identification of multiple functional roles of the H2A/H2B and H3/H4 N-termini in chromatin fiber compaction. *Biochemistry*, 36(38):11381–11388, 1997. [10](#)
- [169] C. Tse, T. Sera, A. P. Wolffe, and J. C. Hansen. Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. *Molecular and Cellular Biology*, 18(8):4629–4638, 1998. [10](#)
- [170] D. W. Ussery. Dna structure: A-, b-and z-dna helix families. *e LS*, 2001. [2](#)
- [171] J. J. Uusitalo, H. I. Ingólfsson, P. Akhshi, D. P. Tieleman, and S. J. Marrink. Martini coarse-grained force field: extension to DNA. *Journal of Chemical Theory and Computation*, 11(8):3932–3945, 2015. [18](#), [20](#)
- [172] K. Van Holde and J. Zlatanova. What determines the folding of the chromatin fiber? *Proceedings of the National Academy of Sciences*, 93(20):10548–10555, 1996. [10](#)
- [173] K. E. van Holde. Chromatin structure and transcription. In *Chromatin*, pages 355–408. Springer, 1989. [8](#)
- [174] D. Vasudevan, E. Y. Chua, and C. A. Davey. Crystal structures of nucleosome core particles containing the ‘601’ strong positioning sequence. *Journal of Molecular Biology*, 403(1):1–10, 2010. [13](#)

- [175] K. Voltz, J. Trylska, V. Tozzini, V. Kurkal-Siebert, J. Langowski, and J. Smith. Coarse-grained force field for the nucleosome from self-consistent multiscale. *Journal of Computational Chemistry*, 29(9):1429–1439, 2008. [17](#)
- [176] J. M. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000. [14](#)
- [177] J. D. Watson, F. H. Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. [2](#)
- [178] K. Wecker, M. Bonnet, E. Meurs, and M. Delepierre. The role of the phosphorus BI-BII transition in protein-DNA recognition: the NF- κ B complex. *Nucleic Acids Research*, 30(20):4452–4459, 2002. [4](#)
- [179] G. Wedemann and J. Langowski. Computer simulation of the 30 nm chromatin fiber. *Biophysical journal*, 82(6):2847–2859, 2002. [5](#), [6](#), [21](#), [22](#)
- [180] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984. [14](#)
- [181] H. Weintraub, K. Palter, and F. Van Lente. Histones H2A, H2B, H3, and H4 form a tetrameric complex in solutions of high salt. *Cell*, 6(1):85–110, 1975. [7](#)
- [182] J. Widom and R. L. Baldwin. Cation-induced toroidal condensation of dna: studies with Co^{3+} (nh₃)₆. *Journal of Molecular Biology*, 144(4):431–453, 1980. [9](#)
- [183] J. Widom and A. Klug. Structure of the 3000 Å chromatin filament: X-ray diffraction from oriented samples. *Cell*, 43(1):207–213, 1985. [5](#)
- [184] R. W. Wilson and V. A. Bloomfield. Counterion-induced condensation of deoxyribonucleic acid. a light-scattering study. *Biochemistry*, 18(11):2192–2196, 1979. [9](#)

- [185] D. Winogradoff and A. Aksimentiev. Molecular mechanism of spontaneous nucleosome unraveling. *Journal of Molecular Biology*, 431(2):323–335, 2019. [13](#)
- [186] A. Wolffe. *Chromatin: structure and function*. Academic press, 1998. [7](#)
- [187] H. Wong, J.-M. Victor, and J. Mozziconacci. An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PloS one*, 2(9), 2007. [6](#)
- [188] C. Woodcock, S. Grigoryev, R. Horowitz, and N. Whitaker. A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. *Proceedings of the National Academy of Sciences*, 90(19):9021–9025, 1993. [5](#)
- [189] A. Worcel, S. Strogatz, and D. Riley. Structure of chromatin and the linking number of DNA. *Proceedings of the National Academy of Sciences*, 78(3):1461–1465, 1981. [5](#), [10](#)
- [190] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–574, 2006. [2](#)
- [191] Z. Yang and J. J. Hayes. The divalent cations Ca^{2+} and Mg^{2+} play specific roles in stabilizing histone-DNA interactions within nucleosomes that are partially redundant with the core histone tail domains. *Biochemistry*, 50(46):9973–9981, 2011. [8](#)
- [192] J. Yoo and A. Aksimentiev. Improved parametrization of Li^+ , Na^+ , K^+ , and Mg^{2+} ions for all-atom molecular dynamics simulations of nucleic acid systems. *J. Phys. Chem. Lett.*, 3(1):45–50, 2012. [14](#)
- [193] J. Yoo and A. Aksimentiev. Improved parameterization of amine–carboxylate and amine–phosphate interactions for molecular dynamics simulations using the charmm and amber force fields. *Journal of Chemical Theory and Computation*, 12(1):430–443, 2015. [30](#)
- [194] H. Zewdie. Computer-simulation studies of diskotic liquid crystals. *Physical Review E*, 57(2):1793, 1998. [22](#)

- [195] M. Zgarbová, P. Jurečka, J. Šponer, and M. Otyepka. A to B-DNA transition in AMBER force fields and its coupling to sugar pucker. *Journal of Chemical Theory and Computation*, 14(1):319–328, 2017. [13](#)
- [196] M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš, and P. Jurečka. Base pair fraying in molecular dynamics simulations of DNA and RNA. *Journal of Chemical Theory and Computation*, 10(8):3177–3189, 2014. [51](#)
- [197] Q. Zhang, D. A. Beard, and T. Schlick. Constructing irregular surfaces to enclose macromolecular complexes for mesoscale modeling using the discrete surface charge optimization (DiSCO) algorithm. *Journal of Computational Chemistry*, 24(16):2063–2074, 2003. [21](#), [22](#)