Alma Mater Studiorum - Università di Bologna

SCUOLA DI SCIENZE

Dipartimento di Chimica Industriale "Toso Montanari"

Corso di Laurea Magistrale in

Chimica Industriale

Classe LM-71 - Scienze e Tecnologie della Chimica Industriale

Simulating the aggregation of DNA oligonucleotides

Tesi di laurea sperimentale

CANDIDATO

Silvia Cristofaro

RELATORE

Prof.ssa Silvia Orlandi

CORRELATORI

Prof. Roberto Berardi

Prof. Luca Muccioli

Prof. Alberto Arcioni

Prof. Claudio Zannoni

Dott.ssa Lara Querciagrossa

Anno Accademico 2018-2019

Contents

Abstract **1**

1 Introduction **3**

2 DNA thermodynamics and self-assembly **7**

2.1 Double Helix Interaction 7

2.2 DNA melting temperature 10

2.2.1 Two-state model 11

2.2.2 Nearest-neighbor method 13

2.3 DNA Liquid Crystal ordering 15

2.3.1 Self-assembly of DNA structures 15

2.4 Liquid Crystal mesophases 16

3 DNA computational models **25**

3.1 State of the art 25

3.1.1 De Pablo coarse-grained model 28

3.1.2 Plotkin coarse-grained model 29

3.2 OxDNA model 31

3.2.1 The averaged base model: oxDNA1 32

3.2.2 The new model: oxDNA2 35

4 Double strand melting transition **43**

4.1 Objectives 43

4.2 Methods 44

4.3 Results 46

4.3.1 Thermodynamic parametrization and comparison . . . 46

4.3.2 Comparison of melting temperatures 49

4.4 Summary 54

5 LC short strand ordering **55**

5.1 Phase diagram reproduction 55

5.2 Methods 57

5.3 Results 58

5.3.1 Orientational parameter 58

5.3.2 Radial distribution function g(r) 60

5.3.3 Local structure 62

5.3.4 Phase organization 65

5.3.5 Introduction of the modified DH potential 69

5.4 Summary 70

Conclusions and future outlooks **71**

Bibliography **i**

Appendices **iii**

A Molecular Dynamics **v**

A.1 Integration of the equations of motion vi

A.1.1 Verlet Integration vi

A.2 Constant temperature molecular dynamics vii
A.3 Finite size effects and boundary conditions vii

Abstract

In this work we have studied, by means of Molecular Dynamics simulations, the process of denaturation and self-assembly of short oligonucleotides. Supramolecular ordering of DNA short strands is a promising field which is constantly enriched with new findings. Examples are provided by micellar and fibrils systems formations and due to the selectivity of DNA bindings, "intelligent" devices have been developed to perform simple logic operations. It is worth to notice that computer simulations of these DNA nanosystems would complement experiments with detailed insight into processes involved in self-assembly. In order to obtain an accurate description of the interactions involved in the complex structure of DNA we used oxDNA, a coarse-grained model developed by Ouldrige *et al* [1], [2]. We simulated the melting transition of 4, 6, and 8 base pair sequences. Sequence and length dependence were analyzed, specifically we compared thermodynamic parameters ΔH, ΔS and the melting temperature T_m with literature results. Moreover, we have attemped to reproduce liquid crystal ordering of the ultrashort sequence GCCG at relatively high saline concentration, until now only experimentally observed in Bellini's *et al.* [3] works. We found that our simple model successfully reproduces the experimental phase sequence (isotropic, nematic, columnar) at $T = 5$ °C as a function of oligonucleotide concentration, and we fully characterized the microscopic structure of the three phases.

Chapter 1

Introduction

DNA is one of the most important bio-polymers, as its sequence encodes the genetic instructions needed in the development and functioning of many living organisms. While we know nowadays the sequences of many genomes, we still know little as to how DNA is organized in 3D inside a living cell, and of how gene regulation and DNA function are coupled to this structure. DNA is a long polymer made of repeating units called nucleotides [4]-[5], comprised of a sugar deoxyribose, a nitrogen base (nucleobase) and a phosphate group as shown in Fig 1.1 .

There are four types of nucleobases in DNA: adenine (A), cytosine (C), guanine (G) and thymine (T). While A and G are purines, which are fused five- and six-membered heterocyclic compounds, C and T are pyrimidines, that are six-membered rings [6]. A always pairs with T by two hydrogen bonds and G always pairs with cytosine C by three hydrogen bonds: this complimentarity is known as the base pairing rule. Each polynucleotide is characterized by a specific nucleobase sequence which carries the genetic information. DNA does not usually exist as a single strand in human being and in vivo, but as a double strand, composed by two single strands, that

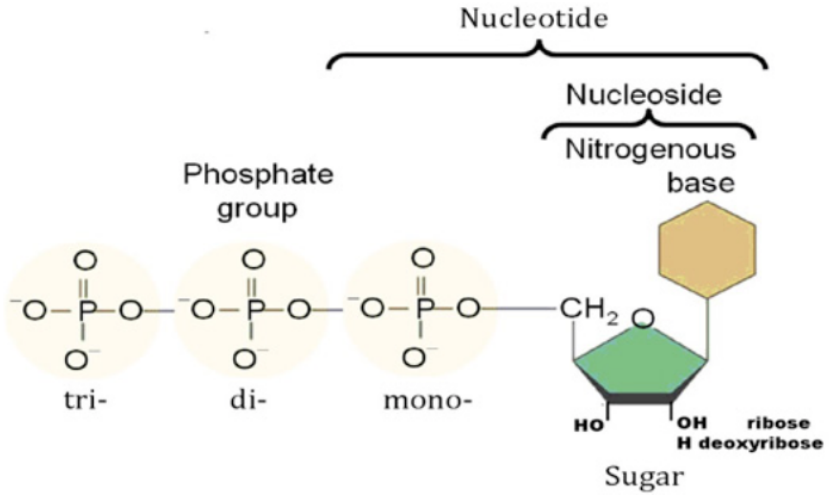

Figure 1.1: A nucleoside is composed of a pentose sugar and a base, while adding one or more phosphate groups a nucleotide is formed.

are spirally twisted around each other and coiled around a common axis to form a right-handed double-helix.The sugar-phosphate backbones remain on the outside, while the core of the helix contains the purine and pyrimidine bases. The DNA helix has a shallow groove called minor groove (1.2 nm) and a deep groove called major groove (2.2 nm) across (see Fig 1.2).

Any DNA strand normally has one end at which there is a phosphoryl attached to the 5' carbon of a ribose (the 5' phosphoryl) and another end at which there is a free hydroxyl attached to the 3' carbon of a ribose (the 3' hydroxyl). The orientation of the 3' and 5' carbons along the sugar-phosphate backbone confers directionality to each DNA strand. The sugar-phosphate backbones of the two DNA strands are antiparallel with the 3' terminal of one strand connected to the 5' terminal of the other (see Fig 1.3). In the most common structure, the B-form, the bases remain on average perpendicular to the molecule axis. Helix periodicity is 10 base pairs, equivalent to 3.4 nm, while the diameter of the bare DNA is 2 nm. Interestingly, DNA

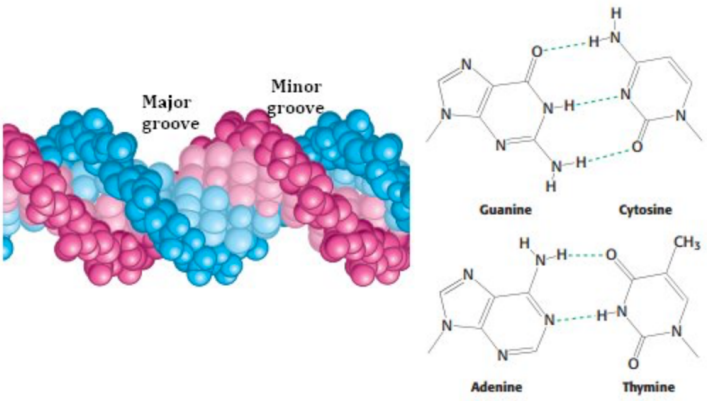

Figure 1.2: The major groove occurs where the backbones are far apart, the minor groove occurs where they are close together. The grooves twist around the molecule on opposite sides.

is a remarkably flexible molecule. Considerable rotation is possible thanks to a large number of bonds in the sugar-phosphate backbone, and thermal fluctuation can produce bending, streching, and unpairing of strands.

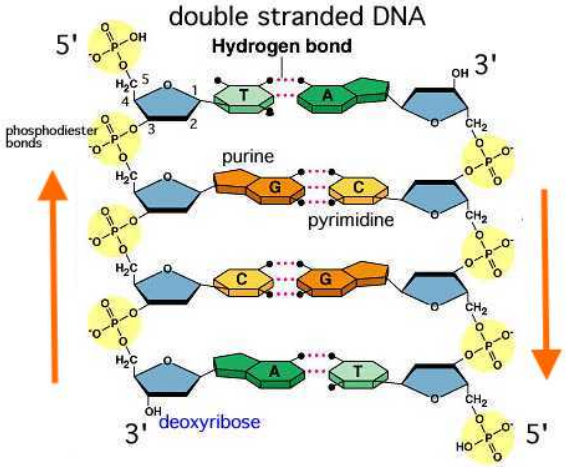

Figure 1.3: Structure of a DNA double strand showing H-bond and strand directionality. The 5' and 3' ends can be clearly distinguished. [Adapted from Pearson Education, Inc., publishing as Benjamin Cummings.]

This thesis is organized as follows: Chapter 2 is divided into two main sections: in the first, we give more details on DNA thermodynamics, while for section two we reported LC short oligonucleotides ordering from literature; Chapter 3 is mainly focused on the determination of the melting temperature of short double strands; Chapter 4 is based on our attempt to reproduce an experimental phase-diagram for a 4 base pair single strand and verify its LC nematic and columnar phase ordering; In Chapter 5 we give a brief summary of our results and we present our future works.

Chapter 2

DNA thermodynamics and self-assembly

2.1 Double Helix Interaction

The DNA double helix is stabilized primarily by two forces: inter-strand hydrogen bonds between nucleotides and intra-strand base stacking interaction between aromatic nucleobases (as shown in Fig 2.1) [7]. Nevertheless, ionic and hydrophobic forces should also be taken into account when discussing the interaction between helices and the aggregation in solution.

The energy associated with the formation of the double helix ("hybridization") depends on the length of the polynucleotide and on the quality of the Watson and Crick (WC) matching. It is important to note that three hydrogen bonds can form between G and C, but only two bonds can be found in A and T pairs (Fig 2.2). This is why it is more difficult to separate DNA strands that contain more G-C pairs than A-T pairs. Even if the description of the pairing of complementary bases is simple, the quantitative evaluation of the energy involved is complicated because, when unpaired, the nucle-

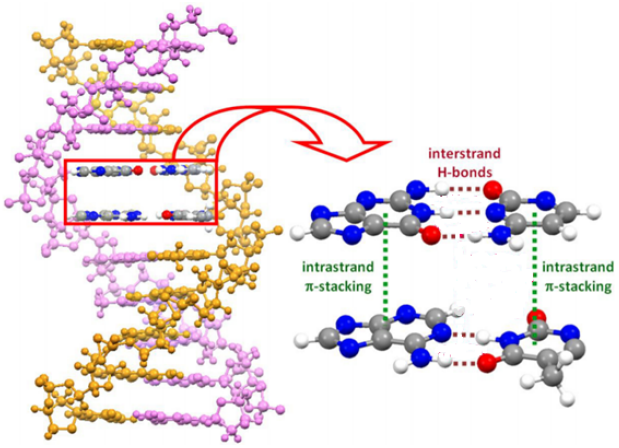

Figure 2.1: Schematic representation of the intra-strand an inter-strand interactions.

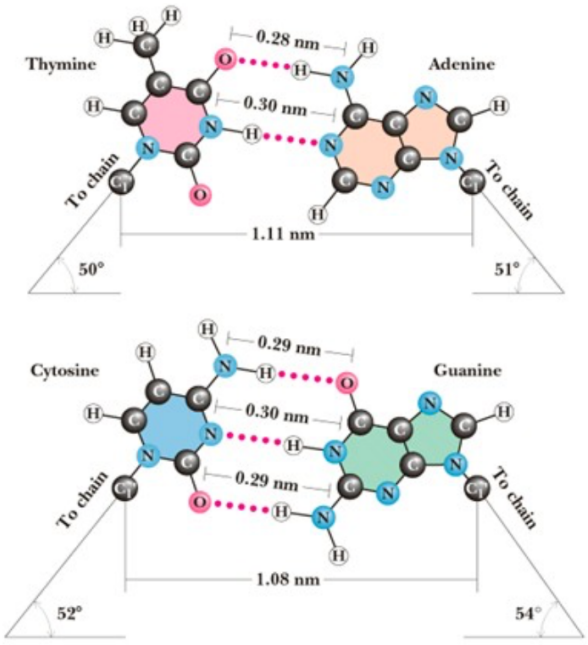

Figure 2.2: Hydrogen bonds between complementary bases according to Watson and Crick theory.

obases form H-bonds with water. Base stacking interactions, instead, act to pull the aromatic hydrocarbon flat surfaces of the nucleobases into contact

with each other. It is accepted that the π interactions are mainly determined by the interplay of the van der Waals dispersion force, electrostatic forces, the exchange (Pauli)-repulsion at short intermolecular distance, however, there are still lot of debates on which one dominates. It would seem that the relative contributions are highly dependent on the geometry of the molecules. Stacking is not selective, i.e. can happen between any bases and its strength mainly depends on the overlap between the aromatic planes. Nevertheless it is the main source of stability of the double helix, while pairing provides the necessary specificity of the bonds. When combined with the constraint of the sugar-phosphate backbone, the stacking forces result in the twist of a base pair relative to the neighboring ones of 36°. For systems with water as a solvent, there is also an hydrophobic effect contribution since the base surfaces tend to minimize their exposure to water. Double helix is stabilized by hydrophobic effects burying the bases in the interior of the helix increases its stability; having the hydrophobic bases clustered in the interior of the helix keeps it away from the surrounding water, whereas the more polar phosphate groups and riboses are exposed and interact with the exterior water. The hydrophobic interactions provide stability to the structure but do not contribute to the specificity. Finally it is worth to notice that ionic interactions play an important role for DNA: ion-ion repulsion of the negatively charged phosphate makes DNA duplex unstable, however the presence of Mg^{2+} or cationic proteins with abundant Arginine and Lysine (zwitterionic) residues stabilize the double helix. A double-stranded helix structure is then promoted by having phosphates on outside which interact with H_2O and counter ions (K^+, Mg^{2+}, etc.).

It is generally found that the formation of a DNA duplex involves negative variations of both enthalpy (ΔH) and entropy (ΔS), where $\Delta G = \Delta H$-$T\Delta S$.

The large number of experimental observations has allowed the development of simplified strategies to quantify ΔG, once the sequence of two strands involved are given. The most commonly adopted approach, the "nearest neighbor" (NN) model [8], is founded on the idea that the minimum unit which contains both the effects of pairing and stacking is a quadruplet formed by two consecutive nucleotides on one strand and the corresponding nucleotides on the other strand. This approach and its use in this thesis are discussed in Section 2.2.2 of Chapter 2.

2.2 DNA melting temperature

The overall values of ΔG, ΔH, and ΔS for each sequence can be evaluated through melting experiments. In fact, as T is increased, the weak non-covalent bonds break and the double helices become unstable and denaturated. The melting temperature (or denaturation temperature) T_m is defined as the temperature at which half of the double helices are actually separated in two isolated single strands (ssDNA) (Fig 2.3). The transition can be inspected by optical (absorbance, fluorescence) or calorimetric methods. T_m is determined as the middle point of the single strands to double helix transition. While in cells denaturation occurs thanks to specific enzymes, in vitro melting DNA by heat is a standard method for preparing ssDNA, then the mixture obtained is cooled to allow strands to rehybridize. Hybrid helices are formed between similar sequences and any differences between those sequences will result in a partial disruption of the base-pairing. On a genomic scale, the method has been used by researchers to estimate the genetic distance between two species, a process known as DNA-DNA hybridization [9].

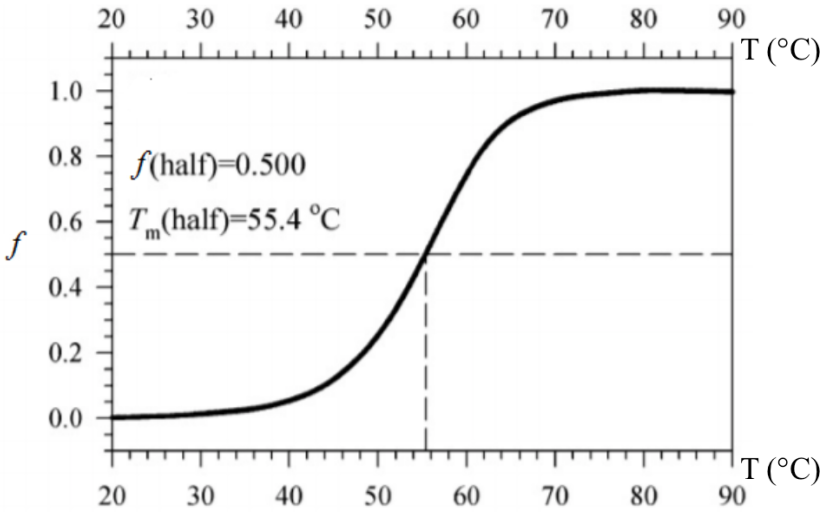

Figure 2.3: Thermal denaturation transition of a DNA helix: fraction of single strands f extracted from experimental data (UV absorbance vs. temperature) which defines the melting temperature corresponding to $f = 0.5$. [Adapted from R. Owczarzy, Melting temperatures of nucleic acids: Discrepancies in analysis, Biophys Chem 117 (2005) 207]

2.2.1 Two-state model

For DNA oligonucleotides, i.e. short sequences of DNA, the thermodynamics of hybridization can be accurately described as a two-state process. In this approximation it is neglected the possibility of intermediate partial binding states in the formation of a double strand state from two single stranded oligonucleotides. The steep part of the melting curve reflects the double strand (dsDNA) to single strand equilibrium (ssDNA1 + ssDNA2).

$$dsDNA \rightleftharpoons ssDNA1 + ssDNA2 \tag{2.1}$$

The equilibrium constant for this reaction is:

$$K = \frac{[ssDNA1][ssDNA2]}{[dsDNA]} \tag{2.2}$$

where [ssDNA1], [ssDNA2], [dsDNA] are the molar concentrations. According to the definition of equilibrium constant the relation between free energy, ΔG, and K is ΔG = -RTln K, where R is the gas constant, and T is the temperature of reaction in Kelvin. This gives, for the nucleic acid system,

$$\Delta G = -RT \ln \frac{[ssDNA1][ssDNA2]}{[dsDNA]} \tag{2.3}$$

If no additional nucleic acids are present, then, at the melting temperature, [ssDNA1], [ssDNA2], and [dsDNA] will be equal, and equal to half the initial concentration of double-stranded nucleic acid. If one starts the experiment at a temperature well below the melting point, this gives an expression for the melting point of a nucleic acid duplex of:

$$T_m = -\Delta G \frac{1}{R \ln(0.5[DNA])} \tag{2.4}$$

With

$$[DNA] = \frac{[ssDNA1] + [ssDNA2]}{2} + [dsDNA] \tag{2.5}$$

Since $\Delta G = \Delta H$ -TΔS, T is also given by:

$$T_m = \Delta H \frac{1}{\Delta S - R \ln(0.5[DNA])} \tag{2.6}$$

The terms ΔH and ΔS are usually given for the association and not for the dissociation reaction and are assumed to be temperature independent.

As mentioned, this equation is based on the assumption that only two states are involved in melting. It considers the duplex and single-stranded states and assumes that there is a constant enthalpy and entropy difference

between the two. In other words, it neglects the variation in enthalpy within the bound and unbound sub-ensembles [2]. However, nucleic acids may melt via several intermediate states. To account for such complicated behavior, the methods of statistical mechanics must be used, especially for long sequences.

2.2.2 Nearest-neighbor method

An in-depth analysis of DNA oligonucleotides and their corresponding experimental T_m has led to the conclusion that not only the relative amount of cytosine and guanine concentration determines the thermal denaturation of DNA, but also the specific nucleotide order defining DNA sequences.
Hence the nearest-neighbour (NN) model was adopted for the calculation of sequence related T_ms and can be extended beyond the Watson-Crick pairs to include parameters for interactions between mismatches and neighboring base pairs [10]. This method allows the estimation of thermodynamic parameters of sequences containing isolated mismatches. The NN model postulates that the free energy for duplex formation depends mostly on two factors: first, the initiation-free energy given by an unfavorable entropy as a consequence of loss of translational freedom after the first DNA/DNA pair is formed; and second, the sum of complementary pairwise terms between the oligonucleotide sequences, which are based on dinucleotide entities. In addition to these two terms, an entropic penalty is also considered, which takes into account the maintenance of simmetry in self-complemetary sequences. The calculation of T_m values by the NN method requires a set of experimental thermodynamic parameters as an input, which are reported in Table 2.4.

A particularly important variant of the NN, which has been shown to reproduce experimental melting temperatures of duplexes ranging from 4

propagation sequence	ΔH° (kcal/mol)	ΔS° (eu)	ΔG°_{37} (kcal/mol)
AA/TT	-8.4 ± 0.7	-23.6 ± 1.8	-1.02 ± 0.04
AT/TA	-6.5 ± 0.8	-18.8 ± 2.3	-0.73 ± 0.05
TA/AT	-6.3 ± 1.0	-18.5 ± 2.6	-0.60 ± 0.05
CA/GT	-7.4 ± 1.1	-19.3 ± 2.9	-1.38 ± 0.06
GT/CA	-8.6 ± 0.7	-23.0 ± 2.0	-1.43 ± 0.05
CT/GA	-6.1 ± 1.2	-16.1 ± 3.3	-1.16 ± 0.07
GA/CT	-7.7 ± 0.7	-20.3 ± 1.9	-1.46 ± 0.05
CG/GC	-10.1 ± 0.9	-25.5 ± 2.3	-2.09 ± 0.07
GC/CG	-11.1 ± 1.0	-28.4 ± 2.6	-2.28 ± 0.08
GG/CC	-6.7 ± 0.6	-15.6 ± 1.5	-1.77 ± 0.06
initiation at G·C	(0)	(-5.9 ± 0.8)	$+1.82 \pm 0.24$
initiation at A·T	(0)	(-9.0 ± 3.2)	$(+2.8 \pm 1)$
symmetry correction	0	-1.4	$+0.4$
5′-terminal T·A bp	$+0.4$	0	$+0.4$

Figure 2.4: Thermodynamic parameters for DNA helix initiation and propagation in 1 M NaCl at 37 °C [Adapted from [8]]

.

to 16 base pairs (bp) in length with a standard deviation of 2.3 K, was introduced by SantaLucia and Hicks [10, 11, 8]. In this model, ΔH_{AB} and ΔS_{AB} are computed by summing contributions for each nearest-neighbor set of two base pairs, together with terms for helix initiation and various structural features, all of which are assumed to be temperature independent. Fig 2.5 illustrates the calculation of ΔG°_{37}, for a sequence GCTAGC at a temperature of 37 °C using the parameters in Table 2.4.

$$\Delta G^{\circ}_{37}(\text{predicted}) = 2\,\Delta G^{\circ}_{37}(\text{GC/CG}) + 2\,\Delta G^{\circ}_{37}(\text{CT/GA}) + \Delta G^{\circ}_{37}(\text{TA/AT}) + \Delta G^{\circ}_{37}(\text{init}) + \Delta G^{\circ}_{37}(\text{sym})$$

$$= 2\,(-2.28) + 2(-1.16) + (-0.60) + 1.82 + 0.4$$

$$\Delta G^{\circ}_{37}(\text{predicted}) = -5.26 \text{ kcal/mol.}$$

$$\Delta G^{\circ}_{37}(\text{observed}) = -5.3 \text{ kcal/mol.}$$

Figure 2.5: The calculation of total ΔG comes form a sum over couples of bases along the strand with orientation 5' to 3' plus the sum over couples of bases along the strand with opposite orientation.

2.3 DNA Liquid Crystal ordering

2.3.1 Self-assembly of DNA structures

Liquid crystals (LCs) are nowadays recognized as a common form of self-organization of soft and biological matter. LC ordering in DNA was first observed in vitro with long double strands; later they were recognized as the in vivo packing mechanism of some organisms, and quite recently they started being considered as a common ordering for DNA oligomers. Self-assembly is the spontaneous and reversible formation through free energy minimization of aggregates of basic building blocks. The size of aggregating units can vary from a few Åto μm. An increasing interest on these materials comes from the possibility to control their physical properties by tuning the interaction of the individual building blocks. Examples are provided by the formation of micellar systems, fibers and fibrils, from solution of long duplex B-form DNA composed of 10^2 to 10^6 base pairs. The self-assembly of short

strands (oligonucleotides) was first demonstrated by the Seeman group [12], who created a four-armed junction. Junction of this type, and more complex motifs have been used to create lattices and ribbons [1] .

2.4 Liquid Crystal mesophases

Liquid crystals are equilibrium phases of matter that present typical liquid properties, such as fluidity, and at the same time those of crystals, like optical and mechanical anisotropy. For this reason, these particular states of matter are called mesophases. We can distinguish two principal classes of mesophases:

- Thermotropic LC are pure systems or mixtures of mesogenes which undergo a phase transition by changing the temperature;
- Lyotropic LC, which depend for their formation on both temperature and mainly mesogen concentration in a solvent.

From a molecular point of view, the key property of liquid crystals resides in the spontaneous alignment of mesogens along a preferred direction, called director. As a consequence, the liquid crystal domains typically present uniaxial symmetry around the director. Depending on the type of order we can identify four main families as seen in Fig 2.6:

- Nematic phases: present long-range orientational, but not positional, order. A relatively high reorientational mobility of the molecules reflects on properties such as density, viscosity and diffusion coefficients

that are similar to those of liquids. The orientational order gives, however, strong anisotropy on electrical, optical, mechanical and magnetic properties;

- Cholesteric phases (or chiral nematic N* phases): they are a positionally disordered fluid in which the molecules align on average their axes along a common direction called the nematic director, but the orientational order develops an additional macro-helical superstructure with the twist axis perpendicular to the local director (see Fig 2.7 with arrows representing the local director).;

- Smectic phases: mesogens show both orientational and monodimensional positional order, typically in layered structures;

- Columnar phases: besides having long-range orientational order, they also present two dimensional order. This is possible thanks to the discotic shape of mesogens that can self-assemble into column-like aggregates.

The ability of duplex DNA to form LC phases was discovered in the late 1940s. Since that time, the LC phases of solution of duplex B-form DNA (B-DNA) have been characterized by polarized optical microscopy, X-ray, and magnetic resonance methods. Two main mesophases have been identified :

- Cholesteric (N*) phase: it starts to appear at a concentration around 150 mg/ml in 100 mM monovalent salt conditions. This LC phase is easily observed in polarized optical microscopy. Since the N* pitch extends to tens of micrometers (that is, across more than 500 molecules), the optical anisotropy of DNA bases leads to characteristic textures such as the droplet in Fig. 2.8 (left), where the stripes correspond to the macrohelical periodicity of the N* ordering.

Phase	Positional order	Orientational order
Solid crystal	Yes (3D)	Yes
Columnar LC	Yes (2D)	Yes
Smectic LC	Yes (1D)	Yes
Nematic LC	No	Yes
Isotropic liquid	No	No
Colesteric LC	No	Yes

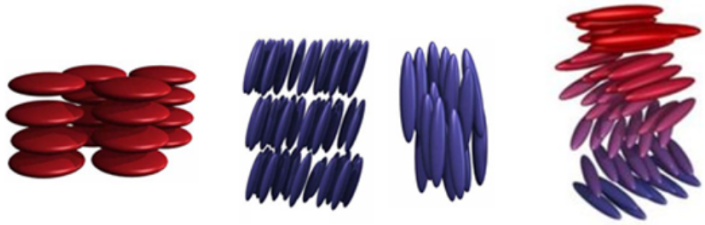

Figure 2.6: Arrangement of molecules in different types of liquid crystals: columnar, smectic, nematic, cholesteric.

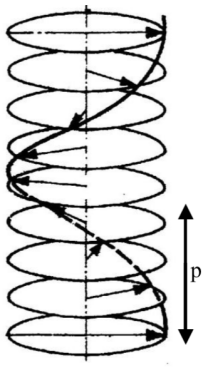

Figure 2.7: Cholesteric phase is made of local nematic "layers" continuously twisted with respect to each other, with periodicity p/2 (where p is the cholesteric pitch).

- Columnar phase: parallel DNA helices align on a 2D lattice (see Fig 2.9), but remain free to slide relative to each other in the orthogonal direction. The nature of the phase and its hexagonal packing symmetry has been established by a combination of polarized microscopy, X-ray diffraction, and freeze fracture electron microscopy experiments. The

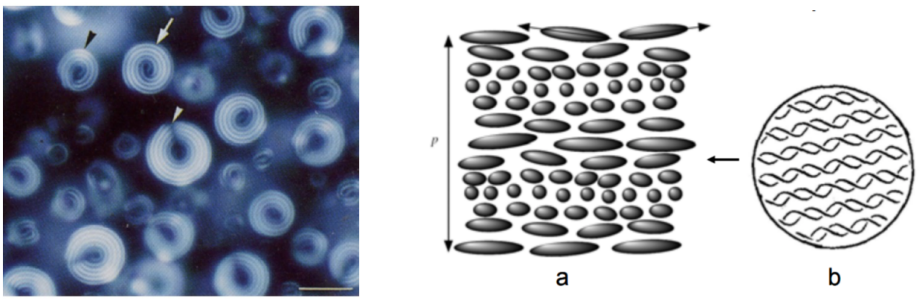

Figure 2.8: On the left: N* droplets observed in polarized microscopy. The bright and dark stripes correspond to p/2 [13]. On the right: schematic representation of the cholesteric organization. Double stranded DNA helices are aligned in parallel (b) and their orientation rotates along the cholestric axis p (a).

columnar phase is typically observed for concentrations higher than 400 mg/mL.

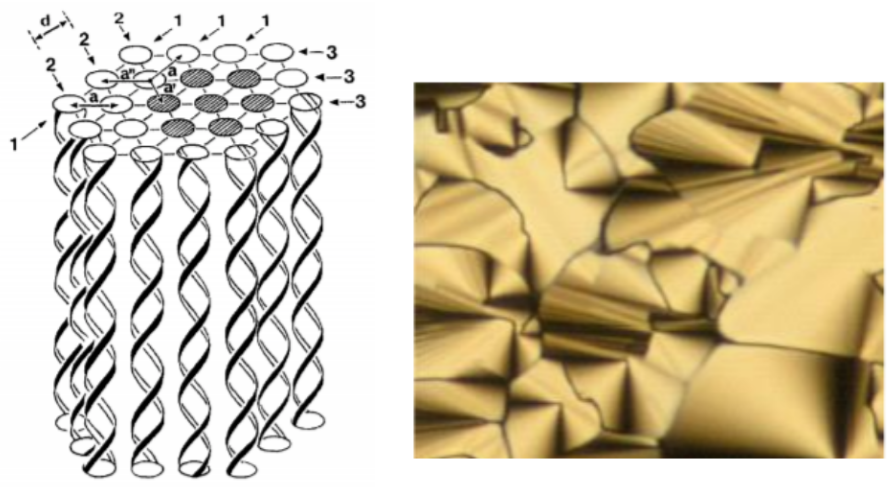

Figure 2.9: Sketch of the structure of the hexagonal columnar phase of DNA, showing parallel helices hexagonally packed in the plane perpendicular to their axis

Experimental evidence was reported for the existence of additional phases

for long DNA chains: a pre-cholesteric order, a hexatic phase that replaces the hexagonal columnar in very long DNA fragment and a structure with orthorhombic symmetry appearing in the transition to crystalline order (see Fig 2.10)

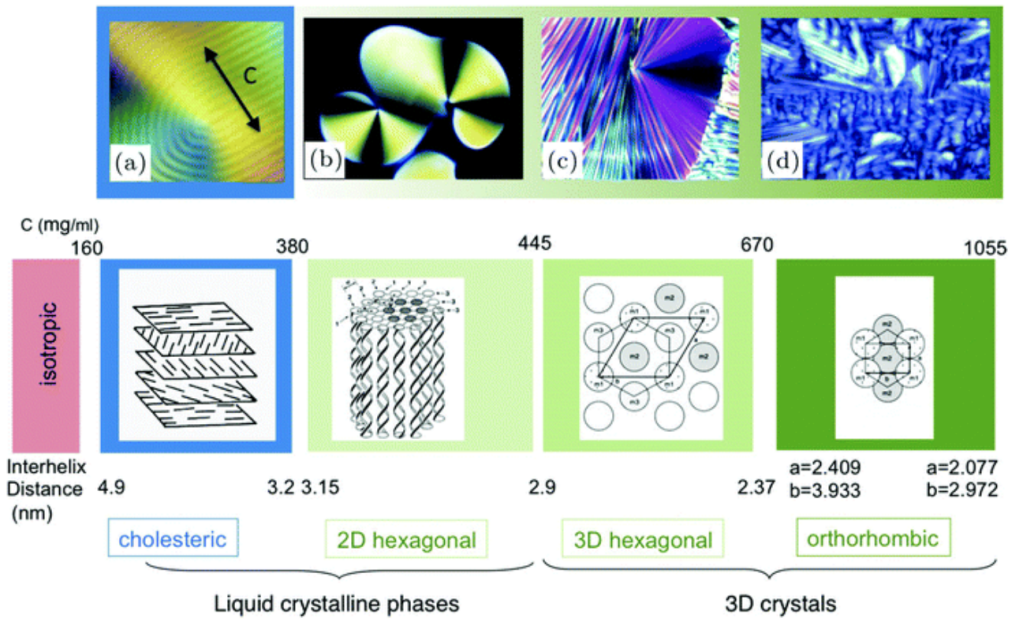

Figure 2.10: The typical sequence of liquid crystalline textures as observed by polarised light microscopy with corresponding structures for an increasing DNA (50 nm fragments-150 bp) concentration: cholesteric phase (a), 2D columnar hexagonal phase (b) and 3D hexagonal and orthorhombic phases (c, d). The structure of each phase is schematized below, using the same color code. Concentrations (in mg/ml) and inter helix distances (in nm) are indicated for each phase transition.[Adapted from: [14]]

According to the Onsager theory and to computer simulations of the behavior of hard sphero-cylinders [15], no LC ordering is predicted for rods with aspect ratio L/D < 4, in the absence of additional interactions, thus DNA double helices with a number of base pairs N < 24 have not enough

anisotropy to display mesophase behavior at any concentration. However, rather unexpectedly, LC ordering has recently been found in concentrated aqueous solutions of DNA self-complementary sequences with N as low as 4-6 bp which combine into helices with aspect ratio well below the Onsager limit [16]. The LC phases exhibited by these short DNA are the same as those observed in long DNA, N* and hexagonal COL, although they are found at higher concentration.

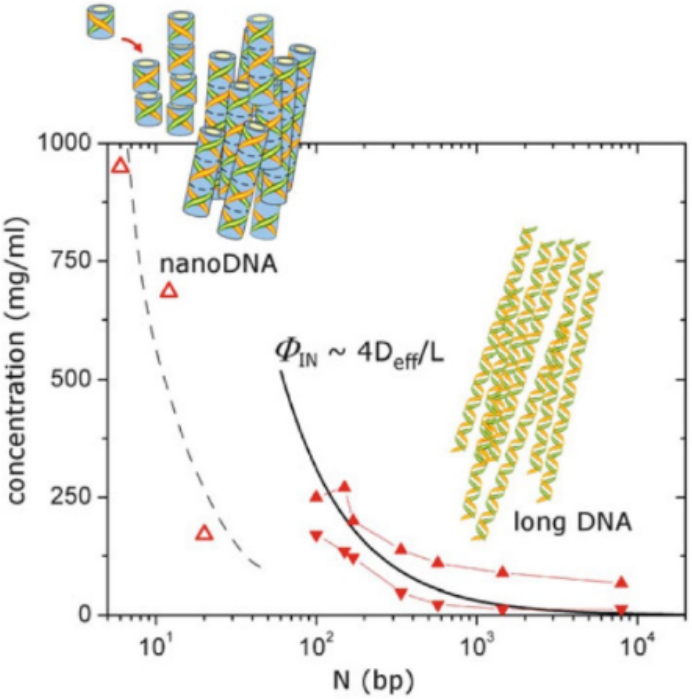

Figure 2.11: Phase behavior of DNA molecules of different lengths. Helices longer than 100 bp (filled triangles) display isotropic-to-nematic transition at concentrations predicted by Onsager (continuous black line). In contrast long DNA and nano DNA data are from [17] and [18], respectively. Image from [19].

The formation of LC phases of short DNA is a consequence of pairing and stacking interactions through a three steps process. First, oligomers hybridize into duplexes, which are stable at T below their melting one; second,

duplexes form linear aggregates whose formation is mediated by either base stacking of terminal bases for blunt-ended duplexes and by pairing of terminal sections, when duplexes terminate with interacting overhangs. Third, linear aggregates align and form a LC phase. At first, no LC ordering was observed for oligomers too short to form stable duplexes, since the LC ordering becomes more stable as oligomer length increases. Moreover, shorter oligomers imply more flexibility of the aggregates, which is a destabilizing factor for LC ordering (see Fig 2.12).

Recently, it has been demonstrate that very short DNA fragments, as small as 4-bp-long, can stack strongly on top of each other and form persistent columns which can show nematic ordering over a large range of temperature [20, 3].

The repulsive side-by-side interaction of blunt-end DNA is mostly governed by electrostatic interactions. It also have been found that adding monovalent salt causes the equilibrium value of nematic parameter to decrease. The effect may be due to the screening of electrostatic repulsion between DNA fragments, leading to an effectively less crowded enviroment resulting from a reduction in the effective excluded volume of DNA columns [3]. Moreover it has been discovered that even temperature can affect LC ordering, but only weakly [20].

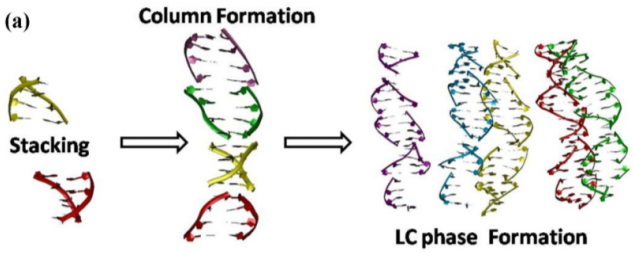

Figure 2.12: Mechanism of LC phase formation in aggregates of ultrashort ds-DNA.

Chapter 3

DNA computational models

3.1 State of the art

Although DNA may seem a relatively simple polymer, it is still difficult to understand some of the fundamental aspects of DNA physics. In the view of this complexity, computer modeling has always played an important role in bridging the gap between the bulk experiments and molecular-level understandings. At the most detailed level, atomistic simulations using force fields such as AMBER o CHARMM offer an intimate representation of DNA. Unfortunately, the number of degrees of freedom (including those of the solvating H_2O molecules) prohibits the simulation of large systems for long periods of time. For example, simulations of double helices (on the scale of 10-20 base pairs) have only recently been extended to time scales of ~ 1 μs. To gain further insight into hybridization, coarse-grained models, which represent DNA through a reduced set of degrees of freedom with effective interactions, are required. In particular, models whose coarse-grained scale is approximately that of the nucleotide may provide the ideal compromise between resolution and computational speed for simulating assembly transitions. The

simplest available coarse-grained are statistical, neglecting structural and dynamical details (Fig 3.1). These models use sequence-dependence parameters that describe the free energy gain per base pair relative to a denatured state, with extra parameters used for initialization of duplex regions and to describe sections whithin the structure.

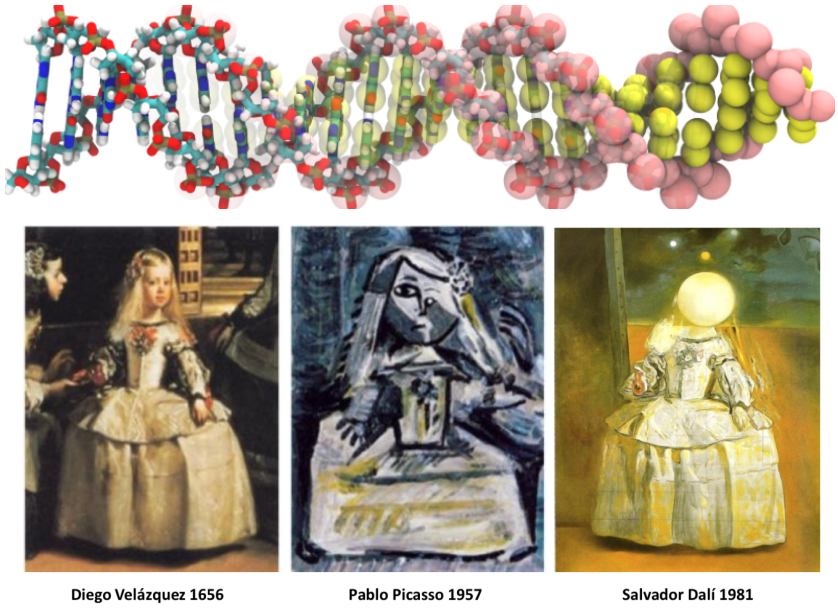

Figure 3.1: From all-atom to coarse-grained simulations illustrated by "Las Meninas" [21].

Molecular Dynamics (MD) is a widely used computational biophysics method which efficiently provides insight into many biochemical processes [22]. MD computer simulations follows the time evolution of a set of interacting particles (generally atoms or molecules) step by step by integrating their equations of motion. MD is therefore a deterministic technique if we use a reversible dynamic algorithm: given an initial set of positions and velocities, the subsequent time evolution is completely determined and in principle reversible. By taking a series of snapshots of the system at different times

it is possible to follow the particles motion from stationary observables and to compute mean thermodynamic properties as time averages. A minimal description of this computer simulation technique is given in Appendix A.
In order to keep a high computational efficiency and yet an accurate description of biomolecular stucture and dynamics, many philosophies on this topic have been developed. It is useful to classify them through the way microscopic informations are coarsened: both top-down and bottom-up approaches are possible. In the bottom-up approach, effective CG interactions are extracted in a systematic and consistent way from reference atomistic simulations. Under the statistical mechanics framework, the many-body Potential of Mean Force (PMF) for any specific coarse-grained system is completely specified by the underlying atomistic model and the chosen coarse-grained mapping. In practice, pair PMFs are used, and the parameters are determined iteratively. Finally, most of the potentials derived with a pure bottom-up approach are fine-tuned *a posteriori*, to reproduce experiments or phenomenological properties of a particular system. On the contrary, the top-down derivation requires that the set of interactions should be empirically parameterized, in a trial-and-error manner, to match experimentally determined thermodynamic properties (i.e. melting temperatures) or structural and mechanical features of double-stranded and single-stranded DNA. Simple equations, usually the same found in atomistic force-fields, are adjusted on the basis of physicochemical intuition to reproduce structural or thermodynamics properties [23]. In the following sections some of the most common coarse-graining strategies for DNA are described.

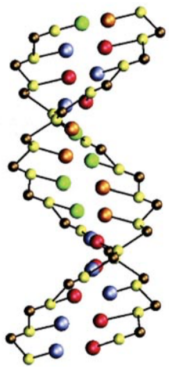

Figure 3.2: Schematic rapresentation of the mesoscale model of DNA by De Pablo [23]

3.1.1 De Pablo coarse-grained model

De Pablo and coworkers [24] have developed a CG representation of DNA, with the purpose of reproducing the thermodynamics of melting, bubble formation, hybridization and salt dependence of the persistence length. This model gives a description of DNA chains with a special biasing potential that penalizes large-scale structural deviations from the reference crystallographic B-form of DNA (see Fig 3.2). The corresponding potential is defined by the following equation:

$$V_{total} = V_{bond} + V_{angle} + V_{dihed} + V_{stack} + V_{bp} + V_{excl} + V_{qq} + V_{solv}. \qquad (3.1)$$

The first three terms represent bonds, angles, and dihedral angles, with a treatment that is commonly used in many molecular mechanics force fields. All the equilibrium values were taken from the crystallographic structures of the B-DNA. The V_{stack} term is responsible for intrastrand base stacking interactions and backbone rigidity, while V_{stack} is responsible for base pairing interaction. The cut-off for V_{stack} is set to 9 Å which extends the range of non-

bonded interactions up to the next neighbor bases. The V_{excl} term accounts for excluded volume interactions. The electrostatic interactions are included in the V_{qq} term, based on a simplified Debye-Hückel (D-H) approximation for the ionic environment, with a temperature and concentration dependent dielectric constant. The V_{solv} term is introduced to reproduce the solvation by water layers, which plays a crucial role in the thermodynamics of the melting transition. Thanks to this coarse-grained model, the authors were able to reproduce a number of thermal properties such as the melting curve and the temperature dependence of heat capacity. However its usage is restricted to systems where the structure of DNA may not deviate from canonical B form. [21].

3.1.2 Plotkin coarse-grained model

Plotkin and coworkers [21] developed a coarse-grained DNA model that stands out from other by the usage of non-isotropic potentials which are more accurate description of the geometry of the nucleotide bases. Their coarse-grained Hamiltonian contains interactions, which are mainly parameterized by relying on corresponding all-atom (AA) MD simulations, with the aim to reproduce the key mechanical properties of DNA. Each nucleotide is represented by three beads: two spherical ones representing the sugar and phosphate groups and one rigid ellipsoid that captures the 'pancake' shape of nucleotide bases (Fig 3.3). Despite the fact that anisotropic potentials are typically more computationally time consuming, they allow straightforward calibration of stacking interactions and by varying their strength one can gain insight into how they influence various structural and mechanical properties. In addition, when using only isotropic potentials, it is difficult to account for structural information about base tilting, twisting and proper

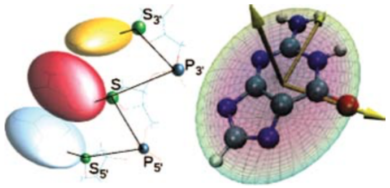

Figure 3.3: On the left the schematic representation of three nucleotides, with ellipsoidal beads corresponding to bases and the beads labeled as S and P to sugars and phosphate groups, respectively. On the right the all atom framework of nucleic base, showing the principal axes of the ellipsoid, which uniquely determine the orientation of the base.

stacking. Having more explicit control over these interactions should allow for finer regulation of DNA chain geometry and local dynamics. Also, as it is known that hydrogen bonds are directional, the non-isotropic potentials in principle do a much better job in representing that aspect of the chemical reality.

The configurational part of the Hamiltonian for the model of Plotkin and coworkers takes the following form:

$$\begin{aligned} V_{total} = V_{bond}(r) + V_{angle}(\theta) + V_{dihed}(\phi) + V_{RE^2}(B_1, B_2) + \\ V_{RE^2}(B_1, res_2) + V_{LJ}(r_{ss}, r_{sp}) + V_C(r_{pp}). \end{aligned} \tag{3.2}$$

The first three expressions describe the bond, angle, and dihedral angle fluctuations; however, no explicit forms are assumed for any of them. Instead the functional forms, along with parameters, are determined by either fitting the AA potentials of mean force (PMF) plots to some either custom or common functional forms. The Gay–Berne style potential [25], proposed by Babadi *et al.* was used to describe the non-isotropic component of potentials, which are denoted with RE^2. The base sugar interactions are

accounted via V_{RE^2} (B_1, res_2) potential which is simply the limiting case of the base–base potential V_{RE^2} (B_1, B_2) when one ellipsoid is replaced by a sphere. Sugar–sugar and sugar– phosphate interactions are modeled by the LJ potential V_{LJ} which only acts between residues that are more than two neighbors apart. Ions are treated in an implicit manner, hence the electrostatic potential is the same as in the de Pablo's model, except that the dielectric constant is here temperature independent [21].

3.2 OxDNA model

Among the coarse-grained model we can find the oxDNA model, developed by Thomas Ouldridge *et al.* [1] with the aim to investigate the thermodynamics of transitions involving ssDNA and dsDNA (in the most common B-form) into a 3-dimensional, dynamical coarse-grained representation. Like in De Pablo's previously described model, this ambition is realized with a top-down approach, since it does not primarily concern with the chemical details of interactions, but rather with their net effect on DNA properties. Thermodynamically, the most important interactions to represent are the stacking of single strands, the formation of single-stranded hairpins and the hybridization of two separate strands to form duplexes. The model takes into account a reasonable representation of the mechanical properties of DNA: single strands should be flexible, and duplexes comparatively stiff [26].
The model consists of rigid nucleotides with four interaction sites, as shown in Fig 3.4. The four interactions sites lie in a line, with the base stacking and hydrogen-bonding/base excluded volume sites separated from the backbone excluded volume site by 6.3 Å and 6.8 Å respectively.

The model is specifically designed to allow an approximate representa-

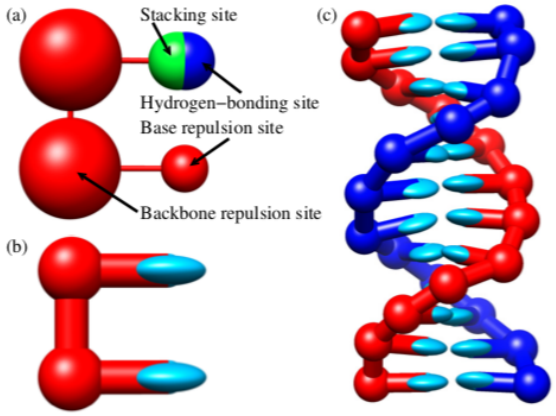

Figure 3.4: Model interaction sites: the stacking/hydrogen-bonding sites are shown on one nucleotide and the base excluded volume on the other. The sizes of the spheres correspond to interaction ranges. The ellipsoidal bases allow a representation of the intrinsic planarity of the model.

tion of B-DNA in its double-stranded state. The relative size of equilibrium backbone separation and ideal stacking distance lead to a pitch of 10.34 bp per turn at 296.15 K similar to experimental estimates of 10-10.5 bp. The length-scale is also choosen so that the average rise per bp at room temperature is equal to 3.4 Å , which results in a helix with a radius of 11.5 Å, comparable to the experimental value of 11.5-12 Å.

3.2.1 The averaged base model: oxDNA1

A first model, called oxDNA1 [1], was designed with an "average-base" representation, neglecting the dependence on the sequence of the single or double strand, and is accordingly suited to study processes for which sequence heterogeneity is not important. In oxDNA1, nucleotides interact in a pairwise fashion with other nucleotides in the system. Interactions between nearest neighbor (NN) on a strand are distinct from all others, allowing for strands

connectivity and stacking. The potential is written as a sum over NN pairs and a sum over all others:

$$V = \sum\nolimits_{NN} \left(V_{backbone} + V_{stack} + V_{exc}^{'} \right) + \sum\nolimits_{otherpairs} \left(V_{HB} + V_{c_{stack}} + V_{exc} \right) \tag{3.3}$$

$V_{backbone}$ is a finitely extensible nonlinear elastic (FENE) spring, with an equilibrium of length of 6.4 $\mathring{A}$ representing the covalent bonds which hold nucleotides in a strand together, whose equation is:

$$V_{FENE} = -\frac{\epsilon}{2} \ln \left(1 - \frac{(r - r_0^2)^2}{\Delta^2} \right). \tag{3.4}$$

V_{stack} represents the tendency of bases to form coplanar stacks: it is a smoothly cut-off Morse potential between base-stacking sites, with a minimum at 3.4 Å. It is modulated by angular terms which favor the alignment of normal vectors, and the alignment of the normal vectors with the vector between stacking sites. As such, the interaction encourages coplanar stacks, separated by a shorter distance than the equilibrium backbone length, leading to helical structures. Right-handed helices are imposed through an additional modulating factor which reduces the interaction to zero for increasing amounts of left-handed twist. $V_{exc}^{'}$ represents the excluded volume of nucleotides preventing the crossing chains and provides stiffness to unstacked single strands.

For V_{exc}, attractive-repulsive Lennard-Jones interactions are included between all sites on the two nucleotides:

$$V_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{3.5}$$

For NN, the backbone/backbone site interaction is not included because the distance between sites is regulated by FENE spring. V_{HB}, represent-

ing the hydrogen bonds which lead to base pairing, is a smoothly cut-off Morse potential between hydrogen-bonding sites, modulated by angular terms which favor the anti-alignment of normal vectors and a colinear alignment of all four backbone and hydrogen bonding sites:

$$V_{Morse} = \epsilon[1 - e^{[-(r-r^0)a]^2}] \tag{3.6}$$

$V_{c_{stack}}$ represents cross-stacking interaction between a base in a base pair and the nearest-neighbor bases on the opposite stand, providing additional stabilization of the duplex. This term has been incorporated through smoothed, cut-off quadratic walls modulated by the alignment of base normals and backbone-base vector with separation vector in such a way that its minimum is approximately consistent with the structure of model duplexes. Although the model incorporates sequence specifity (only A-T and G-C hydrogen bonding are possible), there is no sequence dependence in the potential for either stacking, cross-stacking, hydrogen-bonding, or excluded volume. Unfortunately, most of biological processes involve sequence heterogeneity. An example can be found in DNA itself since A-T and G-C have different relative binding strength, with the latter being significatively stronger because of the presence of three rather than two interbase hydrogen bonds. Moreover, the stacking interactions that drive the coplanar alignment of neighboring bases are known to show significantly different behaviour depending on the sequence. In addition, no explicit electrostatic interaction is present in the model, which may be expected to be important as bare ssDNA has a net charge of -e. For this reason the model could be used to fit experimental data only at high salt concentration [Na^+]= 50 0mM at which the electrostatic properties are strongly screened. Indeed, at these ionic concentrations, the Debye screening length, λ_D is approximately 4.3 Å , smaller

than the excluded volume diameter for backbone-backbone interaction (~ 6 Å) [1]. These drawbacks led to the development of a new model, oxDNA2, able to take into account sequence-dependence interactions and salt dependence, which was actually used to run all the simulations described in the following chapters.

3.2.2 The new model: oxDNA2

In order to go beyond the average sequence parametrization by introducing sequence-dependent interaction strengths into the model, it has been proposed a new potential:

$$\begin{aligned} V_0 = & \sum\nolimits_{nn} \left(V_{backbone} + V_{stack} + V_{exc}^{'} \right) + \\ & \sum\nolimits_{otherpairs} \left(V_{HB} + V_{cr.stack} + V_{exc} + V_{cx.stack} + V_{DH} \right) \end{aligned} \tag{3.7}$$

The interactions between nucleotides are described by the hydrogen bonding (V_{HB}), cross-stacking ($V_{cr.stack}$) , coaxial stacking ($V_{cx.stack}$), and stacking interactions (V_{stack}) explicitly depending on the relative orientations of the nucleotides as well as on the distance between interaction sites. The backbone potential $V_{backbone}$ is again an isotropic spring that imposes a finite maximum distance between neighbors, mimicking the covalent bonds along the strands. The coaxial stacking term, not shown in Fig 3.5, is designed to capture stacking interactions between non-neighboring bases, usually on different strands. All interaction sites also have isotropic excluded volume interactions V_{exc} and $V_{exc}^{'}$.

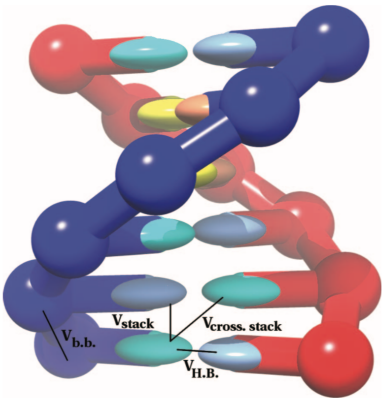

Figure 3.5: The figure shows schematically the interactions between nucleotides in the coarse-grained DNA model for two strands in a duplex. All nucleotides also interact with repulsive excluded volume interactions. The coaxial stacking interaction is not shown.

The introduction of sequence-dependece for stacking interaction

The model was originally parametrized to reproduce the melting temperature of average sequences as predicted by the SantaLucia (SL) model. To introduce a dependence on sequence, the authors added two extra term to the potential V_0: V_{stack} and V_{HB}. The new parametrization V_{stack} has provided the faculty to discern between different bases of the sequence: for example the V_{stack} term has a larger value for GC base pair than for AT, which is characterized by a smaller one. This leads to a stronger interaction between GC which is reflected in a shorter distance. Interestingly, it has also been modified the stacking interaction between complentary bases: in particular, it has been reported that the GC (5'-3') staking interaction is stronger than the CG (3'-5') one. Otherwise, a remarkable innovation constists in the introduction of different values of V_{HB} depending on the base pairing. Since it is known by experimental results that G and C form 3 hydrogen bond-

ing, while A and T only form two, the hydrogen bonds interaction has been modified to represent this aspect of the DNA structure.

The introduction of different widths for major and minor DNA grooves

B-DNA in the original oxDNA model had equal groove widths, while in reality DNA has a larger major groove and a smaller minor one. The presence of major and minor group leads to the necessity of rescaling the position of sites along the backbone, a feature which is characteristic of several DNA motifs. An example is anisotropic duplex bending: the duplex can be expected to bend more easily into the major groove than into the minor, if the groove widths are unequal.

Different groove widths were introduced by changing the position of the backbone site while keeping the duplex radius unchanged (Fig 3.6): the three interaction sites (hydrogen bonding, stacking and backbone site) are placed in a plane, rather than lying on a straight line. The new nucleotide shape introduces an additional parameter into the model, the angle γ between the line from the duplex centre to the backbone site and the line from the duplex centre to the stacking site. Since the sugar-phosphate backbone is composed by 18 different atoms, but in the model it is representing one single interaction, there is no unique decision for assigning the backbone site and the value of the model parameter γ. The authors set $\gamma = 20°$, a value which maps onto a full-atom representation of a DNA duplex by visual inspection, although values of γ between 15° and 25° would give an equally satisfying visual match. For the termodynamics, the introduction of γ provided a change of 1-2 K in duplex melting temperatures, so the hydrogen-bonding and stacking strengths were further modified until the agreement

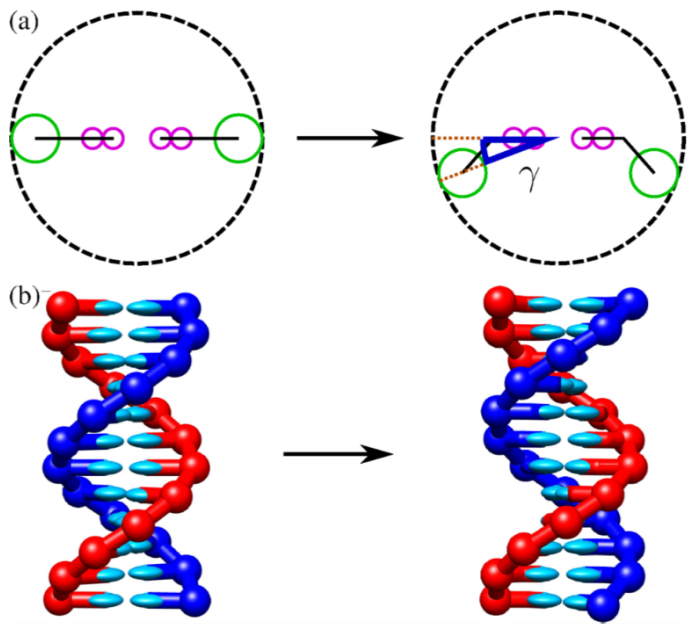

Figure 3.6: Schematic comparison between the original oxDNA model (left) and oxDNA2 (right) (a). In oxDNA all interaction sites are co-linear while in oxDNA2 the backbone interaction site and the stacking and hydrogen-bonding interaction sites are oriented at an angle γ =20°. (b) The non-co-linear arrangement of the interaction sites leads to the formation of the major and minor groove, an important structural feature of DNA. [Adapted from Henrich [27]

with experimental melting temperatures was as good as that of the original model[28].

The introduction of salt dependence

The original oxDNA model was parametrized to [Na^+]=0.5 M, a rather high salt concentration, whose choice comes from the possibility of applying the model to many problems in DNA nanotechnology. However, being able to study a wider range of salt concentration provides a higher interest for this coarse-grained model, since it opens the opportunity to compare computational results with a great number of biological experiments. With this purpose, the authors expandend the usage of oxDNA to lower salt concentration as [Na^+]=0.15 M, typical of biological samples. The treatment of electrostatic for a DNA is not trivial, because of the high negative charge by which the molecule is characterized, caused by phosphate groups localized along the chain. This negatively charged surface may lead to ion condensation, thus a quantitative understanding of these electrostatic effects is a necessary part of any theoretical treatment of biological processes at the molecular level. In the Poisson-Boltzmann (PB) theory [29],[30], which has become a standard tool in biology to describe the electrostatics in biomolecular systems, interactions in dilute (low ion density) electrolytes are found to decay exponentially with distance, with the vharacteristic screening length λ_D, called Debye length. In this framework, in order to reproduce thermodynamic and mechanical effect of salt concentration on DNA developers of oxDNA included in the non-bonded interaction term a Debye-Hückel form V_{DH} :

$$V_{DH}(T, I) = \frac{(q_{eff}e)^2}{4\pi\epsilon_0\epsilon} \frac{\exp\{-r/\lambda_D(T, I)\}}{r} \tag{3.8}$$

where r is the distance between the backbone sites of nucleotides, ϵ_0 is the permittivity of the vacuum, $\epsilon = 86$ is the relative permittivity of water, e is the elementary charge, q_{eff} is the effective charge situated at the backbone site of each nucleotide. The interaction depends on the temperature T and on the (monovalent) salt concentration through the Debye length, λ_D which decreases monotonically with increasing ion concentration, due to effective screening of charges over short distances. In fact the Debye length scales inversely with the square root of bulk ion density, according to:

$$\lambda_D = \sqrt{\frac{\epsilon_0 \epsilon k_B T}{2 N_A e^2 I}} \tag{3.9}$$

where N_A is Avogadro's number and k_B is Boltzmann's constant, and I is the ionic strength of the solution.

It is worth to remember that all the atoms of the sugar and phosphate groups of the backbone are represented by a single interaction site, carrying a charge q_{eff} , that is one of the parameters that have been tuned to reproduce the thermodynamics predicted by the model of SantaLucia [2]. Actually, the use of Debye-Hückel approximation leads to a lower effective charge with respect to the formal charge: this effect is enhanced at lower salt concentration [2]. Notice that in the PB framework, solvated ions are treated only by their valences as point charges that interact with an averaged electrostatic potential. The simplistic treatment eases computation but makes impreticable an accurate treatment of ion behavior. Thus, if we go to higher concentrations, when the Debye length becomes comparable to molecular dimensions, ion size (steric or volume exclusion) effects and ion correlations come into play and the PB theory, which does not include them, is not expected to work properly [31] . In particular, ions of different sizes but equal valences are treated identically in PB theory, but recent experimental work

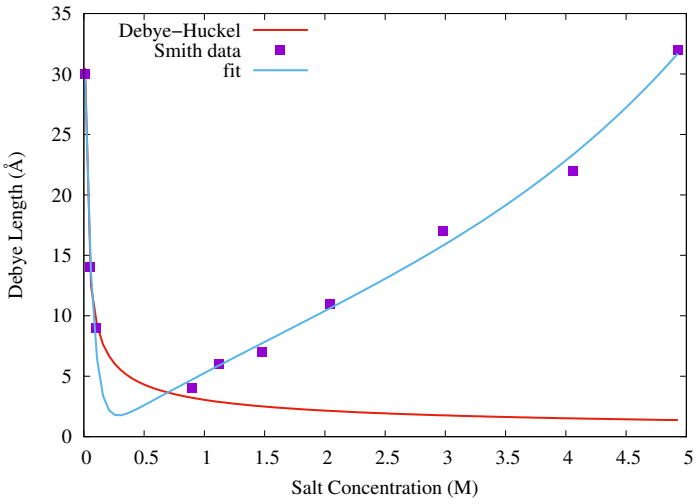

Figure 3.7: Debye length as a function of molar salt concentration (Na^+ in acqueous solution) at T = 300K. The red line represents Debye-Hückel equation, purple dots are the experimental values from Smith *et al* work [35] and the blue line is our curve fitted from their results. Fitting parameters are a = -16.0472, b = 1.59947, c = -0.286956, c = -0.286956, d = 2.13159, e = -0.00761772, f = -0.120095.

has shown significant deviations from it in competitive monovalent and divalent ion binding to a DNA duplex [32]. [33], [34] Since our interest is focused on computing systems where the molar concentration of electrolytes grows above 0.5 M, we developed a new strategy to rapresent DNA electrostatic in acqueous Na^+ solution, starting from experimental results obtained by Smith *et al.* [35]. Their work reports measurements of interaction force between two smooth planar surfaces across electrolytes using a Surface Force Balance (SFB). The authors proved that, contrarily to Debye-Hückel assumption in diluite regime, the decay length does not decrease. It has been shown that when salt concentration exceedes 0.7 M, the Debye length increases with concentration, up to very high values.

On this basis we developed a correction for the Debye length vs molar electrolytes concentration behaviour fitting a curve to the experimental data from [35] (Fig 3.7), in order to apply this new model for DNA interactions in our simulations. The function that best reproduced Smith *et al.* is reported in Equation 2.12 where the parameters *a, b, c, d, e, f* used for the fitting are reported in Fig 3.7.

$$f(x) = \exp(|(ax - b)|) + \exp|(cx - d)| + ex^2 + f \tag{3.10}$$

where x is the molar salt concentration and $f(x)$ has the units of $\text{Å}\text{K}^{-1/2}$.

It is worth to notice that our function poorly describes the DH prediction at intermediate concentration regime (0.1-0.7 M), which is mainly caused by the lack of experimental data. However, we do not pretend to obtain a representing relation for that range, where the Debye-Hückel approximation founds its validation.

All simulations described in the next chapter of this thesis were run using the oxDNA model that has just been described in details.

Chapter 4

Double strand melting transition

4.1 Objectives

The aim of this chapter is the characterization of DNA double strands denaturation. Sequences were designed or selected from the literature [8]. We studied separately the denaturation of five different systems:

- three helices composed of 4 base-pairs: GCCG/CGGC CCGG and CGCG;
- two helices composed of 6 base-pairs: CCGCGG and GACGTC;
- one helix composed of 8 base-pairs : GCTGACG.

Our goal was to discern the difference in melting temperature for double strands, taking into account effects of both sequence strand and length. Short oligonucleotides are described using the oxDNA2 coarse-grained model developed by Ouldridge *et al* (see Fig. 4.1). As previously illustrated in Chapter 3,

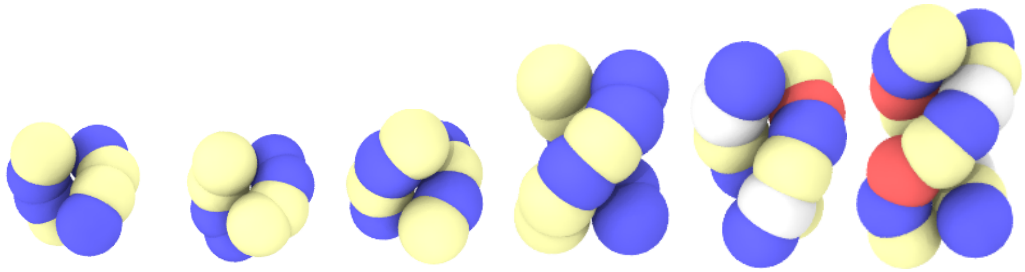

Figure 4.1: Coarse-grained representation of selected double strands for the evaluation of the melting temperature: From left to right: GCCG (with its complementary strand CGGC), CCGG CGCG, CCGCGG and GACGTC. Different coulored beads represents a nucleotide composed of the phosphate group, the deoxyribose sugar and the specific base.

oxDNA2 is widely used to study thermodynamical properties of short DNA strands. We reproduced a computational melting curve for each sequence, which was later fitted with an appropriate function representing the fraction of coupled base-pair. Our results were compared with SantaLucia tabulated termodynamic parameters ΔH, ΔS, and T_m.

4.2 Methods

Simulations were run into an orthogonal box with sides of 230×230×230 box units (i.e. a side length of 1959.14 Å, considering the conversion factor of 8.518 Å[26]), containing only one double strand. The sample volume is much larger than one DNA double strand and corresponds to a concentration of 2.2 x10^{-6} M, even if actually there is only one helix in the box. The only way to introduce real concentration effects would be adding more helices in the same box. In order to obtain a statistically representative result, we executed 100 simulations starting from the same configuration but with different initial velocities for each double strand. At each temperature, two consecutive MD

runs of 2x10^6 or 3x10^6 (depending on the sequence length) were performed, the first one corresponding to the equilibration period and the second one to the production period. We fixed an integration time step of 3.03x10^{-15} s and we saved coordinates every 1000 steps. Periodic boundary conditions were always applied in three dimensions. Simulations were performed in the NVT ensemble using a Bussi thermostat [36] in order to set a constant temperature. Runs at higher temperatures were started from the equilibrated configuration of the lower temperature. We decided to use the same temperature increment of 2 K independently from the sequence length, but we changed the range of temperature to scan according to the expected T_m as predicted from the SL method [8]. The distribution of base-pairing interaction was also recorded, using a code that we developed: we fixed a cutoff value for hydrogen bond energy of -0.1 reduced units, corresponding to -0.60 kcal mol^{-1} [26] and relative to about 1/7 of a typical pair hydrogen bond interaction [1]. When at least one hydrogen bond of the helix was stronger (more negative in energy) than the cutoff value, the duplex was considered completely paired. On the contrary, when each base-pair of the sequence presented a hydrogen bonding interaction energy below the cutoff value, we considered the double strand completely dissociated. We carried out all the simulations with a salt concentration of 1 M in order to precisely compare our values with SL results [8], that were obtained in the same conditions. Moreover, since this electrolytic concentration is in the range of validity for the Debye-Hückel approximation, we decided to use the original treatment of electrostatics for oxDNA2 model (see Chapter 3).

4.3 Results

4.3.1 Thermodynamic parametrization and comparison

We plotted the base-pairing interaction (averaged for 100 simulations at constant T) as a function of temperature. Among the five double strands sequence that we studied, GCCG has no comparative SL values. However, our interest for this strand sequence is linked to the study of its aggregation at high DNA concentrations, for which further details are given in Chapter 5. The temperature at which our curves of the double strand fraction f:

$$f = \frac{[dsDNA]}{[dsDNA][ssDNA]/2} \tag{4.1}$$

assume the value of 0.5 was considered to be the transition point, i.e. when half of the helices are in the single strand form. We extrapolated from our computational curve the thermodynamic parameters ΔH and we compared our results with the ones of SantaLucia method. We fitted the temperature dependence of the fraction of double strands that are actually present in the system:

$$f(T) = \frac{1 + [DNA]\,K_{eq} - \sqrt{1 + 2\,[DNA]\,K_{eq}}}{[DNA]\,K_{eq}} \tag{4.2}$$

where [DNA] is the total DNA molar concentration $[DNA] = [dsDNA] + [ssDNA]/2$ (in order to be coherent with the tabulated value of SantaLucia's article, we choose a value of 1 $\times 10^{-4}$ M) and K_{eq} is the equilibium constant, defined as follows:

$$K_{eq} = e^{\left[\frac{\Delta S}{R} - \frac{\Delta H}{RT}\right]} \tag{4.3}$$

where R is the gas constant (1.987 cal/K).

Our first attempt to fit the melting curve transition from the crude formulation of $f(T)$ in Equation 3.1 gave poor results because of its intrinsic complexity and non-linearity of the parameters. To facilitate the procedure we decided plot the equilibrium constant as a function of $f(T)$ as :

$$K_{eq} = \frac{1}{[DNA]f(T)} \frac{1}{1 - f(T)^2} \tag{4.4}$$

then we calculated $\ln K_{eq}$ and we fitted the slope of $\ln K_{eq}$ as a function of the lone parameter ΔH, which is independent from the DNA total strand concentration [DNA]. We successfully extrapolated ΔH and we determined the melting temperature for each oligonucleotide sequence. The calculation of ΔS comes from equation [8] which relates the T_m of the duplex structure with ΔS, ΔH and the DNA total concentration:

$$T_m = \frac{\Delta H}{\Delta S - R \ln[DNA]} \tag{4.5}$$

This method essentially consists in a Vant'Hoff analysis of our data, which assumes that the transition equlibrium involves only two states: the duplex and the single strand form. Table 4.1 highlights the calculated thermodynamic properties which are compared to SantaLucia's values [8], while Table 4.2 outlines the small difference between the melting points from our simulations and SantaLucia ones. Final results from our plot of Equation 4.2 using the parameters in Table 4.1 are reported in Fig. 4.4 and Fig. 4.3.

Both ΔH and ΔS are given as negative numbers, i.e. for the exothermic association of the two single strands to form a duplex. Usually this convention is used to describe helix formation, however, our studies are conducted in a way to simulate the opposite reaction, i.e. helix denaturation. We discovered that ΔH and ΔS data from the fitting procedure are actually overestimating the SantaLucia reference values, and that the difference between the two set

Sequence	ΔH^{sim} (kcal/mol)	ΔH^{SL} (kcal/mol)
CGTCGACG	-80.8 ± 2.7	-62.9
GACGTC	-87.2 ± 2.9	-42.7
CCGCGG	-78.9 ± 2.91	-44.7
GCCG	-75.1 ± 4.0	/
CGCG	-52.2 ± 2.03	-31.3
CCGG	-58.1 ± 1.07	-23.5

Sequence	ΔS^{sim} (cal/mol)	ΔS^{SL} (cal/mol)
CGTCGACG	-258.0 ± 8.6	-170.4
GACGTC	-297.8 ± 10.0	-119.4
CCGCGG	-256.8 ± 9.4	-44.7
GCCG	-276.2 ± 14.3	/
CGCG	-196.3 ± 7.6	-86.7
CCGG	-220.0 ± 4.1	-64.0

Table 4.1: ΔH^{sim} and ΔS^{sim} values calculated from the simulations and tabulated ΔH^{SL} and ΔS^{SL} [8]. For GCCG sequence there are no comparative values. However further analysis have been carried out for this specific double strand. More details are given in Chapter 5.

of results is approximately constant for each sequence for which we choose to reproduce the denaturation. However, both ΔH and ΔG values are of the same order of magnitude. We think that disagreements between our results and SL probably arises from the choice of the cut-off value of hydrogen bonds (we referred to the value of -0.6 kcal/mol since it is the one reported in Ouldridge work for which oxDNA model was developed). It is worth to notice that even our method to discern from coupled and uncoupled double strand may affect the quality of thermodynamic parameters. It is likely that choosing different cut-off and method to determine the number of base pair interaction along the strand, would provide a more accurate definition of both enthalpies and entropies. Future analyses will be attempted to fulfill this goal.

4.3.2 Comparison of melting temperatures

Sequence	T_m^{sim} (K)	T_m^{SL} (K)
CGTCGACG	337	333
CCGCGG	331	328
GACGTC	312	307
GCCG	291	/
CGCG	293	288
CCGG	288	285

Table 4.2: Melting temperatures are obtained from simulations and values from SantaLucia's work [8]. Again, for sequence GCCG there are no values to compare with.

As reported in Table 4.2, we notice a very good agreement between our computed T_m and SantaLucia set of data. As for thermodynamical parameters, our model seems to overestimate the transition point, but the shift is small, 5 K for the sequence GACGTC, 4 K for CCGCGG and CGTCGAC, while for 4 base paired helices is limited to 3 K.

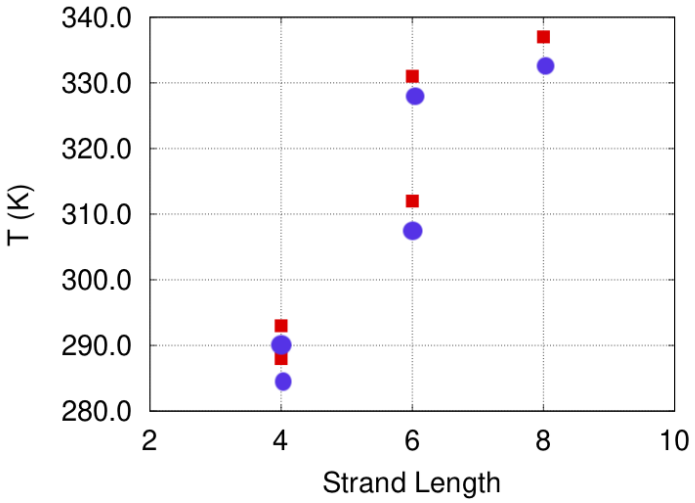

Figure 4.2: Melting temperature versus duplex length as given by SL (circles)[8] and by our simulations (red squares).

We confirmed with our simulations that longer double strands have a higher melting temperature than the ones formed by shorter oligonucleotides. The trend is attributable to the thermodynamics parameters shown in Table 4.1 : enthalpies have a higher module as the number of bases componing the strand increases, i.e. the double strand structure becomes more stable that the single strand form. T_m is affected by complementary bases of the sequence: it is well known, as also reported in Section 1 of Chapter 1, that complementary bases, due to their chemical nature, differ in strength of base-pairing. While G-C forms three hydrogen bonds, A-T is characterized by two. This intrinsic property of DNA is responsible for the higher T required for

denaturation of sequences that lack in A-T complementary bases. This trend is easy to find comparing 6 base paired sequences. For CCGCGG, which is only composed by GC complementary bases, we registered a $T_m = 331$ K. GACGTC sequence resulted in a $T_m = 312$ K for a decrease of 19 K. It is clear that the presence of a single A-T complementary base lowers the melting temperature. Comparing T_m of the 4 base-paired sequences GCCG, CCGG and CGCG, we noticed that even if they have the same content of GC their melting transition shows small differences. The reason should be found in the Nearest-Neighbour method on which are founded our computer simulations and SantaLucia ones. In addition, we point out that even base order is determinant in DNA duplex stability; in fact we found $T_m^{CGCG} > T_m^{GCCG} > T_m^{CCGG}$, that is consistent with the trend in nearest-neighbour stabilities are GC > CG > GG reported in Table 2.4 of Chapter 2.

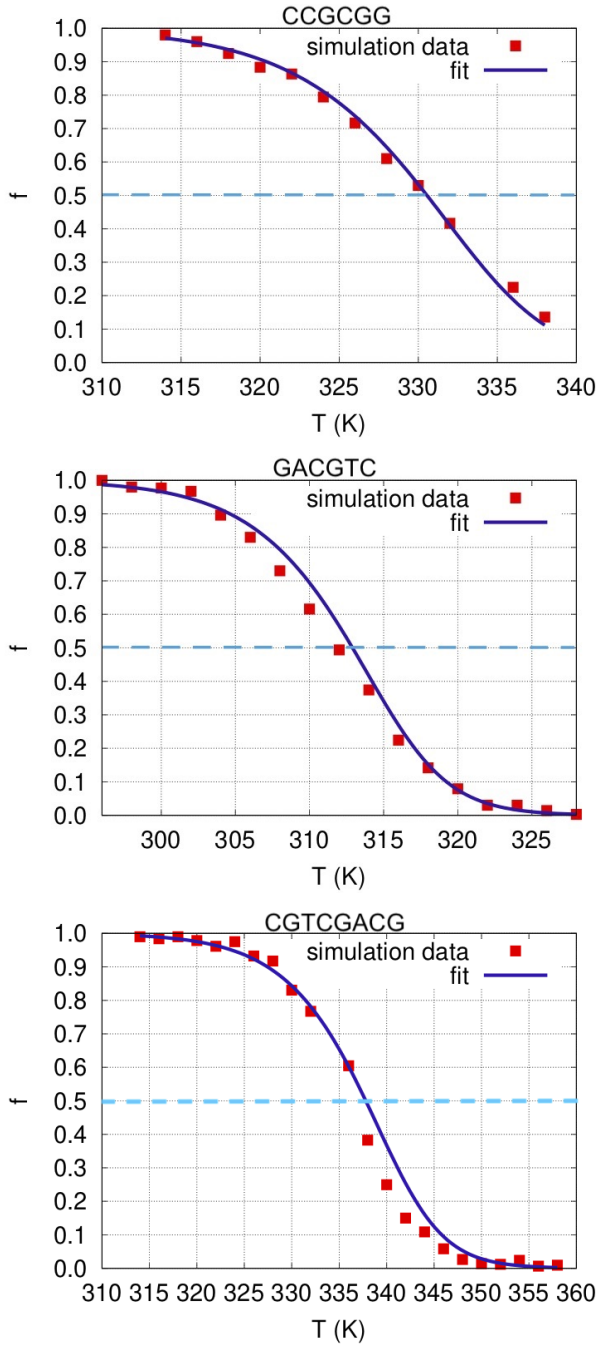

Figure 4.3: Fraction f of undissociated double strands as function of temperature. Red dots represent sequence-pairing fraction resulting from our simulation. The blue line is the function reported in Equation 3.1, from which we extrapolated ΔH. The dotted horizontal line f(T)= 0.5 was employed to determine the melting transition temperature. The sequences analyzed are CCGCGG (top), GACGTC (middle) and CGTCGACG (bottom).

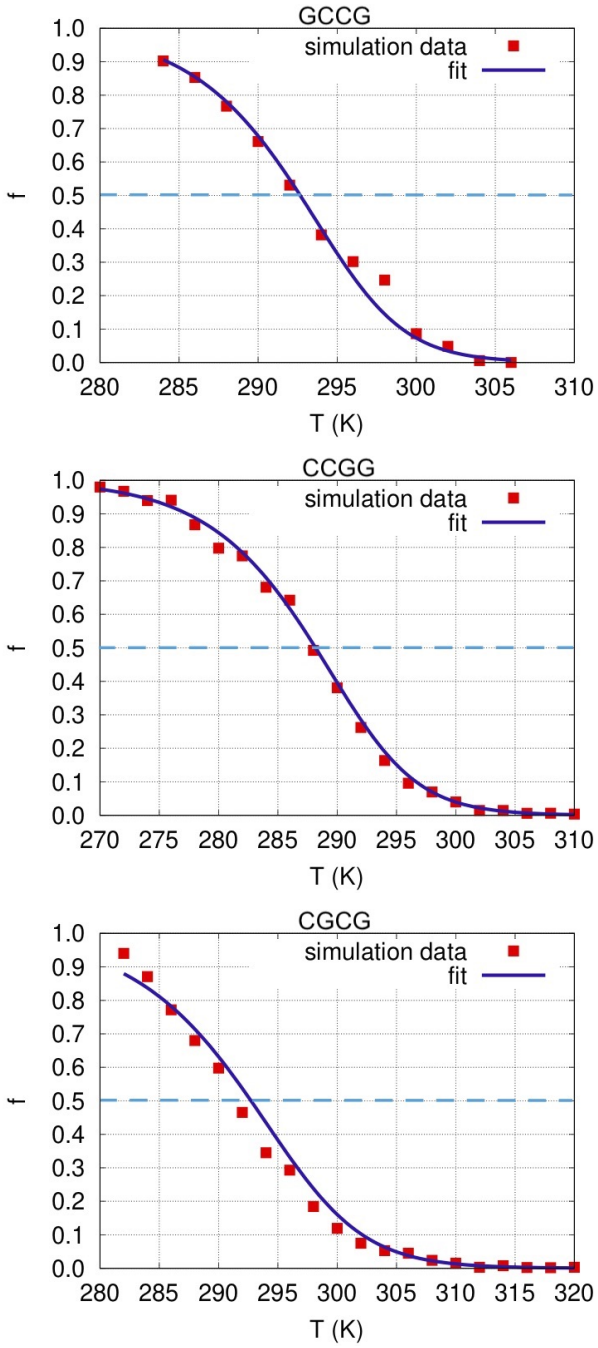

Figure 4.4: Fraction f of undissociated double strands as function of temperature. Red dots represent base-pairing value resulting from our simulation. The blue line is the function reported in Equation 3.1 from which we extrapolated ΔH and ΔH. The dotted horizontal line highlights the importance f(T)= 0.5 to determine the melting transition. The sequences reported are GCCG (top), CCGG (middle) and CGCG (bottom).

4.4 Summary

Our results confirmed that T_m for short oligonucleotide sequences is mainly affected by three factors:

- number of bases componing the double strand, i.e. its length;
- base pairing, i.e. the content of C-G;
- stacking interaction, i.e. the combinantion of bases along the strand;

Essentially, the good agreement of our MD melting temperature with SantaLucia ones is a proof of validation of the oxDNA coarse-grained model.

However, we mainly conducted MD simulations in order to evaluate the accuracy of the intermolecular potential which is defined in the oxDNA2 model. The agreements we found out with SL reported melting temperature reassure us on the validation of the model we choose to represent oligonucleotides interaction. Consequently, in Chapter 5, we are going to simulate the self-assembly for a GCCG single strand.

Chapter 5

LC short strand ordering

5.1 Phase diagram reproduction

Supramolecular ordering of oligonucleotides constitutes a field of investigation constantly enriched with new findings, in which interest is still growing. Several examples have been described in literature such as sequence-directed self-assembly [37], base stacking driven structures [38] and various forms of LC longe-range ordering [16]. This chapter describes our attempt to reproduce the experimental phase diagram (see Fig 5.1) obtained from a solution of ultrashort DNA oligomers, as published by Bellini *et al* in reference [3].

Starting from a water solution 4 bases single strand the authors described the mechanism of LC ordering (see Fig 5.2). The work considers only two sequences (ATTA and GCCG), but we focused on the single strand GCCG because its phase diagram shows both nematic chiral (N*) and columnar phases (COL), while for sequence ATTA no N* it is found. In order to reproduce experimental results, we run MD simulations at constant temperature for each concentration DNA.

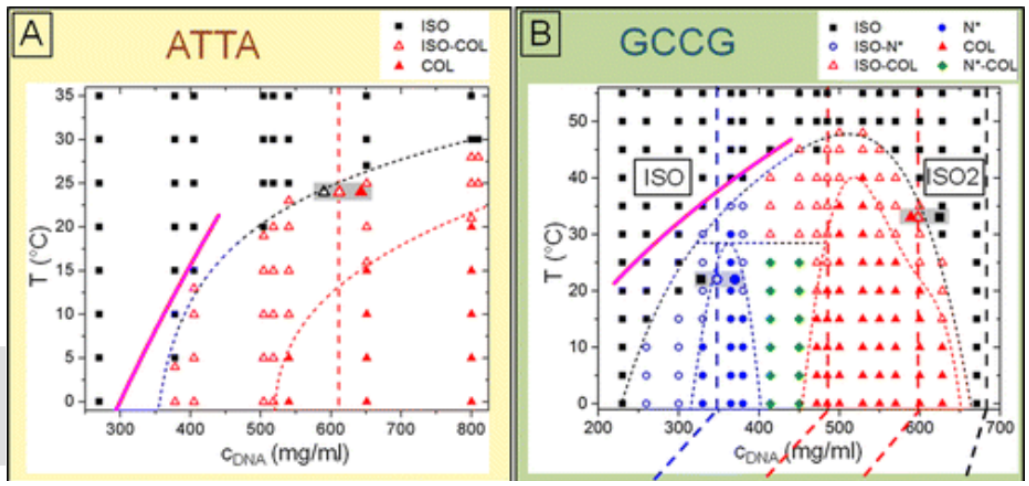

Figure 5.1: Experimental phase diagram for ultrashort sequences ATTA (left) e GCCG (right) [Adapted from [3]].

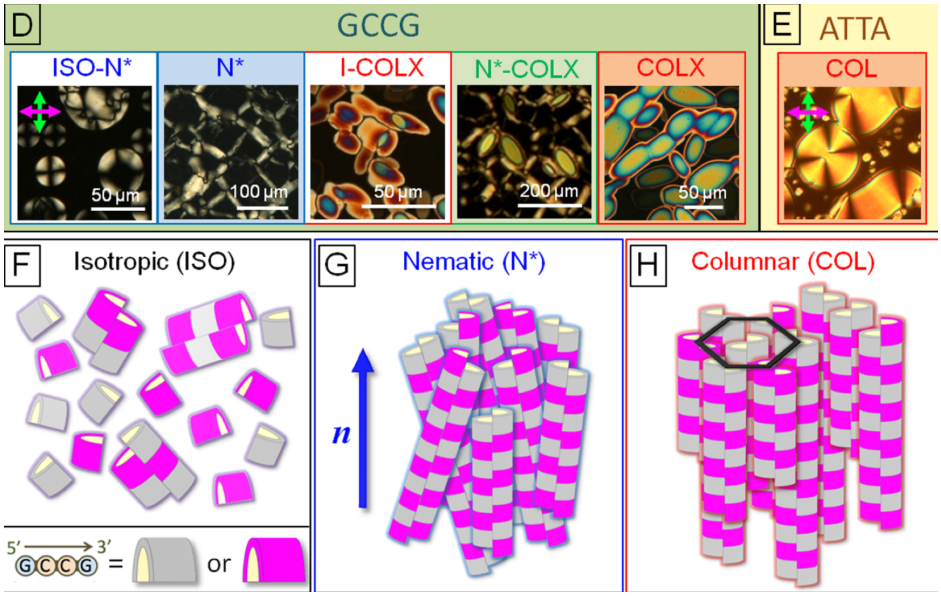

Figure 5.2: Textures of LC GCCG and ATTA visualized through Polarized Optical Microscopy (top). Schematic representation of isotropic (ISO), chiral nematic (N*) and columnar (COL) phases (bottom) [Adapted from [3]].

5.2 Methods

We performed a simulation in the NVT ensemble using a Bussi thermostat [36], at a temperature T = 5° C while varying the DNA concentration. First, simulation at lowest concentration of 200mg/mL were run. Then, the concentration value was gradually increased until 550 mg/mL. Simulations were conducted into an orthogonal sample box with periodic boundary conditions always applied in all three dimensions. Side lengths have been modified according to differents GCCG concentrations, while salt concentrations were calculated in order to balance the negative charge of the phosphate group from the DNA (see Table 5.1), the number of oligonucleotides was kept constant.

The sample was constituited by 512 GCCG single strands for a total of 1728 sites using the oxDNA2 representation. MD simulation of $5\text{x}10^8$ steps was performed for the first concentration of 200 mg/mL, because we started from a 10 times larger sample box which we reduced. For concentrations in the range 230-550 mg/mL, shorter simulations were required ($1\text{x}10^8$ steps). The integration time step was fixed to $3.03\text{x}10^{-15}$ s and we saved coordinates every 10000 steps. We calculated the average order parameter $\langle P_2 \rangle$ for each DNA concentration (see Table 5.1) and we analysed the positional order through radial correlation function *g(r)* and $g(r, \cos\beta_r)$. Further details for these analyses are reported in the following section.

DNA conc (mg/mL)	Side Length (Å)	Salt conc (M)
200	178.88	0.60
230	170	0.69
260	164	0.78
300	156	0.90
330	151	0.99
380	144	1.14
414	141	1.24
450	137	1.35
472	135	1.42
550	128	1.66

Table 5.1: Different concentrations have been reproduced changing the side of the box of our simulations. Here are reported the different sizes in box units and in Å, according to the conversion factor (1 box unit = 8.518 Å).

5.3 Results

5.3.1 Orientational parameter

The determination of the orientational degree of alignment has a central role in the characterization of mesophases. However, when no external stimuli are present (e.g. an electric or a magnetic field) it is necessary to find a director orientation $\mathbf{n}$ to compute the order parameter. $\mathbf{n}$ is generally different from the box reference frame. The uniaxial order parameter $\langle P_2 \rangle$ can be calculated through the second rank order tensor $\mathbf{Q}$. This tensor is obtained by averaging

over all the N molecules. For simplicity, assuming them to be uniaxial, the Krönecker product of the unit vectors $\mathbf{u}^i$ defining the molecular axes is:

$$\langle \mathbf{Q} \rangle = \frac{1}{N} \sum_{i=1}^{N} \mathbf{u}^i \otimes \mathbf{u}^i \tag{5.1}$$

Any element of matrix $\langle \mathbf{Q} \rangle$ is defined as:

$$[\langle Q_{ab} \rangle] = \langle u_a u_b \rangle \tag{5.2}$$

The order tensor $\langle \mathbf{Q} \rangle$ can be diagonalized in the canonical form:

$$\langle \mathbf{Q} \rangle = \mathbf{U} \langle \mathbf{q} \rangle \mathbf{U}^T \tag{5.3}$$

Where U is the unitary matrix rotating from the laboratory frame to the director frame whose columns are eigenvectores of $\langle \mathbf{Q} \rangle$. The elements of the diagonal eigenvalues $\langle \mathbf{q} \rangle$ are the squared averages of the director cosines that represent the orientation θ_i of each molecule with respect to the mesophase directors $\mathbf{n^x}, \mathbf{n^y}, \mathbf{n^z}$, i.e. the eigenvectors of $\langle \mathbf{Q} \rangle$.

$$[\langle q_{ab} \rangle] = \delta_{ab} \langle \mathbf{u} \cdot \mathbf{n}^a \rangle \tag{5.4}$$

where δ_{ab} is the Krönecker delta. The director is defined as the direction of maximum molecular alignment, and it corresponds to the eigenvector of $\langle \mathbf{Q} \rangle$ associated to the highest eigenvalue of $\langle \mathbf{q} \rangle$. The common convention is that of aligning the director reference frame with respect to the space-fixed frame such as $\mathbf{n}_z \equiv \mathbf{z}$. The order parameter $\langle P_2 \rangle$ can then be defined as :

$$\langle P_2 \rangle = \frac{3}{2} \langle q_{zz} \rangle - \frac{1}{2} \tag{5.5}$$

where $\langle q_{zz} \rangle$ is the highest eigenvalue of matrix $\langle \mathbf{Q} \rangle$. For a perfectly aligned crystal we have $\langle P_2 \rangle = 1$, while $\langle P_2 \rangle \to 0$ as $\mathrm{N}^{-1/2}$ in isotropic phase, due to

the choice of the maximum eigenvalue $\langle q_{zz} \rangle$ to define the director. Roughly speaking, nematic phases are normally characterized by an order parameter $\langle P_2 \rangle > 0.4$, while $\langle P_2 \rangle > 0.7$ typically corresponds to highly ordered smectic or columnar phases. We calculated $\langle P_2 \rangle$ values for all DNA concentrations we investigated. We choose as molecular axis the normal to the aromatic planes of the bases, and as director $\mathbf{n}$ the z axis which corresponds to the direction of double helix formation, i.e. the column growth direction. Higher values of $\langle P_2 \rangle$ were found for higher density of samples. The data reported in Fig 5.3 shows a seemingly first order liotropic phase transition between a disordered phase below a GCCG concentration of 260 mg/mL and an ordered one starting at 300 mg/mL, while experimentally the same transition occurs between 230 mg/mL and 260 mg/mL. Since $\langle P_2 \rangle$ value rapidly growed from a disordered to an high ordered configuration, the assignement of the specific LC ordered phase based solely on the order parameter values is not possible. In the next section, we report further analyses we performed to better characterize the ordered phase.

5.3.2 Radial distribution function g(r)

The Radial Distribution Function, *g(r)*, is an important observable which provides the distribution of particle centres as a function of the molecular separation $\mathbf{r}$.

$$g(r) = \frac{1}{(4\pi r^2 \rho)} \langle \delta(r - r_i) \rangle_{r_i} \tag{5.6}$$

The visualization of a snapshot representing the instantaneous configuration of the sample and the quantitative evaluation of the pair correlation function are two very useful tools to investigate the organization of our oligonucleotides solutions. The *g(r)* is strongly dependent on the physical

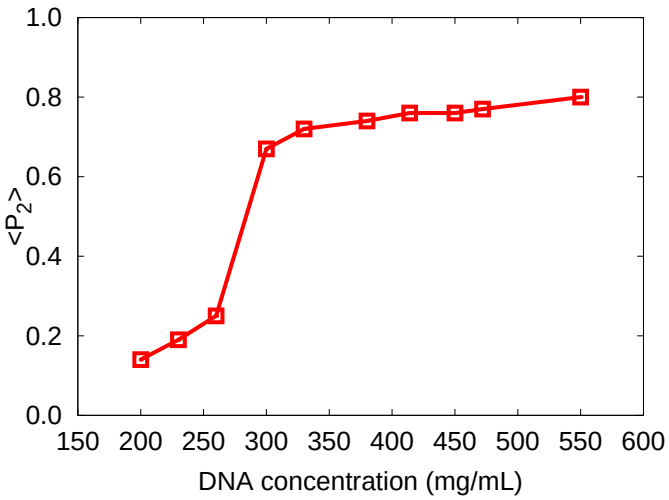

Figure 5.3: Average nematic order parameter $\langle P_2 \rangle$ as function of GCCG concentration at T = 5 °C.

state of the matter, so will vary greatly for solids, gases and liquids. The average density at any point in a liquid is referred to as the number density, $\rho = n/V$. The radial distribution function curve can be obtained by counting the number of atoms or molecules in a shell volume of $d\mathrm{V} = 4\pi r^2 dr$ divided by the expected value for an isotropic distribution ρdV[39] (Fig 5.4).

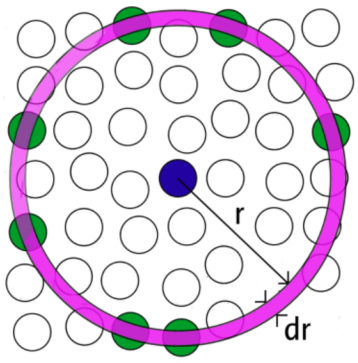

Figure 5.4: Space discretization for the evaluation of the radial distribution function.

As a matter of fact, the *g(r)* quantifies how the particle of interest is surrounded by other particles. For rarefied gas, which do not have a regular structure, the *g(r)* shows only a single broad peak; instead, for solids and liquids, where molecules have a certain regular location, it provides a spectra with several peaks which can be more or less narrow-shaped depending on parameteres as temperature or, as in our case, concentration. In order to get an in-depth picture of the phases that were formed, we moved from the coarse graining representation of single strand to an atomistic view using the tacoxDNA code [40] developed by some of the creators of oxDNA. Radial distribution functions were all obtained using the mass centre of the bases as a reference.

5.3.3 Local structure

At first, we focused our attention on the local order of our sample. By visualizing a screenshot for the simulation, we observed that the single strand structure is not stable at 200 mg/mL and $T = 5$ °C, i.e, we verified self-assembly of short oligonucleotides (see Fig. 5.5). Oligonucleotides first hybridize into duplexes, which are stable at T below their melting, then helices form linear aggregates, which posses random orientation.

In Fig. 5.6, we plotted the radial correlation function for bases of the same type (G-G and C-C) and complementary G-C bases. Each peak of the radial distribution function was assigned to the corresponding distance along the double strand structure:

- the first peak at ~ 4 Å is the distance between the stacking interaction sites C-C , G-G and C-G along the same strand. We noticed that stacking for G-G has a longer distance than C-C since they belong to different sequences, even if self-assembled. While C-C is intrastrand,

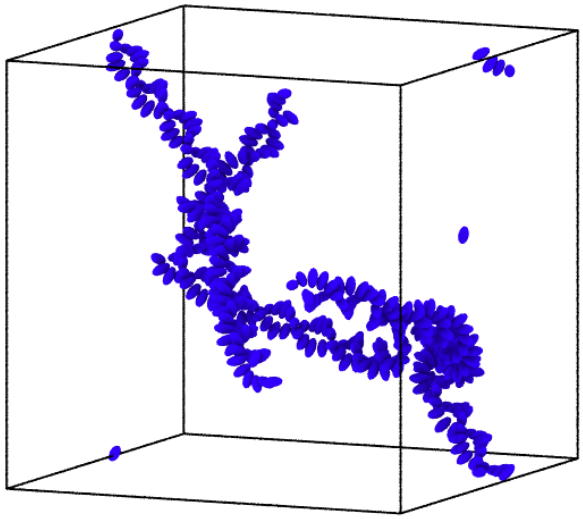

Figure 5.5: Snapshot of an istantaneous MD simulation at DNA concetration = 200 mg/mL that highlights the largest cluster formed.

G-G requires the vertical stacking of two GCCG strands.

- the second and third peaks correspond to the distance between one C (or G) belonging to one strand and the first and second neighbours of the same base type located on the complementary strand, i.e. between crossing interaction sites. Once again, we observed that for G-G distances are longer, because of the higher volume occupied by the guanine base (composed of two condensed aromatic cycles instead of one as for cytosine). For the mixed correlation function C-G, the attribution of peaks is not trivial. Our interpretation is that the peak at shorter distances is related to the base pairing distace interaction, while the longer distance corresponds to the second neighbour stacking interaction distance between complementary bases on the same strand.

To obtain a higher accuracy in finding the correspondence between the helix structure and the mixed radial correlation function G-C, we have ex-

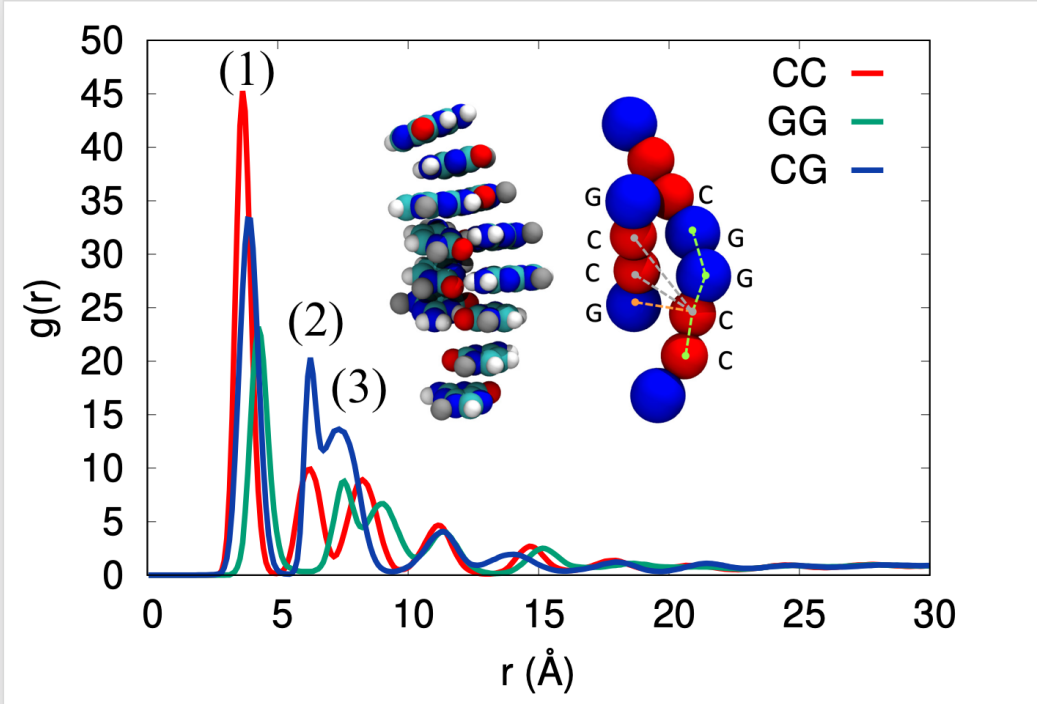

Figure 5.6: Radial correlation function for GCCG sample at 200 mg/mL and DNA double strand structure.

amined the mixed radial orientational distribution function $g(r, \cos\beta_r)$ (see Fig. 5.7), which estimates the average arrangement of C and G basis as a function of their distance r (the modulus of the vector $\mathbf{r}$ connecting their centers of mass) and of the angle β_r, that $\mathbf{r}$ forms with the phase director $\mathbf{n}$

$$g(r, \cos\beta_r) = \frac{1}{4\pi r^2 \rho} \langle \delta((r - r_{12}) \delta((\cos\beta_r - \cos\beta_{r_{12}}) \rangle \tag{5.7}$$

Specific sections of $g(r, \cos\beta_r)$ for $\cos\beta_r = 0, 1$ correspond to correlations for C-G pairs having the intersite vector orthogonal or parallel to the director.

In agreement with the purely radial distribution function, we confirmed that the first peak at shorter distances is related to the intra-strand bond between complementary bases. The distribution of $\cos\beta_r$ centred on 0.8 is due to the helix structure itself: bases are not perfectly piled on top of each

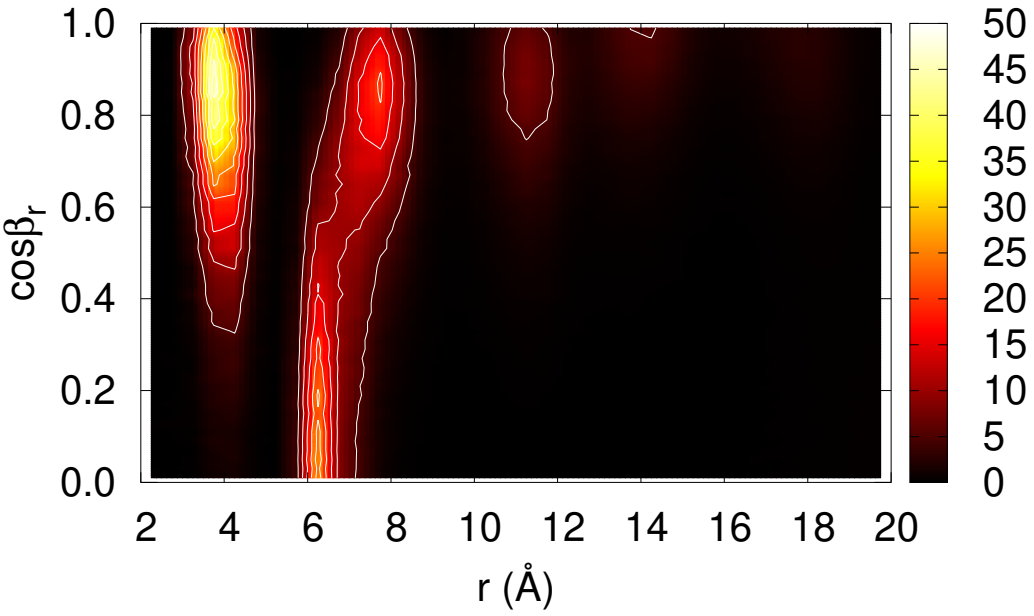

Figure 5.7: Mixed radial-orientational distribution function of the cytosine-guanine center of mass intermolecular distance r as a function of its orientation with the phase director $\cos\,\beta = \mathbf{r} \cdot \mathbf{n}$.

other, but twisted by an angle of 36° as implemented in the oxDNA model. We also confirmed the correspondence between the peak at 6 Å and the base pairing interaction distance, since the distribution of $\cos\,\beta_r$ is centred on ~ 0, i.e. 90° , while the third peak at 8 Å refers to the second neighbour stacking interaction between C-G bases along the strand: $\cos\,\beta_r \sim 1$ is typical of an intra-strand interaction.

5.3.4 Phase organization

We reported in Table 5.2 the number of clusters formed during the simulation to investigate the phase organization deriving expecially from the highest

concentrations. We noticed that enhancing DNA concentration, the number of formed cluster decreaseas, i.e. double helices are constituted by a higher number of base pairs. Moreover, the increasing of negative values of energy confirmed that a more stable and ordered phase is formed.

DNA conc (mg/mL)	N° of clusters	N° bases per largest cluster	Energy per strand (kcal/mol)
200	65	198	-35.10
230	52	280	-35.57
260	29	824	-36.03
300	16	928	-36.44
330	13	1013	-36.64
380	8	1219	-36.69
414	6	1356	-36.65
450	5	980	-36.77
472	5	1120	-36.80
550	4	1286	-36.81

Table 5.2: For each DNA concentration the number of clusters formed, the number of bases composing the largest cluster and the average energy per strand are reported.

To determine the relative position of neighbouring columns and to assign the phase with orientational order in addition to the RDF, we plot also the quantity $[g(r) - 1]\, r^2$ as a function of r (shown in Fig. 5.8 with the corresponding snapshot).

We previously observed that for a concentration equal to 200 mg/mL an isotropic phases arises and, as an further proof, above 30 Å no definite

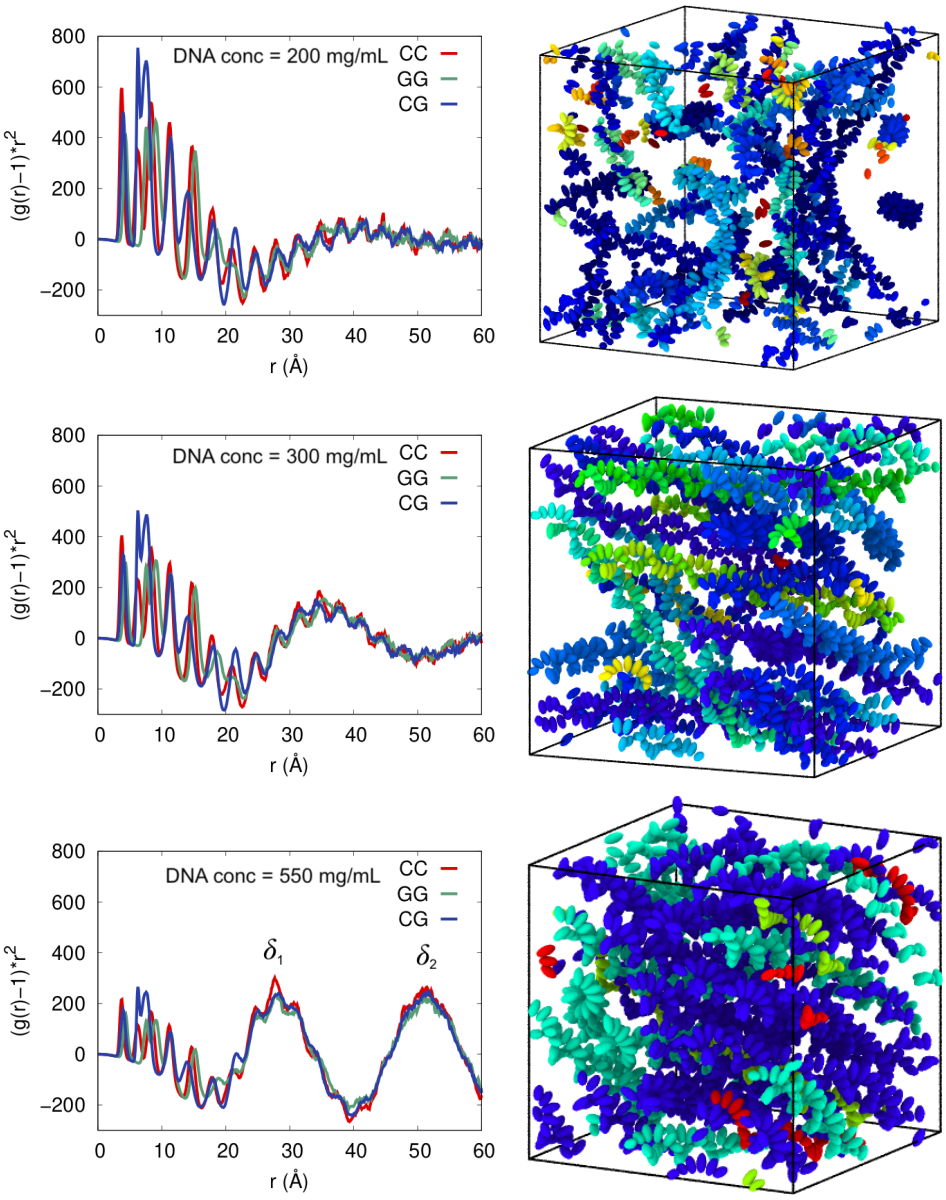

Figure 5.8: The $[g(r)-1]\,r^2$ are reported on the left , while the corresponding snapshot registered at the same DNA concentration are reported on the right.

peaks appear. When the concentration reaches 300 mg/mL, a broad peak is observed at 40 Å, while at 550 mg/mL two significant peaks are visible: one at $\delta_1 \sim 26$ Å and the other $\delta_2 \sim 52$ Å, the latter being at twice the distance than the former. We attributed to 300 mg/mL a nematic phase: even if double helix assemblies are quite longer than those in isotropic phase, we could not find a periodicity in distance between them as for 550 mg/mL, where instead a columnar phase is formed. The distance between columns is defined by δ_1.

By combining $\langle P_2 \rangle$ values with the registered *g(r)* we finally assigned the corresponding LC phase for each concentration (see Fig. 5.9). Our simulations were run in the concentration range from 200 mg/mL to 550 mg/mL.

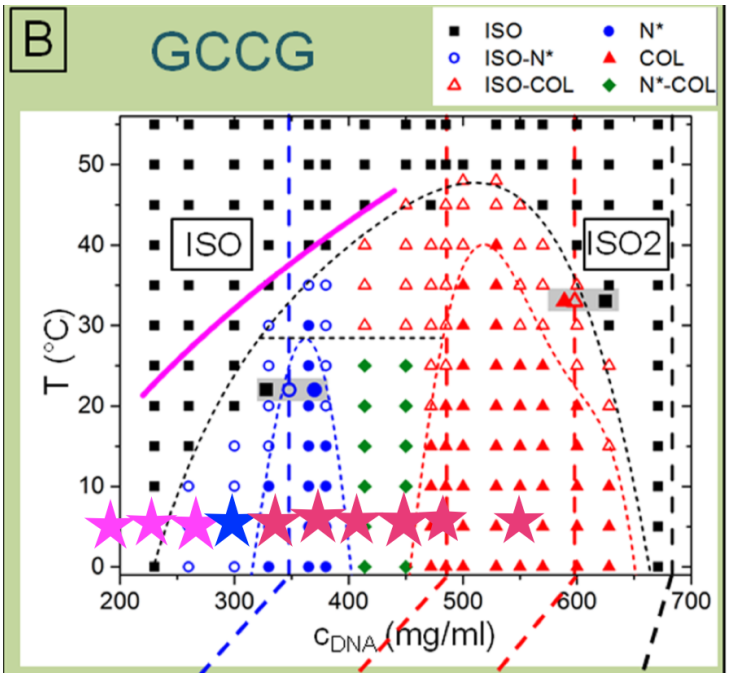

Figure 5.9: (Left) plots of $[g(r) - 1]\, r^2$ as a function of the distance for three different DNA concentrations. (Right) corresponding snapshots of instantaneous configurations of the same DNA concentrations. [Adapted from [3]].

A good agreement with experimental results is observed: all three different LC phases reported in Bellini's article were detected. However, we noticed that the existence of the chiral nematic phase for our simulations is limited to the single 300 mg/mL concentration, while in the phase diagram is visibile for a wider range of concentrations, from 260 mg/mL to 380 mg/mL.

5.3.5 Introduction of the modified DH potential

A modified description of the Debye lenght λ_D was introduced in Chapter 3 to obtain a more accurate formulation of electrostatic interactions at high electrolyte concentrations. We thus carried out simulations with this modified λ_D in the same conditions described in section 2 of this Chapter. We observed that when the saline concentration rises above 0.7 M, a more ordered phase occurs. Specifically, we noticed that while using the original Debye-Hückel potential for the sample at 260 mg/mL (corresponding to a saline concentration of 0.76 M, i.e. above the limits of DH application) an isotropic phase was detected. Instead, using the new description for λ_D, a $\langle P_2 \rangle = 0.66$ typical of the nematic phase was registered, according to experimental results [3]. In addition, a higher orientational order was found in the next DNA concentration compared to the results showed in the previous section. We are planning further analysis to investigate if for higher concentration the isotropic phase, experimentally observed above 650 mg/mL, takes place. Our aim is to demonstrate that longer Debye length, i.e. stronger electrostatics interactions, may allow to reach the isotropic phase at very high concentrations (above 650 mg/mL), until now observed only experimentally.

5.4 Summary

Despite Onsager theory prevision, which forbids LC ordering for rigid objects shorter than DNA double helices with a number of base pairs N $<$ 24, as reported in Chapter 2 of this thesis, we proved that ultrashort single strand oligonucleotide GCCG form chiral nematic and columar phases, because of their head-to-tail self-assembly into helices. By combining the calculation of the order parameter $\langle P_2 \rangle$ with the radial distribution function, we were able to determine the local structure of the newly formed double helices. Morever, we were able to characterize and assign to each concentration the correspective LC phases.

Conclusions and future outlooks

In this work, we have studied the melting transition for short double strand DNA sequences and the self assembly into LC ordered phases for GCCG single strands. MD simulations were run using the oxDNA model developed by Ouldrige *et al.* [26], which provides a coarse-grained description of the DNA structure. We have calculated the melting temperature of 4, 6 and 8 base-pairs sequence scanning a different range of temperature depending on the sequence. By plotting the fraction of double strands as a function of temperature, we determined T_m as the temperature at which curves assume the value of $f(T) = 0.5$, i.e. when half of the helices are in the single strand form. Our results, together with the thermodynamic parameters ΔH and ΔS calculated from the $f(T)$ curve, were compared with the reference work of SantaLucia [8]. The agreement we found with SantaLucia reported melting temperature reassures us on the validation of oxDNA2 model to represent oligonucleotides' interactions.

Based on this verification, we successfully demonstrated the lyotropic self-assembly for GCCG single strands in spite of Onsager theory prevision, which forbids LC ordering for double strands of length below 24 base pairs. Both chiral nematic and columnar phases formation were confimed by combining

the quantitative assesment of orientational and positional order. Specifically, we confirmed that the helix formation occurs even for isotropic phases. The same succession of LC ordered phases was observed, and a certain accuracy was found with respect to experimental results for the concentrations at which those phases appear.

In order to obtain a more accurate agreement with the reference results, we also applied a newly developed formulation of electrostatic interactions valid at very high electrolyte concentrations. A better agreement was found with the experimental results. As for the outlooks, we plan to study in deep the influence of the longer Debye length provided by the application of the recently measured electrostatic potentials.

Bibliography

[1] T. E. Ouldrigde, A. A. Louis, and J. P. K. Doye. Structural, mechanical and thermodynamic properties of a coarse-grained DNA model. *J. Chem. Phys.*, 134:085101, 2011.

[2] P. Sulc, F. Romano, Ouldrigde T. E., L. Rovigatti, J. P. K. Doye, and A. A. Louis. Sequence-dependent thermodynamics of coarse-grained DNA model. *J. Chem. Phys.*, 137:135101–3, 2012.

[3] T. P. Fraccia, G. P. Smith, L. Bethge, G. Zanchetta, G. Nava, S. Klussman, N. A. Clark, and T. Bellini. Liquid crystal ordering and isotropic gelation in solutions of 4-base-long DNA oligomers. *ACS Nano*, 10:8508–16, 2012.

[4] W. Saenger, editor. *Principles of Nucleic Acid Structure*. New York: Springer-Verlag, 1984.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, editors. *Molecular Biology of the Cell*. New York and London: Garland Science., 4th ed. edition, 2002.

[6] A. Panjkovich and F. Melo. Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics*, 21:711–722, 2005.

[7] P. Yakovchuk, E. Protozanova, and M.D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.*, 34:564–74, 2006.

[8] J. SantaLucia, H. T. Allawi, and A. Seneviratne. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 35:3555–3562, 1996.

[9] C. G. Sibley and J. E. Ahlquis. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J. Mol. Evol.*, 20:2–15, 1984.

[10] J. Jr. SantaLucia and D.J. Hicks. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.

[11] J. Jr. SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95:1460–5, 1998.

[12] N. C. Seeman. Nucleic acid junctions and lattices. *J. Theor. Biol.*, 99:237–347, 1982.

[13] F. Livolant and A. Leforestier. Condensed phases of DNA: structure and phase transitions. *Prog. Polym. Sci.*, 21:1115, 1996.

[14] M. de Frutos, A. Leforestier, and F. Livolant. Relationship between genome packing in the bacteriophage capsid and the kinetics of DNA ejection. *Biophys. Rev. and Lett*, 9:81–104, 2014.

[15] P. Bolhuis and D. Frenkel. Tracing the phase boundaries of hard spherocylinders. *J Chem Phys*, 106:666, 1997.

[16] M. Nakata, G. Zanchetta, B. D. Chapman, C. D. Jones, J. O. Cross, R. Pindak, T. Bellini, and N. A. Clark. End-to-end stacking and liquidi crystal condensation of 6 to 20 base pair DNA duplexes. *Science*, 318:1276–1279, 2007.

[17] C. R. Calladine, H. Drew, B. Luisi, and A. Travers, editors. *Understanding DNA: the molecule and how it works.* Acadamic, New York, 2004.

[18] T. Strzelecka, M. W. Davidson, and R.L. Rill. Multiple liquid crystal phases of DNA at high concentrations. *Nature*, 331:457–470, 1988.

[19] T. Bellini, R. Cerbino, and G. Zanchetta. DNA-based soft phased. *Top Curr Chem*, 318:225–79, 2012.

[20] S. Saurabh, Y. Lansac, Y. Jang, M. Glaser, N. Clark, and P. Maiti. Understanding the origin of liquid crystal ordering of ultrashort double-stranded DNA. *Phys. Rev. E*, 95, 03 2017.

[21] D. A. Potoyan, A. Savelyev, and G. A. Papoian. Recent successes in coarse-grained modeling of DNA. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3:69–83, 2013.

[22] Ankita Naithani. *Molecular Dynamics Study of the Allosteric Control Mechanisms of the Glycolytic Pathway.* PhD thesis, University of Edinburgh, UK, 2014.

[23] P. D. Dans Puiggròs. Coarse-grained models for biomolecular simulations: Theory and applications for proteins, DNA, RNA and beyond. *PRACE Simulation enviroment for life sciences*, 2018.

[24] G. S. Freeman, D. M. Hinckley, and J. J. de Pablo. A coarse-grain three-site-per-nucleotide model for DNA with explicit ions. *J. Chem. Phys.*, 135:165104, 2011.

[25] J. G. Gay and B. J. Berne. Modification of the overlap potential to mimic a linear site–site potential. *J. Chem. Phys*, 74:3316, 1981.

[26] Thomas Ouldridge. *Coarse-grained modelling of DNA and DNA self-assembly.* PhD thesis, Oxford University, UK, 2011.

[27] O. Henrich, Y.A. Gutiérrez Fosado, and T. Curk. Coarse-grained simulation of DNA using lammps. *Eur Phys J.*, 41:57, 2018.

[28] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Sulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. K. Doye. Introducing improved structural properties and salt dependence into a coarse-grainded model of DNA. *J. Chem. Phys.*, 142:234901, 2015.

[29] D. E. Draper, D. Grilley, and A. M. Soto. Ions and RNA folding - annual review of biophysics and biomolecular structure. *Annu. Rev. Biophys. Biomol. Struct.*, 34:221–243, 2005.

[30] P. Trylska, J. add Grochowski. Continuum molecular electrostatics, salt effects, and counterion binding—a review of the Poisson–Boltzmann theory and its modifications. *Biopolymers*, 89:93–1131,2,3, 2008.

[31] C. S. Santos, N. S. Murthy, G. A. Baker, and E. W. Castner. Communication: X-ray scattering from ionic liquids with pyrrolidinium cations. *J. Chem. Phys.*, 134:121101, 2011.

[32] I. Bai, Y. K. Travers, V. B. Chu, J. Lipfert, S. Doniach, and D. Herschlag. Quantitative and comprehensive decomposition of the ion atmosphere around nucleic acids. *J. Am. Chem. Soc.*, 2007.

[33] U. P. Strauss and Leung Y. P. Volume changes as a criterion for site binding of counterions by polyelectrolytes. *J. Am. Chem. Soc.*, 87:1476–1480, 1965.

[34] M. L. Bleam, C. F. Anderson, and M. T. J. Record. Relative binding affinities of monovalent cations for double-stranded DNA. *Proc. Natl. Acad. Sci. USA*, 77:3085–3089, 1980.

[35] A. M. Smith, A. A. Lee, and S. Perkin. The electrostatic screening length in concentrated electrolytes increases with concentration. *Phys. Chem. Lett.*, 2:2157–2163, 2016.

[36] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:14–101, 2007.

[37] M. R. Jones, N. C. Seeman, and C. A. Mirkin. Programmable materials and the nature of the DNA bond. *Science*, 80:1260901–1260901, 2015.

[38] J. T. Davis and G. P. Spada. Supramolecular architectures generated by self-assembly of guanosine derivatives. *Chem. Soc.*, 36:296–313, 2007.

[39] Z. Q. Hu, editor. *Modern Inorganic Synthetic Chemistry*. Ruren Xu, Wenqin Pang and Qisheng Huo, 2011.

[40] A. Suma, E. Poppleton, M. Matthies, P. Sulc, F. Romano, and Lu-ois A. A. TacoxDNA: A user-friendly web server for simulations of complex DNA structures, from single strands to origami. *J. Comp. Chem.*, 40:2517–2595, 2019.

[41] L. Verlet. Computer experiments on classical fluids. *Phys. Rev.*, 165:201, 1967.

Appendices

Appendix A

Molecular Dynamics

For a classical Molecular Dynamic simulation, the forces are usually obtained as the gradient of a potential energy function, depending on the position and possibly on the orientations of the particles. The realism of the simulation depends, firstly, on the choice of the potential, which should reproduce the one experienced by the system under the thermodynamics condition at which the simulation is run. Secondly, simulation rely on the numerical accuracy of the integration of the equations of motion and on the time length explored, chosen basen on the time scale of the studied phenomena. However, in reality, we only consider systems under some forms of constraints, in which case only a small portion of phase space, called *ensemble*, is accessible. Thermodynamic properties can be obtained by taking their average value over the simulation time: this technique is analogous to obtaining *ensemble* averages based on probability distribution functions and can be rationalized with the help of statistical mechanics theory.

A.1 Integration of the equations of motion

Solving the equations of motion requires a numerical integration of the differential equations. The integration is typically done by discretizing the variable t in small timesteps of length dt using finite difference methods. These are explicit methods, based on a Taylor expansion of the position and momenta at a time $t + dt$:

$$\begin{aligned} \mathbf{r}(t+dt) &= \mathbf{r}(t) + \dot{\mathbf{r}}(t)dt + \frac{\ddot{\mathbf{r}}(t)}{2}dt^2 + ... \\ &= \mathbf{r}(t) + \mathbf{v}(t)dt + \frac{\mathbf{f}(t)}{2m}dt^2 + ... \end{aligned} \tag{A.1}$$

A.1.1 Verlet Integration

The most common integration algorithm in Molecular Dynamics is the Verlet integrator [41], which is based on the addiction of two Taylor expansions in time, one forward and one backward:

$$\mathbf{r}(t+dt) = \mathbf{r}(t) + \mathbf{v}(t)dt + \frac{\mathbf{f}(t)}{2m}dt^2 + ... \tag{A.2}$$

$$\mathbf{r}(t-dt) = \mathbf{r}(t) - \mathbf{v}(t)dt + \frac{\mathbf{f}(t)}{2m}dt^2 + ... \tag{A.3}$$

$$\mathbf{r}(t+dt) = 2\mathbf{r}(t) + 2\mathbf{r}(t-dt) + \frac{\mathbf{f}(t)}{2m}dt^2 + O\left(dt^4\right) \tag{A.4}$$

Advantages and drawbacks of Verlet's algorithm are the following:

- Integration does not require the velocities, which are nevertheless required for the calculation of the energy. These can be extimated with the formula obtained subtracting the equation A.2 to A.3:

$$\mathbf{v}(t) = \left[\mathbf{r}(t+dt) + \mathbf{r}(t-dt)\right] / (2dt) + O\left(dt^3\right) \tag{A.5}$$

- Only a single evaluation of forces is required at each time step;
- The formulation is time reversible.

A.2 Constant temperature molecular dynamics

Simulating a system at constant temperature, thus in a canonical (NVT) ensemble, has the thermodynamical meaning of bringing the system into thermal contact with a large heat bath. In any case, the simulation temperature exploiting the equipartition can be calculated from the average kinetic energy of the system $\langle K \rangle = 1/2 m_i v_i^2$

$$\begin{aligned} \frac{3}{2} NkT &= \langle K \rangle \\ T &= \frac{2}{3k} \langle K \rangle \\ &= \frac{1}{3kN} \left\langle \sum m_i \, v_i^2 \right\rangle \end{aligned} \tag{A.6}$$

A.3 Finite size effects and boundary conditions

The finite–size of the simulated sample introduces systematic deviations from bulk (infinite size) behaviour. In order to reduce their influence on simulations, we employ the common artifact of periodic boundary conditions (PBC). This way, the simulation box is replicated in all directions to form an infinite lattice.

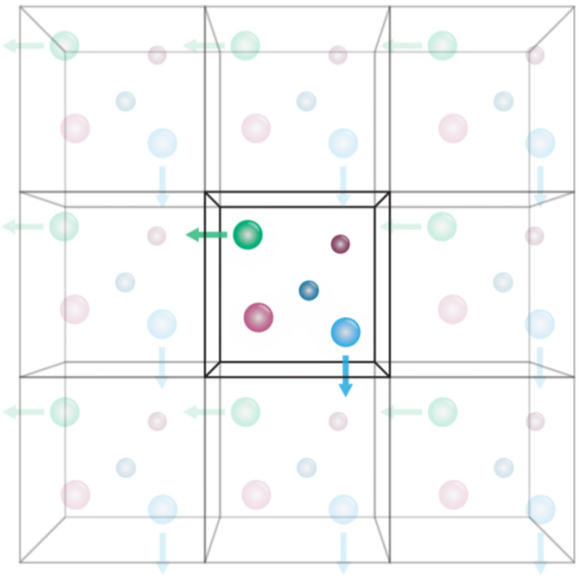

Figure A.1: Graphical representation of periodic boundary conditions