STUDY OF THE FLEXIBILITY OF DNA USING MOLECULAR DYNAMICS SIMULATIONS

Víctor Manuel Velasco Berrelleza

Doctor of Philosophy

University of York

Physics, Engineering and Technology

March 2022

Abstract

Biological processes manipulating DNA test its physical properties. Atomistic molecular dynamics simulations are a powerful tool to study the mechanical properties of DNA at atomic resolution, which is beyond the reach of single-molecule experiments. To deepen our understanding of these mechanical properties and their biological impact, a multi-approach combining simulations and experiments becomes crucial. Due to the lack of computational tools bridging these approaches, this thesis introduces two softwares for systematic analysis of nucleic acids structure and elasticity from numerical simulations, generating outputs compatible with single-molecule experiments.

The first software, SerraLINE, allows the analysis of bending angle and compaction parameter distributions from simulations. We explored the structural effects of supercoiling on DNA minicircles. Our findings indicate the level of superhelical stress found in vivo induces DNA defects, providing a mechanism to relieve torsional stress and causing the shrinking of the molecule.

SerraNA, the second software, provides local structural and flexibility parameters, along with global elastic constants, delivering a comprehensive mechanical description. Analysing the 136 unique tetramer sequences at the tetranucleotide length-scale reveals highly sequence and length-dependent elastic properties, with some sequences being 200% more flexible than others. Furthermore, exploring flexibility properties of complex DNA structures reveals that DNA-protein complexes are more rigid as proteins restrain the DNA into particular conformations, while DNA sequence mismatches act as flexible hinges.

Additionally, we conducted a pioneering analysis of DNA elastic couplings as a function of length using SerraNA. We found that twisting and stretching deformations are coupled to bending through the roll, while tilt remains uncorrelated. Principal component analysis reveals that the transition from local to bulk flexibility is driven by a stretching mode at the length of 1.5 DNA turns. Our findings reveal that the DNA elastic couplings are intrinsic to essential movements and that these can yield opposite elastic couplings.

Contents

List of Tables

List of Figures

Acknowledgements

The completion of my doctoral dissertation would have not been possible without all the people that supported and accompanied me throughout these four years. First, I am deeply grateful to Agnes Noy for her supervision and guidance throughout all these years. I would also like to extend my gratitude to the Physics of Life Group, which I was fortunate enough to be able to work and relate with the people that conform to it. I am deeply indebted to CONACyT for funding my PhD studies. And I would also like to thank the University of York for lending me their local facilities.

En lo personal, primero quiero agradecer a mis papás, que sin su apoyo no hubiera podido estudiar Física ni haber llegado hasta este punto de mi formación. También quiero agradecer a la casa 15 Flaxman Avenue, donde conocí a mi segunda familia Eduardo y Mauricio. Mi estancia en York no hubiera sido tan divertida y cálida sin la 'Reina de Cumbia', gracias por ayudarme a llenar ese hueco que dejó en mí la lejanía. Gracias a todos mis amigos mexicanos y latinoamericanos de York, que no puedo nombrarlos a todos porque se me acaba la página, pero con ustedes uno no se siente tan lejos de casa. Y por último, pero por supuesto que no menos importante, gracias Tinty por acompañarme y apoyarme todos estos años, estoy seguro que sin ti no hubiera llegado hasta aquí.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

All the results, analysis and software presented in this work, were produced by me with the guidance and support of my supervisor Agnes Noy. However, the simulations analysed here were not made by me. Particularly, the DNA minicircle and linear trajectories were produced by Agnes Noy. Simulations of DNA-protein complexes and sequence mismatches were obtained from the BIGNASim database, while the tetramer simulations were obtained from the ABC databases.

Chapter 3 shows AFM experimental results produced by Dr Alice Pyne. It is necessary that some of this work be discussed in this thesis as the SerraLINE software was developed in parallel and in collaboration.

Publications

The following publications emerged from this project:

1. Yoshua, S. B., Watson, G. D., Howard, J. A., Velasco-Berrelleza, V., Leake, M. C., and Noy, A. Integration host factor bends and bridges DNA in a multiplicity of binding modes with varying specificity. *Nucleic Acids Res. 49*, 15 (2021), 8684–8698.
   14 citations according to Google Scholar.

2. Pyne, A. L. B., Noy, A., Main, K., Velasco-Berrelleza, V., Piperakis, M. M., Mitchenall, L., Cugliandolo, F., Beton, J. G., Stevenson, C., Hoogenboom, B., Bates, A., Maxwell, A., and Harris, S. Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides. *Nat. Commun. 12* (2021)
   65 citations according to Google Scholar.

3. Velasco-Berrelleza, V., Burman, M., Shepherd, J. W., Leake, M. C., Golestanian, R., and Noy, A. SerraNA: a program to determine nucleic acids elasticity from simulation data. *Phys. Chem. Chem. Phys. 22* (2020), 19254–19266.
   19 citations according to Google Scholar.

Chapter 1

Introduction

This thesis is focused on the study of DNA structure and flexibility from the analysis of ensembles obtained by numerical simulations. Most of the work presented here revolves around SerraLINE and SerraNA, two softwares that we developed for the systematic analysis of mechanical properties from simulations of nucleic acids. This chapter provides a concise overview of nucleic acids and their biological relevance, where we then review the flexibility of DNA as well as the experimental techniques that are used to quantify its elastic properties. We then revise the most relevant models of DNA elasticity, discuss the current challenges in the field accompanied by the motivation of this work, as well as the research aims and objectives of this work. Chapter 2 provides all the mathematical tools that our two programs implement for the analysis of simulations of NA, as well as other tools such as principal component analysis, which complement the last section of this project.

In chapter 3 we introduce SerraLINE, which is a program that calculates bending angle distributions at different length-scales, as well as global parameters that characterise the shape of MD simulations of NA, whose results are suitable for comparison with experimental high resolution imaging techniques such as electron microscopy (including cryo), scanning force microscopy and atomic force microscopy (AFM). We demonstrate the usefulness of the program by analysing simulations of supercoiled DNA minicircles in combination with atomic force microscopy (AFM) performed by a counterpart experimental team, where we were able to characterise how DNA supercoiling affects the structure of double-stranded DNA. Chapter 4 introduces SerraNA, which also analyses simulations of NA but with a greater emphasis on its elastic properties. This program allows us to observe the transition from local to global flexibility, where one of its main features is to calculate global elastic constants that characterise the overall flexibility of the molecule, and which estimations are suitable for comparison with results from single-molecule experiments such as AFM, magnetic or optical tweezers. A variety of systems with different sizes and conditions are analysed with SerraNA in order to demonstrate its valuableness. Lastly, chapter 5 is focused on the analysis of the elastic couplings of DNA, and in uncovering the mechanisms that originate its flexibility. To this end, we used principal component analysis in combination with SerraNA to identify and analyse the essential modes that mainly affect the flexibility of DNA.

Finally, chapter 5 provides a final conclusion that condenses all the work presented in this project, where it then discusses possible future areas of study, including the particular cases of study that the aims of our project did not cover.

1.1 Nucleic Acids

Nucleic acids (NA) are among the most important biomolecules found in life, whose job is to encode the genetic information of every living organism. Deoxyribonucleic acid, DNA, and ribonucleic acid, RNA, are the two main types of nucleic acids. For most organisms, the instructions of life (genetic information) is encoded within the DNA molecule (encoded in RNA in case of retroviruses), which it is then transcribed to RNA molecules to be read by ribosomes in order to translate them into proteins, which allow the proper functioning of biological organisms [22].

Nucleic acids are biopolymers made of a sequence of monomers, which are called nucleotides, and are composed of three parts that are chemically connected: a pentose sugar, a phosphate group and a nitrogenous base [14] (see figure 1.1). The pentose sugar in the RNA molecule is ribose $C_5H_{10}O_5$ while in the DNA is deoxyribose $C_5H_{10}O_4$. Usually there are five kinds of nucleotides found in NA, which share the same sugars and phosphates but have different bases. For DNA, these bases are adenine (A), guanine (G), cytosine (C) and thymine (T), while in case of RNA thymine bases are replaced by uracil (U). Adenine and guanine are called purines (R) and are bigger than cytosine, thymine and uracil bases which are called pyrimidines (Y). All of these bases are aromatic rings, which make their structure particularly flat and rigid.

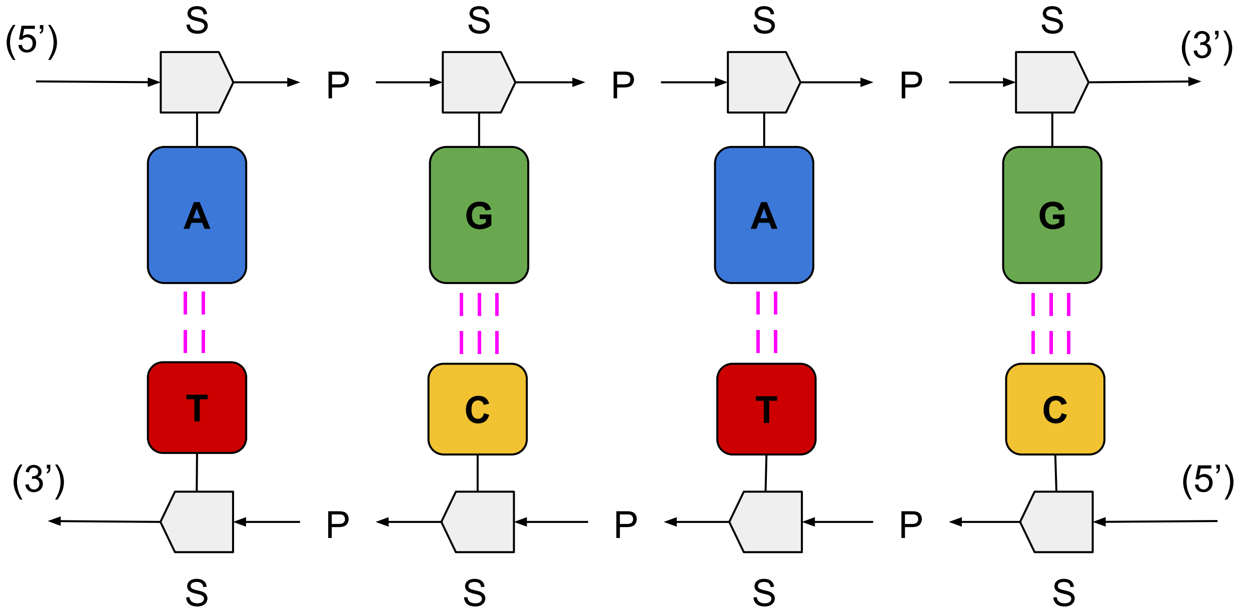


Figure 1.1: DNA structure. In the Watson-Crick model, adenine (A) can only be paired with thymine (T), and guanine (G) with cytosine (C). Letters S represent sugars and letters P represent phosphate groups. Dashed lines represent hydrogen bonds that maintain the bp together. G-C bp forms three hydrogen bonds while A-T forms two. Sugars point in the 5' to 3' direction, where both strands run in opposite directions.

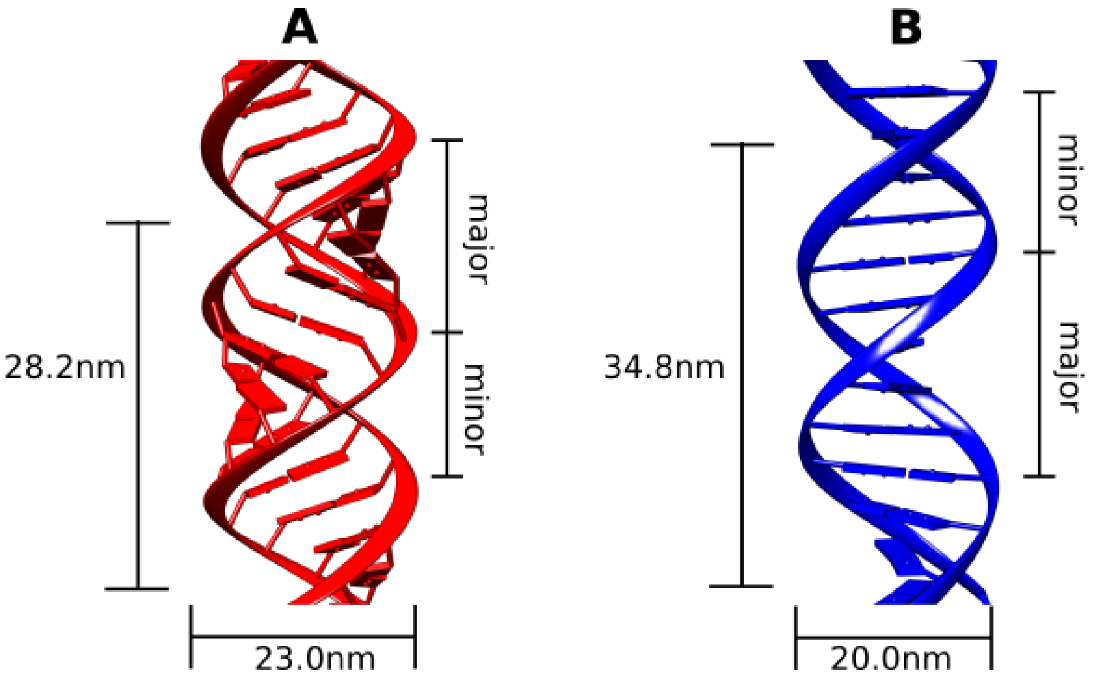


Figure 1.2: Dimension comparison between A-DNA (red) and B-DNA (blue) with major and minor grooves highlighted.

1.1.1 The structure of double-stranded DNA

In 1953, thanks to the famous X-ray diffraction image taken by Rosalind Franklin [41], Watson and Crick were able to determine the double helical structure of DNA [164]. The double helix is formed by two strands that coil around each other and are held together by hydrogen bonds between complementary bases. Generally, two types of base-pairs (bp) are found in nature, and in each type a purine pairs with a pyrimidine either in the form of A-T or G-C (A-U in case of RNA). These two types of base-pairs are roughly of the same size, where the A-T base-pair is characterised by two hydrogen bonds and G-C base-pairs by three (see figure 1.1). The fewer hydrogen bonds in A-T base-pairs (in combination with a low twist [14]) makes successive A-T sequences easier to unwind and separate. Biological processes benefit from this behavior; for example, during transcription initiation, most of the transcription machinery enzymes called RNA polymerases, tend to bind promoters with TATA sequences as they are easier to melt [145]. Due to the fact that cells are in an aqueous environment and the nitrogenous bases are hydrophobic, the planarity of the bases allow the base-pairs to stack and form a stable double helix, which protects the genetic information inside by leaving the phosphates and sugars facing the outside waters [14].

In the early 1950s, many scientists were confused with X-ray diffraction patterns that corresponded to a mix of different DNA states [68]. By defining specific humidity conditions, Rosalind Franklin resolved two main states in which the DNA molecule could exist: A- and B-DNA [40]. Currently, we know that DNA can adopt more states such as Z-DNA, hairpin loops, cruciform DNA, and more [14]. However, the DNA states most common in nature are A- and B-DNA.

Both A- and B-DNA are right handed structures with similar sizes, where A-DNA

Parameter	A-form	B-form
Base-pairs per turn	11.0	10.5
Vertical rise per bp (Å)	2.55	3.4
Rotation per bp (°)	33.0	34.3
Diameter (Å)	23.0	20.0

Table 1.1: Average structural parameters of A- and B-DNA [140].

is characterised to be shorter, with a slightly bigger diameter and more compact than B-DNA [43] [168] (see table 1.1). In B-DNA, the base-pairs are perpendicular to the helical axis, while in A-DNA, the base-pairs are tilted (see figure 1.2). Under normal physiological conditions in the cell (high salt and high humidity), the B-form is the most stable conformation, while the A-form is stable under low salt and low humidity conditions [79]. Due to this reason, B-DNA is the conformation most commonly found in cells, although the molecule may adopt transitional conformations between A- and B-forms. The double helical shape of DNA defines two different grooves: the major groove, which is wider and deeper than the minor groove (figure 1.2). The major groove in B-DNA is characterised to be wider and shallower than the major groove in A-DNA, while the minor groove is deeper and narrower in B-DNA than in A-DNA [71]. Many DNA-binding proteins benefit from this groove asymmetry as they bind specific sequences in the major groove, where bases are more accessible than in the minor groove [14].

1.2 The roles and organization of Nucleic Acids in the cell

Two processes essential for reading the genetic information in cells are transcription and replication, where the two strands need to be separated. The first process is when a gene coded within the DNA sequence is copied into a messenger-RNA by a RNA polymerase enzyme. The messenger-RNA carries the instructions of an amino-acid sequence, and transfer-RNA molecules carry the corresponding amino-acids to the protein-making machinery called ribosome to form the polypeptide chain that will end up forming a protein [14]. The second process takes place during cell division, where the DNA sequence is read by a DNA polymerase enzyme and both strands are copied into a new DNA.

Biological processes such as transcription are highly complex as they involve a myriad of supportive proteins that bind the DNA in order to regulate transcription rate [14]. In fact, in nature the DNA is rarely held in its ideal B-form as DNA-protein interactions bend, twist and stretch the DNA. For example, the transcription factor GCN4 (from the bZIP transcription factor family) [37, 72], binds just upstream of genes in order to regulate their transcription rate. The two arms of this protein hold the DNA by its major grooves and smoothly bend and untwist the DNA structure (see figure 1.3a). On the other hand, proteins called repressors, bind and prevent the RNA polymerase from transcribing the DNA [130]. Crystallographic analysis of the 434 repressor reveals that it binds DNA in a region of two helical turns, causing a compression in the DNA minor grooves and overtwisting its structure [129] (see figure 1.3b). These proteins

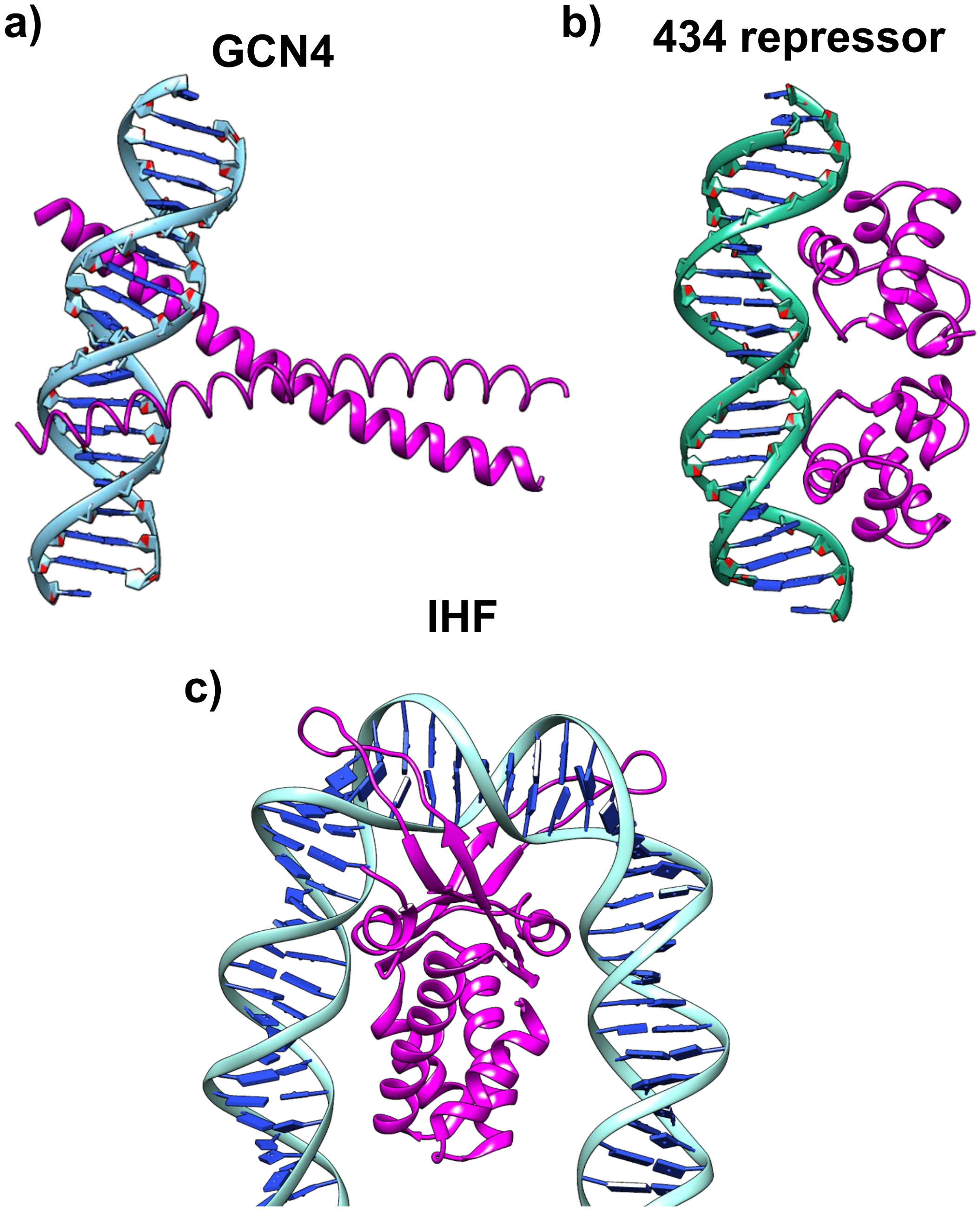


Figure 1.3: Several examples of proteins (in magenta) interacting with double-stranded DNA (in blue): The transcription factor GCN4 (PDB 2DGC [66]) (a); the transcription repressor 434 (PDB 1RPE [137]) (b) and IHF (PDB 5J0N [78]) (c).

smoothly deform the DNA, which conformations oscillate around B-DNA. However, there are other proteins such as the nucleoid associated protein IHF [1] which strongly bends DNA in a distance of just one helical turn (see figure 1.3c). This sharp bend is sufficient to induce conformations that do not longer correspond to B-DNA and in fact IHF is capable of causing a kink (unstacking a bp).

Most of the DNA sequences do not code for genes, but even non-coding regions have very important roles in the genome, as functionality is intimately related to sequence. DNA-binding proteins such as the transcription factor GCN4, the 434 repressor and IHF (see figure 1.3), bind specific sequences that allow the DNA to acquire certain conformation that facilitates their binding [39]. In chromosomal DNA, most of the DNA is organized and packaged in condensed structures called nucleosomes. These nucleosomes prefer to form in characteristic AT-rich sequences as it allows the DNA to adopt conformations that are ideal for the DNA wrapping around protein spools called histones [14]. Chromosomes can further fold into highly organised structures that can bring together distant genes, enabling a coordinated regulation of the genetic material [45].

All these factors and processes make the natural biological environment of DNA highly complex and dynamic, as the molecule is constantly manipulated by proteins that benefit from its local flexibility and sequence. Experimental techniques such as atomic force microscopy (AFM) [36] offer high resolution at the nanometer scale, but still lack the detail needed to fully comprehend the structure and interactions of DNA. Atomistic molecular dynamics (MD) simulations, on the other hand, provide atomic resolution and are often considered as super resolution microscopes. While MD simulations have been extensively used to study the DNA [23], several limitations exists when attempting to capture DNA dynamics under biological conditions.

The key limitations of MD simulations in capturing the natural environment of DNA include the restricted timescales, length scales, force field parametrization, and the lack of interactions with cellular machinery. Biological processes, such as DNA replication and transcription, occur on timescales of miliseconds, seconds and even minutes, which are challenging to simulate with current atomistc MD protocols as the timescales they can access typically range from nanoseconds to a few miliseconds [23]. Additionally, the accessible length-scales of simulations are limited to a few hundred base-pairs, being protein-DNA complexes such as nucleosomal DNA [143] or DNA minicircles some of the largest systems they can handle [8], while the study of some of the smallest genetic systems such as DNA plasmids [89], ranging from a few kbp remain impossible to simulate at the moment. Force fields are essential components of MD simulations, describing interactions between atoms based on empirical approximations. While force fields have undergone significant refinements and improvements, they still have limitations in fully capturing complex interactions involved in DNA. For instance, solvent, ionic, and protein interactions may not be fully represented by current force field parametrizations [152], affecting the accuracy of DNA simulations. Furthermore, most MD simulations primarily focus on the isolated DNA molecule without accounting for the influence of cellular factors such as proteins, multiple ions and solvent molecules, which significantly impact DNA structure, dynamics, and function in a true in vivo setting [152]. Incorporating all these factors remain a challenge to be addressed.

Despite these limitations, MD simulations have provided valuable insights into the structure, dynamics, and interactions of DNA and have contributed to our understanding of various biological processes [23]. Researchers often complement MD simulations with other experimental techniques to obtain a more comprehensive understanding of the DNA in its natural biological environment. These complementary approaches offer great advantages, as MD simulations provide atomistic detail beyond what experiments can achieve, and experiments directly measure DNA interactions with its environment and with other biomolecules. That is why we need new computational tools that facilitate the integration and comparison of both approaches, as this integration would allow us to deeper our understanding into the fundamental mechanisms governing DNA interactions under different natural conditions, and will contribute to significant advancements in this particular field of biophysics.

1.2.1 DNA supercoiling

In biological systems, DNA is frequently held under torsional stress, where the application of torsion results in an increase or decrease in the number of helical turns [172]. This characteristic state is called supercoiling [14], as DNA tends to coil around itself in order to relieve the imposed torsional stress. Negative supercoiling is referred to when the applied torsion decreases the number of helical turns, while positive supercoiling is when it increases them.

One of the important functions of supercoiling is that it is an essential mechanism for DNA packaging [62]. Imposing superhelical stress to a DNA molecule makes it more compact, which allows the DNA to be packaged. It is known that in chromosomal DNA, supercoiling acts as a gene regulation mechanism as it is folded into a hierarchy of structures, which causes a coordinated expression of genes within the same topological domain [45]. Nucleoid associated proteins like IHF (see figure 1.3c), can form DNA topological domains in which supercoiling DNA cannot diffuse outside the topological barriers [31].

In nature, most DNA is held in a negative supercoiling state [14]. Biological processes constantly introduce superhelical stress to the DNA molecule. DNA-binding proteins can also induce supercoiling, as they deform the double helix when binding (see figure 1.3). In chromosomal DNA, torsion stress is induced when the double helix is wrapped around histones [20]. Supercoiling is also intimately related with transcription and replication, as negative supercoiling can reduce the energy required to unwind and open the double helix, allowing biomolecules to access the encoded genetic information. In contrast, positive supercoiling can repress transcription initiation as well as jamming the transcription/replication machinery [19]. Topoisomerase proteins can cut the DNA strands in order to relieve the excessive torsion (either positive or negative), and hence maintain a steady superhelical state [172].

Supercoiling DNA cannot exist in linear free DNA, as it would quickly dissipate the applied torsional stress as the molecule is allowed to rotate. Therefore, the topological constraint for supercoiling DNA is to restrain both ends. There are two ways for fixing the DNA ends; by joining the two ends and hence forming a closed structure; or by anchoring the ends to a surface [14]. Both of these types of restraints are found in nature. For example, plasmids are extrachromosomal DNA typically found in bacteria, which have a closed structure (circular DNA). On the other hand, in chromosomal DNA there are looped regions in which ends are anchored to a protein surface. In either case, supercoiled DNA can adopt two types of structures: toroidal DNA, which is when the DNA acquires a circular shape, and plectonemic DNA, which is when the DNA strands coil between them due to supercoiling. In nature, it is more typical to find plectonemic structures as these are the shapes that plasmids tend to adopt; however, it is also common to find a combination of both [14].

DNA supercoiling is mathematically described through its topology with the linking number Lk [14]:

$$Lk = Tw + Wr \tag{1.1}$$

where twist (Tw) quantifies the number of helical turns and writhe (Wr) quantifies the number of coils (number of times the DNA crosses over itself).

The linking difference is a conservative quantity, so Tw and Wr have a one-to-one relationship between them; if Tw increases, then Wr decreases with the same magnitude.

In practice, it is preferable to quantify supercoiling by comparing the current linking number Lk with respect to the linking number of the relaxed state Lk_0. This is quantified through the linking difference ΔLk:

$$\Delta Lk = Lk - Lk_0 \tag{1.2}$$

The linking number of the relaxed state can be seen as the number of turns that the molecule would have if it was in the B-DNA form $Lk_0 = N/10.5bp$, where N is the total number of base-pairs in the molecule and 10.5 correspond to the number of base-pairs in B-DNA turn. Notice that if the DNA is negatively supercoiled, then $\Delta Lk < 0$, while positive supercoiling would correspond to $\Delta Lk > 0$.

Lastly, for comparing systems of different sizes, it is preferable to use the superhelical density σ:

$$\sigma = \frac{\Delta Lk}{Lk_0} \tag{1.3}$$

These are the basic mathematical tools that are used for describing DNA supercoiling. In recent years, our knowledge of DNA supercoiling has been greatly improved as we are currently aware of its role in biological processes such as: DNA packaging, genome organisation and in gene regulation. However, due to limitations in spatial resolution, it is still unclear how supercoiling affects the global and local structure of DNA in detail.

1.3 Flexibility of double-stranded DNA

The DNA molecule is one of the most important biopolymers on Earth. Similar to other polymers, the DNA exhibits elastic properties at the molecular level, as it can be deformed when subjected to external forces and revert to its relaxed conformation once the external forces are removed. The DNA possesses bending, stretching and twisting stiffness, which are quantified through the persistence length, stretch modulus and twist persistence length, respectively [109]. These quantities characterise the molecule global response to the three types of mechanical deformations. At the local level, the flexibility of DNA is significantly influenced by its sequence, playing a critical role in essential biological processes, including transcription, replication, and DNA-protein interactions, which are highly sequence-dependent [23].

In the past two decades, a variety of single-molecule techniques have been developed to visualise the DNA structure, both in isolation and under biological conditions. Among the most widely used high resolution techniques include atomic force microscopy [36], magnetic tweezers [158] and fluorescence resonance energy transfer (FRET) [132]. While these methods offer high resolution imaging capabilities, they are not able to provide atomistic length scale details and can only obtain parameters that globally characterise the analysed structures. Luckily, molecular dynamics (MD) simulations can provide the atomistic detail that single-molecule experiments lack. However, in order to complement experiments and validate these simulations by comparing with experimental data, it is necessary to calculate compatible global parameters. Examples of such parameters include the persistence length or stretch modulus for evaluating flexibility, and the radius of gyration or aspect ration for quantifying compaction. Currently, there are softwares like curves+ [76] and 3DNA [90] that can calculate parameters from simulations that evaluate the local structure of DNA, but they lack the ability to compute global parameters suitable for comparison with single-molecule experiments. To gain a comprehensive understanding of how the DNA structure is affected by biological processes, a combination of computer simulations and experimental data is crucial. Nonetheless, there remains a significant gap in the availability of computational tools that effectively bridge the gap between these two approaches. One of the main research aims of this thesis is to address this deficiency by developing novel computational frameworks that enable a more integrated and comprehensive characterization of the DNA structure and flexibility.

Single-molecule experiments have very well established the elastic constants for double-stranded DNA at the kilobase-pair level (bulk flexibility) and under biological conditions, where experimental setups such as magnetic tweezers, optical tweezers and atomic force microscopy have determined that the DNA has an average persistence length of 50 nm ($\sim$ 150 bp) [9, 56, 85, 103, 165], force-extension experiments yield a stretch modulus between 1100-1500 pN [50, 147] and experimental torque measurements obtained a twist elastic constant ranging between 90 and 120 nm [13, 85, 107]. However, between 2005 and 2010, experimental measurements of the flexibility of short DNA fragments showed discrepancies with the well established values for B-DNA. For instance, Yuan and collaborators combined the techniques of FRET with small-angle X-ray scattering (SAXS) to study the bending properties of short double-stranded DNA, which results revealed persistence lengths of less than half the accepted values (approximately 11 nm) [171]. Furthermore, atomic force microscopy (AFM) images displayed a high number of strong bends, contradicting predictions of polymer models [167]. Another SAXS experiment measured the stretch stiffness by analysing the end-to-end distance of short DNA fragments (less than 40 bp), which yielded a dramatic stretch modulus that was less than one order of magnitude lower (approximately 91 pN) than values reported in the literature [101]. These controversial results suggested the intriguing possibility that the mechanical properties of DNA might depend on the length of the molecule, implying that the DNA behaves as a flexible rod at short lengths and becomes more rigid at longer lengths when bulk flexibility is reached. Addressing these discrepancies then became a critical task for researchers in this field.

Double-stranded DNA has features such as asymmetrical grooves, hydrogen bonds and stacking interactions that not only affect its flexibility, but are also strongly sequence dependent. The stiffness of DNA also depends on multiple factors such as the salt concentration or the type of ions in the environment (monovalent or multivalent) [9]. However, in this thesis we will primarily focus on the flexibility of DNA under normal biological conditions with constant temperature (T $\sim$ 300K), while also considering other aspects that affect the flexibility such as the DNA sequence, DNA supercoiling, sequence mismatches or proteins interacting with the double-helix. In the following sections, we review in more detail the current DNA elasticity models, current knowledge regarding how the DNA sequence affects the mechanical properties of DNA at the dinucleotide level, how flexibility changes at different length-scales from local to bulk flexibility, and the current understanding of the DNA elastic couplings, which also play an essential role in biological processes. Through this review, we aim to gather deeper insight into the current knowledge of the flexibility of double-stranded DNA, highlighting the main challenges in this field which this thesis aims to address.

1.3.1 Elastic models for describing the flexibility of DNA

Here, we briefly review the most relevant elastic models of DNA used in both theoretical and experimental approaches.

Wormlike chain model (WLC)

The wormlike chain model (WLC) is the simplest mathematical description of DNA [109], where its flexibility is quantified through its persistence length A. If we define DNA as a continuous rod, and consider relatively small bending deformations, the elastic energy dE can be approximated by the following harmonic function [109]:

$$dE = \frac{k_B T}{2} A(\theta - \theta_0)^2 ds \tag{1.4}$$

where ds is an infinitesimal element of the rod length, k_B is the Boltzmann constant, T the temperature, and the term $(\theta - \theta_0)^2$ represents the fluctuations of the bending angle θ around its averaged value θ_0.

Twistable wormlike chain model (TWLC)

In practice, the WLC model can be used to estimate the persistence length of DNA, however it fails to provide a broader description of the DNA flexibility as it neglects the stretching and twisting stiffness [109]. Therefore, scientists characterise the DNA as an elastic rod with bending, stretching and twisting stiffness. The stretching and twisting stiffness can be quantified through the elastic constants of the stretch modulus B and the twist elastic constant C (also called twist persistence length or twist modulus). Adding these two elastic terms to the WLC equation 1.4 we obtain the following energy function [109]:

$$dE = \frac{1}{2}k_BT\left[A(\theta-\theta_0)^2 + B(L-L_0)^2 + C(\Omega-\Omega_0)^2\right]ds \tag{1.5}$$

where the L and Ω parameters, are the end-to-end distance and twist angle respectively. These structural parameters quantify the change in length extension and the twist rotation with respect to the average structure values, whose parameters are indicated by the 0 subscript.

This model is often referred to as the twistable wormlike chain model (TWLC) since the elastic rod is allowed to twist. If we neglect stretch and twist deformations (B=C=0) one returns to the WLC model [109] (see equation 1.4). Regarding sequence effects on DNA flexibility, they only appear at small length-scales while at longer lengths they are averaged out. Thus sequence effects are usually neglected in single-molecule experiments as they measure flexibility parameters at relatively long length scales. Lastly, the approximation of equation 1.5 describes the elastic energy of a rod that has been weakly deformed, hence higher order terms beyond the quadratic approximation should be considered in order to describe strong deformations.

Marko and Siggia (MS) model

In 1994, Marko and Siggia [98] derived an elastic theory of DNA that considers its groove asymmetry. Basically, due to the asymmetry introduced by the major and minor grooves, the bending angle (θ) has two components; roll (ρ), which quantifies bends towards the grooves, and tilt, which quantifies bends towards the backbones. The Marko and Siggia (MS) model predicted that roll (ρ) is the bending component that is coupled with twist deformations. The free energy stored in an element ds along a DNA molecule with major and minor grooves is then approximated by:

$$dE = \frac{1}{2}k_BT\left[A(\theta-\theta_0)^2 + B(L-L_0)^2 + C(\Omega-\Omega_0)^2 + 2G(\Omega-\Omega_0)(\rho-\rho_0)\right]ds \tag{1.6}$$

where G is the twist-roll coupling (also known as the twist-bend coupling). Notice that this equation does not consider other coupling terms, and that by neglecting the anisotropy of DNA ($G = 0$), we return to the TWLC model (see equation 1.5). Previous evidence from MD simulations [74] has suggested that tilt is not coupled with twist nor stretch. Similar to the TWLC model, the MS model deviates for high deforming forces, where higher order terms might correct these deviations [111]. .

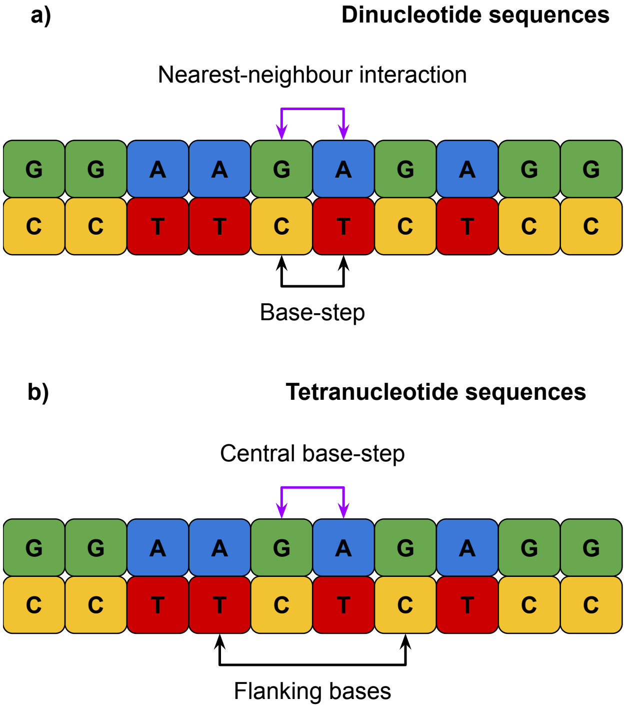


Figure 1.4: Representation of the nearest-neighbour approximation on (a) dinucleotide sequences and (b) tetranucleotide sequences of double-stranded DNA. Purple arrow indicates the nearest-neighbour interactions of the current sub-sequence. In general there are a total 10 dinucleotide sequences and 139 tetranucleotide sequences.

TWLC with external stretching force

Force-extension experiments typically consist in restraining one end of a DNA molecule while pulling the opposite end with an external and constant force [48, 49, 84]. In these types of experimental setups, scientists usually only take into account stretching and twisting deformations while neglecting bending fluctuations. In force-extension experiments, the TWLC is not able to accurately describe the physical responses unless the twist-stretch coupling (D) is introduced [49]. Similarly to continuous models, the total elastic energy stored in a stretching molecule corresponds to [48]:

$$E = \frac{1}{2}\frac{C}{L^{CL}}\Omega^2 + D\frac{x\Omega}{L^{CL}} + \frac{1}{2}\frac{B}{L^{CL}}x^2 - xF \tag{1.7}$$

where L^{CL} is the contour length when no force is applied, x the elongation beyond L^{CL} and F the applied force. Both single-molecule experiments [48,84] and computational studies [7,81,96] relay on this model for estimating the twist-stretch coupling D.

1.3.2 Sequence dependence of DNA flexibility at the dinucleotide level

The arrival of crystallographic structures in the 80s [33], created the need of defining standard protocols and nomenclatures for the description of atomic resolution NA structures. During a workshop held in Cambridge in 1989, the parameters to describe the geometry of a base-pair and base-step (dinucleotide) were defined [29]. These conventions, known as the 'Cambridge conventions' or the 'CEHS scheme' (Cambridge University Engineering Department Helix computation Scheme), continue to be the prevalent choice for describing the local structure of DNA (see figures 2.2 and 2.3). Curves+ [76] and 3DNA [90] are two of the most widely used softwares for the analysis of NA molecules, where Curves (previous version of curves+) was the pioneer software.

The CEHS scheme [29], has served as a pillar for mathematical procedures aimed at investigating the local sequence-dependent flexibility of DNA using the nearest-neighbour approximation [23]. Over time, this scheme has been extended to study flexibility at longer length scales [114]. Here, we review the existing literature concerning the double-stranded DNA sequence-dependant flexibility following the nearest-neighbour approximation.

Nearest-neighbour approximation: dinucleotide sequences

The nearest-neighbour approach for determining the flexibility of DNA consists in assuming the molecule is composed by a sequence of rigid base-pairs that only interact with their nearest neighbour. One of the first studies to use this approximation was performed by Olson and co-workers in 1998 [118]. For the first time, they empirically calculated the complete set of 10 dinucleotide energy functions (see figure 1.4a), from 92 DNA-protein crystal complexes. They approximated the energy of each base-step with the following harmonic function:

$$E = E_0 + \frac{1}{2}k_B T \sum_{i=1}^{6} \sum_{j=1}^{6} F_{ij} \Delta x_i \Delta x_j \tag{1.8}$$

where E_0 is the minimum energy, Δx_i the fluctuations of the six base-step parameters (see figure 2.3) and F_{ij} the elastic constants associated with each base-step parameter.

By carefully selecting crystal structures that fell within the harmonic behaviour and within B-DNA tendencies, Olson's team were able to assume that the calculated elastic parameters would capture the DNA natural response to imposed deformations. In general, they found that YR dinucleotides are more variable than other base-steps and that they apparently act as flexible hinges when interacting with proteins.

Other experimental studies have attempted to deduce sequence dependence properties like cyclization probabilities. However, these are indirect methods that require the fitting of measured parameters into theoretical models and are incapable of providing an atomistic description of DNA flexibility [44, 176].

It is at this stage where molecular dynamics (MD) simulations at atomic resolution become relevant and incredibly useful as they allow the systematic analysis of physical properties of DNA [25, 74]. A full list of the sequence dependence elastic constants obtained via the stiffness matrix of equation 1.3.2 was calculated by Lankas et al., [74], where they calculated the stiffness parameters from MD simulations of 20 ns long.

Later in 2007, the Parmbsc0 force-field [122] was introduced to the scientific community and, with it, the first microsecond MD simulation. This trajectory was used to provide sequence-dependant structural parameters that were in good agreement with previous experimental results from nuclear magnetic resonance spectroscopy (NMR) and X-ray structures [126]. This study demonstrated the potential of multidisciplinary studies, where computational and experimental approaches are combined to provide a more complete overview of DNA mechanics.

However, a critical limitation of atomistic MD approaches is that every atom is explicitly modelled, and that comes with a high computational cost, where it no longer becomes feasible to simulate large length scales or long time scales. To overcome this challenge, coarse-grained models in combination with Monte Carlo (MC) or MD algorithms are usually employed to gain access to both larger length and time scales at the cost of atomic resolution. Popular softwares like cgDNA [47, 123] utilise nearest-neighbour parameters obtained from atomistic MD simulations to generate sets of structural configurations for the input DNA molecule. Although coarse-grained models lose atomistic resolution, they can be rather accurate in capturing global behaviors and parameters as outputs they are comparable with both experimental and MD results. Moreover, coarse-grain models offer significant speed advantages, enabling simulations several times faster than atomistic MD.

Importantly, both atomistic and coarse-grained simulations should not be treated as two independent approaches as they have provided great insights about DNA flexibility and its role in biological procedures, both at the local level through the nearest-neighbour approximation and at the bulk level by calculating global parameters like the persistence length [104,114]. For instance, both approaches have been implemented to explain the controversial properties of A-tracts, where in some cases they present high stiffness opposing nucleosome formation but then seem to be flexible in DNA looping [34]. The origin of this behaviour lies in sequence-dependant features, where the specific AT-rich sequences drastically change the overall flexibility. In general, it was found that symmetric A-tracts (A_nT_n) are efficient for nucleosome exclusion and are more rigid than asymmetric A-tracts (A_2n), which are more flexible in terms of bending and twisting and are relevant in DNA looping.

Overall, the sequence-dependant features of DNA at the dinucleotide level and through the nearest-neighbour approach have been extensively investigated through crystallographic experiments and computational approaches, including atomistic MD simulations and coarse-grained models.

Nearest-neighbour approximation: tetranucleotide sequences

It quickly became evident that characterising the DNA flexibility from only 10 dinucleotide sequences was not enough to provide a broad picture as flanking bases influenced the flexibility of the central base-steps [77]. Therefore, the Ascona B-DNA (ABC) consortium created a microsecond MD simulations library that consisted in 39 oligomers of 18 bp that together contain all 136 unique tetranucleotide sequences [119] (see figure 1.4). This study revealed that there is indeed a strong sequence effect not only in the central base-step itself, but also in the sequences that flank them, and confirmed that many tetranucleotides have multimodal distributions in their base-step parameters. In general, it was found that the bimodal behaviour was intrinsic to some tetranucleotide sequences, particularly with central CG steps (XCGX). High resolution experimental data from crystal X-ray diffraction supported the multimodal behaviour found in DNA at the base-step level [93].

As it could be expected, the sequence effects are not limited to just the flanking bases of a dinucleotide sequence, but can be extended to the flanking bases of a tetranucleotide sequence. In fact, recent MD simulations have found evidence where the tetranucleotide sequence "CTAG " presents multiple sub-states which change according to the XCTAGX flanking sequences [4]. These findings complicate things even more as it suggests there are still more 2,080 unique hexanucleotide sequences to analyse and going a step further, 32,826 unique octanucleotide sequences. Although, creating atomistic simulations that cover all these possible sequences may not be feasible, strategies can be implemented for analysing representative sequences that would provide a broad view of the sequence space. This shows that there is still much ground to cover for understanding what key aspects of sequence contribute to DNA flexibility and at what length-scales these sequence effects are suppressed.

All these findings have contributed to our understanding of DNA and how sequence affects its flexibility and structure, where the sequence effects are not only subject to dinucleotide sequences but the flanking bases can also induce the nearest-neighbour interaction. All these gained knowledge have been used in the development of modern coarse-grained models, where they consider multimodal distributions of helical parameters from tetranucleotide sequences in order to simulate/predict conformations of B-DNA in greater length and timescales than conventional MD simulations [28, 163].

1.3.3 Flexibility at longer length scales: beyond the nearest neighbour approximation

In 2012, Noy and Golestanian [114] tried to bridge the gap between the local description of DNA elasticity given by MD simulations and crystallographic data, with the global description achieved by single molecule experiments. Their aim was to explain the controversial soft values of the persistence length (∼ 11nm) and stretch modulus (∼91 pN) calculated by SAXS experiments on short DNA fragments [101, 171]. To achieve this, they developed the Length-Dependent Elastic Model (LDEM), which was capable of describing how the bulk elastic properties of DNA emerge from base-pair fluctuations using atomistic MD simulations. Following Olson's approach [118] (equation 1.3.2), they extracted the stiffness constants beyond the dinucleotide level (see figure 1.5). These elastic constants correspond to the bending persistence length, the stretch modulus and the twist persistence length.

By analysing how these flexibility variables evolve as a function of length, the LDEM revealed that the transition from local to global flexibility occurs within one helical turn of B-DNA [114]. This transition was subsequently observed in other studies, including the reproductive work performed by Wales and collaborators [169], as well as coarse-grained simulations of OxDNA [141]. Furthermore, the LDEM model revealed that tangent-tangent correlations of base-pairs have a periodic behaviour that reflects the "crookedness" [95] of the static curvature of DNA. These tangent vec-

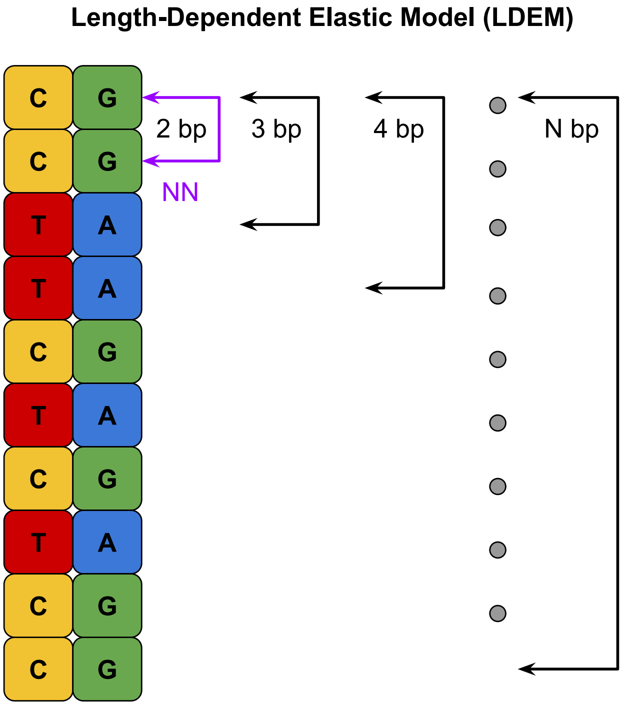


Figure 1.5: Visual representation of the LDEM on a DNA fragment composed by N bp. Purple arrows indicate nearest-neighbour (NN) interactions which are at the base-step level (2 bp), while black arrows indicate interactions at longer length-scales.

tors describe the polymer's orientation and quantify bending deformations. By employing the WLC model, Noy and Golestanian [114] were able to predict a bending persistence length around 50 nm, which is in agreement with the consensus value [30, 51, 95, 104, 136, 149, 169].

Regarding the stretch modulus as a function of length, the LDEM revealed that it followed a non-monotonic behaviour that was characterised by a high stiffness (2000-3000 pN) for short lengths less than one DNA turn, followed by a stabilisation region which values were close to force-extension measurements between 1100-1500 pN [50, 114, 147, 169]. At lengths beyond two DNA turns, the stretch modulus exhibited incredibly soft values (less than 1000 pN). They deduced that the high stiffness observed at short lengths was originated by the prevalence of strong base-stacking interactions, captured by the rise base-step parameter [29, 114]. Regarding the soft stretch modulus observed at long lengths beyond one helical turn, they found that it was mainly caused by end-effects, which also explained the low stretch modulus previously measured by SAXS experiments [101]. Finally, the results from the LDEM demonstrated that torsion elasticity follows a monotonic behaviour and transitions from soft local (30-50 nm) to rigid long-range (bulk) flexibility (90-120 nm). Their bulk flexibility predictions were in agreement with experimental predictions of the twist persistence length [13, 85, 107, 136].

Additionally, other coarse-grained models have also measured flexibility properties at length-scales beyond the nearest-neighbour level. MC simulations [104] have calculated persistence lengths from a length-dependant and sequence-dependant perspective. As previously mentioned, MC simulations allow access to longer length-scales than conventional atomistic MD, and these MC results also observed similar trends with respect to the periodic tangent-tangent correlations, where they revealed that the static curvature of DNA is the most affected by the DNA sequence. The coarse-grained model in which the OxDNA software is based [177] considers DNA interactions beyond the nearest-neighbour approximation, where nucleotides are treated as rigid bodies that mutually interact via backbone, stacking and hydrogen-bonding interactions. The OxDNA model is parametrized to reproduce mechanical and thermodynamical properties of DNA as observed in experiments. The improved version OxDNA2 implements the MS elastic model [98] (see equation 1.6) to take into account the grooves asymmetry via the two bending components: tilt and roll. Their coarse-grained MD simulations results showed similar trends in the bending and twist components of the elastic matrix, where bulk flexibility is reached within one DNA turn and the value of elastic constants agree with experimental data [141]. In a more recent study performed by Skoruppa and coworkers [144], they developed an analytical framework to estimate the bending and twisting persistence lengths from elastic models with interactions beyond the dinucleotide level. They compared MD simulations from all-atom and coarse-grained (OxDNA2) models. In both cases, the length-dependent elastic curves of bending and twisting have great similarity with predictions of the LDEM, and their persistence lengths are in agreement with the consensus values.

These collective findings highlight the importance of a length-dependent and sequence-dependent approach to achieve a deeper understanding of DNA mechanics. While DNA static curvature has been explored concerning its dependence on both length and sequence, other elastic properties such as bending and twisting persistence lengths remain to be further investigated. Atomistic MD analysis is a reliable tool to investigate the elastic properties of DNA, as current force-fields and protocols have been developed to produce trustworthy simulations of B-DNA [23]. However, despite the availability of powerful tools like computer simulations and protocols such as the LDEM [114], there is still a need for computational tools that can integrate these methods, provide both length-dependent and sequence-dependent descriptions of mechanical properties, and facilitate comparison with experiments.

1.3.4 DNA elastic couplings

Double-stranded DNA is usually characterised as a simple rod that can be bent, stretched and twisted independently. However, in reality this is not entirely true as there is experimental evidence that demonstrates these deformations are coupled. Specifically, the twist-stretch (D) and twist-bend (G) couplings have been investigated by both experimental and computational approaches. On the other hand, the bend-stretch (H) coupling has largely remained relatively unexplored.

Understanding these elastic couplings is crucial for comprehending biological processes, as DNA constantly undergoes deformations induced by interacting proteins, which can alter its local shape and, consequently, impact its overall functionality [140]. In this context, here we conduct a comprehensive review of the existing literature regarding the DNA elastic couplings, shedding light on their importance in biological systems.

Twist-stretch coupling

Single-molecule experiments usually apply external forces that stretch the DNA while measuring its rotation to estimate the twist-stretch coupling. These experimental setups implement the elastic model of equation 1.7 as the twist-stretch coupling needs to be accounted for correctly fitting the model to force-extension measurements [48,49,84].

In a pioneering study in 2006, Gore et al. [48] utilised a rotor bead tracker to measure torsion while stretching the DNA molecule, revealing a counter-intuitive behavior where the DNA overwinds when stretched, indicating a negative twist-stretch coupling (D $<$ 0). Similarly, Lionnet et al. [84] conducted a study in that same year, where they overtwisted the DNA using magnetic tweezers, and observed a twist-stretch coupling value around D=-22. Both of these studies applied biological forces under 35 pN for their force-extension experiments.

Other single-molecule experiments have used optical traps to overtwist/overstretch the DNA at higher forces (>35pN) in order to characterise regimes in which the DNA transitions to other conformational states [49, 134]. They observed that under low forces, the DNA overwinds when stretched but at higher forces (>35pN) it underwinds due to the formation of bubbles which act as a mechanism for relaxing the imposed mechanical stress. A more recent magnetic tweezers study revealed that dRNA and dsDNA have opposite twist-stretch couplings as the dsRNA molecule shortens when overwound [86]. Atomistic MD simulations of RNA/DNA molecules have achieved qualitative agreement with experimental observations, and have been implemented to provide a microscopic explanation of the striking difference between the twist-stretch couplings of dsRNA and dsDNA [7, 81, 96]. They concluded that the opposite signs in D are due to the inter-strand distance, which is correlated with the slide base-step parameter [96].

Overall, single molecule experiments have measured a negative twist-stretch coupling around D=-20 at the bulk level, while MD simulations have estimated its value to be around D=-50 [86,96] at the dinucleotide level (nearest-neighbour approach). However, despite these findings, investigations into the length-dependant and sequence-dependant features of the twist-stretch coupling remain lacking. Currently, there is no established framework for estimating this elastic coupling at the bulk level from simulations, which would quantify the overall correlation between twist and stretch deformations in the DNA molecule. Understanding and characterising the twist-stretch coupling is of vital importance since multiple biological processes involve the stretching of DNA. The stretching forces acting on the DNA molecule are capable of inducing local changes in its structure, influencing its functionality and potentially modulating the binding of proteins [140].

Twist-bend coupling

Regarding the twist-bend coupling, it is surprising that it has not been given as much attention as the twist-stretch coupling since its existence was already predicted in 1994 by Marko and Siggia [98] (see equation 1.6). One of the first works that focused on the study of the bend-twist coupling was performed by Golestian and co-workers in 2005 [106]. They developed an elasticity model similar to the MS model [98] (equation 1.6) that included the twist-stretch coupling. This model was fitted to data from a crystal structure of nucleosomal DNA, where they estimated a twist-bend coupling of G=25nm.

More than a decade later, a new study performed by Carlon and collaborators [111] investigated results from force-extension experiments [85, 87]. They found that the effective twist modulus changes as a function of extension-force, deviating from the TWLC model. Through computer simulations, they demonstrated that the MS model corrected these deviations, since it accounted for the twist-bend coupling, which they estimated to have a value of $\mathrm{G} = 40 \pm 10$ nm. That same year, the anisotropy of DNA was introduced in OxDNA1, resulting in the OxDNA2 software which is also based on the MS model [148]. OxDNA2 is capable of reproducing length-dependent elastic curves of the twist-bend coupling, which reaches a plateau (bulk behaviour) within one helical turn with G=30 ± 1nm, which agrees with previous estimations [111].

Recent investigations have shifted focus from the direct estimation of the twist-bend coupling to studying its effects on the structure of DNA under various circumstances. It has been found that bending stress induces oscillations in the twist angle along the molecule with a period equal to one helical turn (approximately 10.5 bp) [110]. These waves are termed "twist waves' ' and are induced by bending stress via the twist-bend coupling. Twist waves have been observed in crystal structures of nucleosomal DNA [143]. In fact, it was thought that the twist oscillations found in nucleosomal DNA were caused by protein interactions, but it turns out that it is an intrinsic effect of bent DNA. Coarse-grained MC and MD simulations of DNA minicircles and DNA loops have reproduced these twist waves [110], showing different behaviours. The amplitude of twist oscillations in minicircles remained constant, while in DNA loops, the amplitude increases as it reaches the apex of the loop. Other coarse-grained simulations of over and undertwisted DNA minicircles, have found that the excess of twist induces circular polygonal shapes within the DNA structure [15]. These findings are biologically relevant, as proteins might bend DNA in specific regions induce twists, facilitating protein binding. Furthermore, it is known that nucleosomes slide along the DNA, and it is believed that the diffusion of twist defects may be the driving mechanism [38], which would induce changes in the local curvature.

These investigations underscore the importance of the twist-bend coupling in biological processes, such as nucleosome formation, where DNA is smoothly bent around histone proteins. By fitting force-extension measurements to elastic models, the bulk twist-bend coupling is estimated to have a value between 20 to 40 nm. Similar to the twist-stretch modulus, there are currently no established frameworks for predicting the twist-bend coupling at the bulk level through numerical simulations. Additionally, computational tools allowing the computation of the twist-bend coupling at distinct length-scales or emphasising sequence-dependent properties are currently lacking.

Bend-stretch coupling

Regarding the bend-stretch coupling, most of the current elastic models predominantly focus on either the twists-stretch coupling [48, 49, 84] or the twist-bend coupling [98, 111]. Some studies have incorporated a bend-stretch component into their elastic models [106], but the observed effects were not significant. It is worth noting that this latter study primarily analysed elastic correlations in nucleosomal structures, where the principal deformations were twisting and bending, not stretching. However, some attempts have been made to calculate the bend-stretch coupling from atomistic MD simulations and crystallographic structures [74, 116], which results have provided evidence of the existence of this coupling. According to their results, similar to the twist-roll coupling, the bending and stretching deformations are coupled via the roll component, indicating a roll-stretch coupling instead of a direct bend-stretch coupling. This roll-stretch coupling arises due to the groove assymetry of DNA.

Even though some investigations have shed some light on the existence of a bend-stretch coupling, there is no consensus on its magnitude, and there are no current experimental setups that allow for its direct estimation. Given that the three main flexibilities of DNA are coupled, it is crucial to invetigate the magnitude and behavior of these couplings, as they play important roles in fundamental biological processes as transcription or replication, where proteins severely deform the DNA by twisting, stretching and bending its structure. This also includes the bend-stretch coupling. Recent studies provide evidence that proteins deform the DNA along its essential movements, suggesting that proteins benefit from the sequence-dependent flexibility properties of DNA to induce deformations on the DNA [163]. This raises the opportunity to further expand our understanding on DNA mechanics through the analysis atomistic MD simulations of free B-DNA, as infering the flexibility properties of the isolated molecule would shed light into its functionality and potential interactions with other biomolecules.

1.4 Research aims and objectives

The aims and objectives of this dissertation revolve around studying the structural and elastic properties of DNA through the detailed analysis of atomistic MD simulations. The motivation behind this work arises from the lack of computational tools that enable direct comparison between atomistic MD simulations and experimental assays. To address this gap, a significant aspect of this thesis focuses on the creation of two software tools that will aim to facilitate the integration of experiments and simulations, enabling comprehensive analysis of local and global structural and elastic properties of DNA, offering valuable insights beyond experimental limitations due to their resolution. Additionally, the parameters provided by our software can be utilised for mechanistic characterization of the analysed molecules, enabling their comparison with experimental data as well as validation.

The development of such software tools is crucial as atomistic MD simulations serve as powerful microscopes with atomistic resolution, making them ideal for studying mechanical properties in detail. Thus, another significant aspect of this dissertation involves studying mechanical properties of DNA through MD simulations, employing the softwares tools that we developed. Using these tools, we will investigate various aspects, including the mechanical response of DNA under superhelical stress, DNA-protein interactions, and the exploration of sequence and length-dependent properties. Furthermore, this research aims to explore the widely unexplored DNA elastic couplings. We will employ our software tools to provide a mathematical description of the elastic couplings. Additionally, essential dynamics analysis will be employed to identify principal movements that govern DNA flexibility, establishing a link between essential dynamics and DNA flexibility. The development of the two software tools as well as the investigations of different aspects of DNA flexibility will be distributed into three results chapters.

1.4.1 SerraLINE

The first main objective of this thesis is to develop SerraLINE, a software tool that will mimic measurements from single molecule experiments such as atomic force microscopy (AFM), where the DNA is visualized in a 2D plane. SerraLINE will incorporate various vectorial techniques to calculate bending angles at different lengths from a simulation of the molecular contour of a polymer. It will also be able to project the structure onto a 2D plane, enabling the extraction and analysis of bending angle distributions and compaction parameters like the aspect ratio. These structural parameters are suitable for comparison with experimental techniques used to characterise DNA molecules.

To ensure efficient processing of the large trajectory files utilized in atomistic MD simulations, SerraLINE will be programmed in Fortran 90. Fortran 90 has fast processing speeds that make it an ideal choice for handling such data, enabling the required mathematical procedures to be executed within minutes.

In the initial chapter of this thesis, we will be using SerraLINE to analyse MD simulations of DNA minicircles at various physiological supercoiling levels. By comparing the simulation results with AFM experiments conducted at similar supercoiling levels, we aim to establish a more direct and comprehensive comparison between both approaches, gaining valuable insight about detailed features that experiments cannot provide alone while at the same time validating the simulations. The objectives of this chapter include evaluating the capabilities of SerraLINE in analysing MD simulations of DNA and acquiring new significant insights into the structural characteristics of supercoiled DNA as the superhelical density increases. Particularly, we aim to uncover the deformations induced in the DNA structure by supercoiling, which could have crucial implications in biological processes, such as DNA packaging.

1.4.2 SerraNA

The second main objective is to develop SerraNA, a software tool that extracts structural and elastic properties in detail from atomistic MD simulations of nucleic acids. Unlike SerraLINE, which mimics single molecule experiments, SerraNA implements the LDEM model [114] to calculate structural parameters at different length scales to then use them to infer elastic properties. These parameters provide a more comprehensive structural and elastic description of the analysed molecule at the local level. These local properties will then be used to infer global elastic parameters that evaluate the molecule overall flexibility. Such parameters will be the stretch modulus, twist persistence length and bending persistence length. These parameters are suitable for comparison with experimental techniques as well, and can be used for validation and characterisation of DNA molecules.

Similar to SerraLINE, SerraNA will be written in Fortran 90 to efficiently handle MD trajectory files and perform the required mathematical procedures. In the second chapter, SerraNA will be tested by analysing short fragments of linear DNA, extracting global elastic parameters and comparing them with literature values to evaluate the program's capabilities. Furthermore, we will calculate local structural properties at various length scales from the set of linear DNA simulations, allowing us to observe the transition from local to bulk flexibility. This comprehensive analysis will showcase the potential of SerraNA in providing detailed insights into mechanistic properties of MD simulations of DNA, spanning from local to global levels.

Next, we will employ SerraNA to analyse more complex structures, such as DNA bound by proteins and DNA with sequence mismatches. Through these investigations, we expect to uncover valuable insights into the changes in DNA flexibility under the influence of interacting proteins and the impact of sequence mismatches. These findings will shed light on their biological relevance in processes like DNA-protein interactions and DNA repair. Additionally, these investigations will further explore the program's limitations and its ability to provide valuable insights even when analysing mechanically deformed structures. Furthermore, SerraNA will be used to analyse a database containing MD simulations of all 136 tetranucleotide sequences [119], studying their flexibility at the tetranucleotide level, which has not been explored previously. This research will provide new insights into sequence and length-dependent mechanical properties of DNA, which directly influence various biological processes.

1.4.3 Investigating the DNA elastic couplings

The third main objective of this dissertation focuses on the study of the DNA elastic couplings, an area that has received limited exploration. To this end, SerraNA will be able to calculate the twist-bend, twist-stretch, and bend-stretch elastic couplings. In the third chapter of this thesis, we aim to produce the elastic profiles of these couplings by analysing MD simulations of free DNA, providing a new mathematical description and calculating parameters that characterise the overall elastic couplings as a function of length. The set of elastic parameters including couplings, are particularly important for estimating elastic energies (see equations 1.6 and 1.3.2) as well as parametrizing flexibility models of DNA [97], considering the DNA flexibility is highly sequence dependant.

Next, to study the movements that originate the flexibility of DNA, including elastic couplings, we will employ principal component analysis (PCA) to analyse the essential dynamics of DNA [138], and obtaining the set of essential movements that cause most of the variance in the simulations. These essential movements will be associated with flexibility variables based on the deformations they induce. This analysis will be performed on various sequences and DNA fragments to compare similarities between essential movements. Furthermore, we will analyse the conformational space accessible by DNA along these essential movements to study the natural deformations they induce.

Lastly, the calculated set of DNA essential modes will be used to project relaxed

DNA structures onto more complex structures, such as supercoiled DNA or DNA-protein complexes. This analysis aims to expand our knowledge of DNA dynamics through flexibility and essential dynamics analysis, and could have potential implications in DNA-protein recognition as previous studies have suggested that proteins deform DNA along its essential modes [163].

Overall, the objectives of this PhD dissertation focus on developing two software tools, SerraLINE and SerraNA. These tools serve multiple purposes, such as enabling the comprehensive analysis of structural and elastic properties of DNA at different scales, facilitating the comparison and validation of simulations with experimental data. These software tools will be employed to characterise the structural response of DNA under superhelical stress, as well as studying sequence and length dependant properties of free DNA, DNA-protein complexes and sequence mismatches. Additionally, SerraNA will also be employed to investigate the widely unexplored area of DNA elastic couplings, as well as associating 3D deformations with essential dynamics. By achieving these objectives, this research aims to contribute to the advancement of knowledge in the area of DNA structure and flexibility, and to provide valuable computational tools to the scientific community to further expand this particular biophysical field.

Chapter 2

Methods

Synopsis

In this thesis, we aim to study the structure and flexibility of DNA by implementing multiple methods. To aid this task, in this chapter we introduce some key processes that are directly required for our research. For studying the structural properties of DNA, it is essential to provide a description of the DNA geometry at the dinucleotide level. We further extend the parameters that describe the geometry between a pair of bp, by following the procedures of the Length Dependence Elastic Model (LDEM). This geometric model allows us to study the elastic properties of DNA in detail across multiple length scales. We then introduce a powerful technique called Principal Component Analysis that is used to calculate the essential modes from a trajectory that capture most of the fluctuations from the simulation. This technique can be used to compress trajectory files, however, here we implement it to associate essential modes to flexibility parameters. Finally, we describe the methods of the WrLINE software, which we use to calculate DNA molecular contours from simulations that are suitable for comparison with high resolution experiments such as AFM.

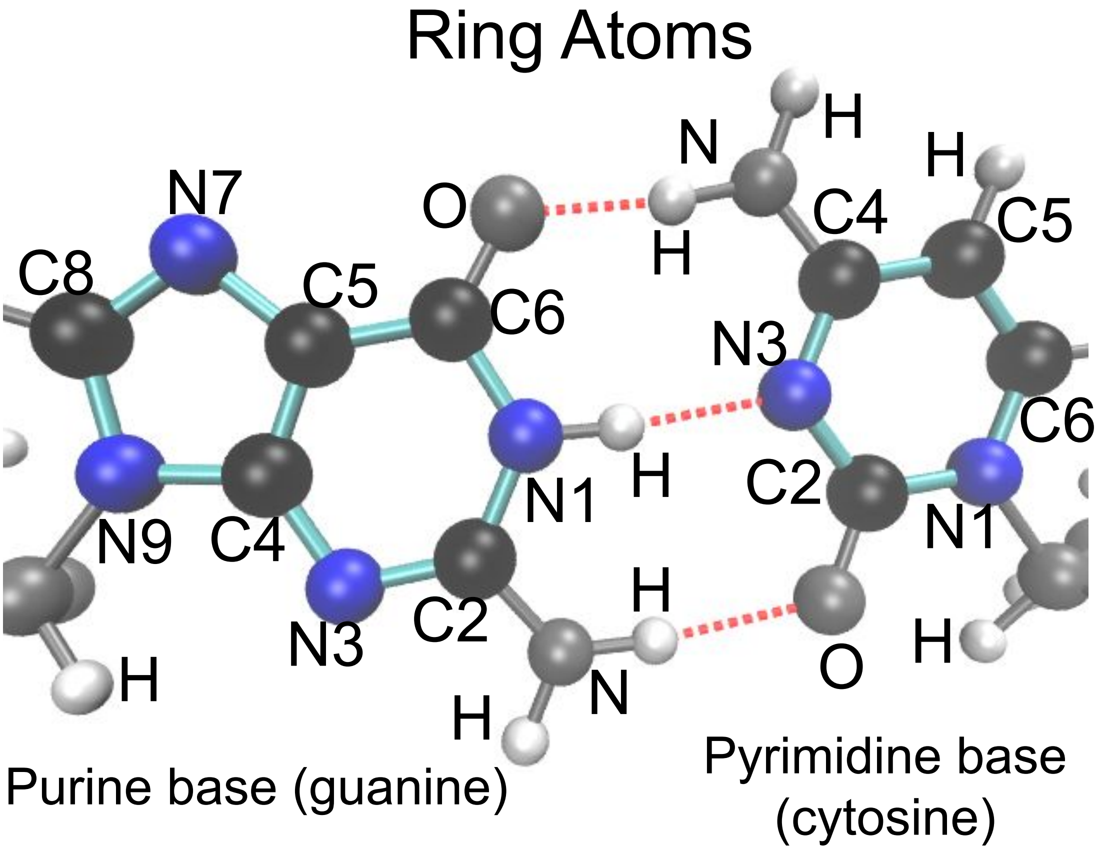


Figure 2.1: An example of purine (guanine) and a pyrimidine (cytosine) bases forming a base-pair. Highlighted carbon (black) and nitrogen (blue) atoms participate in the fitting process, where their covalent bonds are colored as cyan. Three hydrogen (white) atoms maintain the bases paired through their hydrogen bonds (red lines). Oxygen (O) and nitrogen (N) atoms that are not involved in the fitting process are represented as gray spheres.

2.1 Base-pair Geometry

In this section, we provide the tools for describing the geometry of NA molecules by following the CEHS scheme [91], which is computationally implemented by the 3DNA program [90]. CEHS describes the geometry of NA molecules through a set of parameters called the base-pair parameters (BPP), which describe the geometry between two bases (nucleotides) that compose a bp, and a set of parameters called the base-step parameters (BSP), which are used to describe the geometry between two consecutive bp. These two sets of parameters are each composed of 3 translations and 3 rotations, and have been widely used in the literature due to the fact that the CEHS scheme is mathematically rigorous and reversible [92].

2.1.1 Base fittings

Before calculating the BPP and the BSP, we need to fit a standard base S to each observed base E. Standard bases (S) are previously defined and contain the coordinates of ring atoms from bases which are approximately planar, while observed bases (E) contain the ring atoms in the atomistic simulation. We will do so by following the procedures of the 3DNA program [90]. The aim of this process is to calculate a reference frame for base i, which is defined by a position vector $\vec{O}_i$ and a orientation matrix R_i.

Then, a close-form solution of absolute orientation using unit quaternions [58] can be used for the least-squares problem of fitting a standard base S to an observed base E.

Purine bases (G & A) are composed of nine ring atoms and pyrimidines (T, C and U) of six (see figure 2.1). Once matrices S and E are built, the first step for fitting S onto E is to calculate the following 3×3 covariance matrix C:

$$C = \frac{1}{N-1}\left(S^T E - \frac{1}{N} S^T J J^T E\right) \tag{2.1}$$

Where N is the number of atoms in each base (9 for purines and 6 for pyrimidines), and J is a $N \times 1$ column vector where each element is equal to one.

Then a 4×4 matrix M is calculated using the elements of C:

$$M = \begin{bmatrix} C_{1,1}+C_{2,2}+C_{3,3} & C_{2,3}-C_{3,2} & C_{3,1}-C_{1,3} & C_{1,2}-C_{2,1} \\ C_{2,3}-C_{3,2} & C_{1,1}-C_{2,2}-C_{3,3} & C_{1,2}+C_{2,1} & C_{3,1}+C_{1,3} \\ C_{3,1}-C_{1,3} & C_{1,2}+C_{2,1} & -C_{1,1}+C_{2,2}-C_{3,3} & C_{2,3}+C_{3,2} \\ C_{1,2}-C_{2,1} & C_{3,1}+C_{1,3} & C_{2,3}+C_{3,2} & -C_{1,1}-C_{2,2}+C_{3,3} \end{bmatrix} \tag{2.2}$$

To continue with the fitting process, the eigenvector with the corresponding largest eigenvalue needs to be calculated. There are many approaches to find the eigenvectors of matrix M, but here we benefit from the fact that M is a real symmetric matrix (hence diagonalizable) to apply the Jacobi algorithm [102]. Through the Jacobi diagonalization algorithm (see section 2.5), we can find the eigenvector $\vec{v}$ with the largest eigenvalue of matrix M. The elements v_i of this eigenvector are used to calculate the following rotation matrix:

$$R = \begin{bmatrix} v_1v_1+v_2v_2-v_3v_3-v_4v_4 & 2(v_2v_3-v_1v_4) & 2(v_2v_4+v_1v_3) \\ 2(v_3v_2+v_1v_4) & v_1v_1-v_2v_2+v_3v_3-v_4v_4 & 2(v_3v_4-v_1v_2) \\ 2(v_4v_2-v_1v_3) & 2(v_4v_3+v_1v_2) & v_1v_1-v_2v_2-v_3v_3+v_4v_4 \end{bmatrix} \tag{2.3}$$

R is precisely the orientation of the base i fitted to the reference frame, hence we refer to it as R_i. The first column vector of the matrix R corresponds to the x-axis, the second to the y-axis and the third to z-axis. Lastly, the position vector of base i is calculated as:

$$\vec{O} = \bar{E} - \bar{S}R^T \tag{2.4}$$

Where $\bar{E}$ are averaged coordinates of ring atoms in E, and $\bar{S}$ the average coordinates of S. Following this procedure, an orientation R_i and an origin vector $\vec{O}_i$ is assigned to each base i, which can be used to calculate the BPP and BSP.

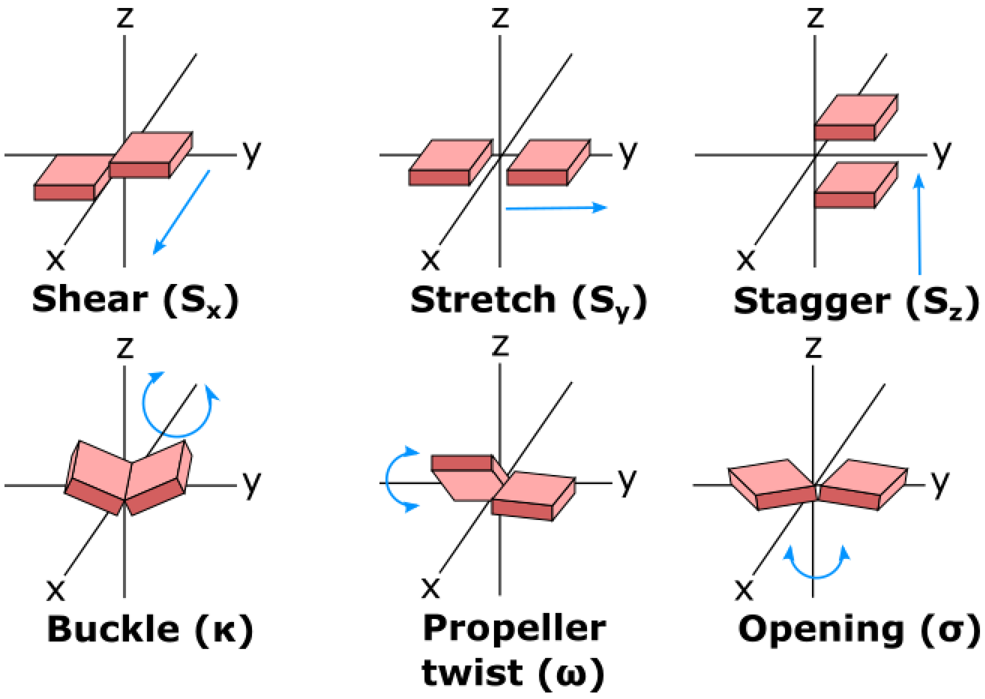


Figure 2.2: Representation of the six base-pair parameters (BPP). Each red square represents a nucleic acid base, where the shaded areas indicate the minor grooves. Blue arrows indicate the direction of the movements.

Lastly, an important thing to note is that the bp index i increases from 5'-to-3' and that the orientations R are constructed so the x-axes are always pointing towards the minor groove. In case of a double stranded structure, the directions of the y- and z-axes of the second strand need to be reversed in order to calculate the BPP and BSP parameters. With this sign conversion the z-axes point towards 5'-to-3':

$$R_y = -R_y \tag{2.5}$$

$$R_z = -R_z \tag{2.6}$$

2.1.2 CEHS scheme: base-pair parameters (BPP)

Once the standard bases S have been fitted to each observed base E, and each base has its origin $\vec{O}_i$ and orientation R_i, the corresponding BPP can be calculated. The BPP describes the geometry between two bases that form a bp through a set of parameters composed of three translations and three rotations. Figure 2.2 shows these six parameters, where the shear (S_x), stretch (S_y) and stagger (S_z) translations describe relative movements between bases in the x, y, z directions, respectively, and, similarly, the buckle (κ), propeller twist (ω) and opening (σ) angles describe rotations in the x, y, z directions, respectively. Each pair of bases have their respective x, y and z axes, which are constructed with a matrix called the mid-base triad (MBT) that will be described in the following process. Lastly, it is worth pointing out that the BPP can only be calculated for double-stranded structures.

Given the position vectors $\vec{O}_I$ & $\vec{O}_{II}$ and orientation matrices R_I & R_{II} of the base of the first strand I and the base of the second strand II, the first step is to calculate the BucklePropeller angle, which is defined as the angle between the z-axes of these two bases:

$$\delta = \arccos(\hat{z}_I \cdot \hat{z}_{II}) \tag{2.7}$$

Note that $\hat{z} = R_z$ for each base, and, from now we will use the same nomenclature for the other two axes ($\hat{x} = R_x$ & $\hat{y} = R_y$). Then the BucklePropeller axis is defined as the vector perpendicular to the two zth directions:

$$\hat{bo} = \hat{z}_I \times \hat{z}_{II} \tag{2.8}$$

Then, each orientation matrix is rotated by half the BucklePropeller angle δ around the BucklePropeller axis $\hat{bo}$:

$$R'_I = R_{bo}\left(+\frac{\delta}{2}\right) R_I \tag{2.9}$$

$$R'_{II} = R_{bo}\left(-\frac{\delta}{2}\right) R_{II} \tag{2.10}$$

where matrices $R_{bo}(+\delta)$ & $R_{bo}(-\delta)$ are general rotation matrices [54]. The column vectors of the new matrices R' represent new directions that we denote as $\hat{x}', \hat{y}', \hat{z}'$.

We refer to these new matrices R'_I & R'_{II} as the triads of the first and second strands, and their averaged directions are used to construct the MBT:

$$T_{mbt} = \frac{1}{2}\left(R'_I + R'_{II}\right) \tag{2.11}$$

Same as orientation matrices, each of the column vectors of the MBT represents a direction, which we denote as $\hat{x}_{mbt}$, $\hat{y}_{mbt}$, $\hat{z}_{mbt}$.

With these triads constructed (R'_I, R'_{II}, T_{mbt}), we can proceed to calculate the three angles that compose the BPP. The opening angle σ is defined as the angle between the y' axes:

$$\begin{gathered} \sigma = \arccos(\hat{y}'_I \cdot \hat{y}'_{II}) \\ if\ \ (\hat{y}'_{II} \times \hat{y}'_I) \cdot \hat{z}_{mbt}\ < 0, \quad \sigma < 0 \\ if\ \ (\hat{y}'_{II} \times \hat{y}'_I) \cdot \hat{z}_{mbt}\ > 0, \quad \sigma > 0 \end{gathered} \tag{2.12}$$

To define the other two angles, we need the angle between BuckleOpenning $\hat{bo}$ and the $\hat{y}_{mbt}$ vectors:

$$\begin{gathered} \phi = \arccos(\hat{bo} \cdot \hat{y}_{mbt}) \\ if\ \ (\hat{bo} \times \hat{y}_{mbt}) \cdot \hat{z}_{mbt}\ < 0, \quad \phi < 0 \\ if\ \ (\hat{bo} \times \hat{y}_{mbt}) \cdot \hat{z}_{mbt}\ > 0, \quad \phi > 0 \end{gathered} \tag{2.13}$$

Then, the buckle κ and the propeller ω angles are calculated as:

$$\kappa = \delta \sin(\phi) \tag{2.14}$$

$$\omega = \delta \cos(\phi) \tag{2.15}$$

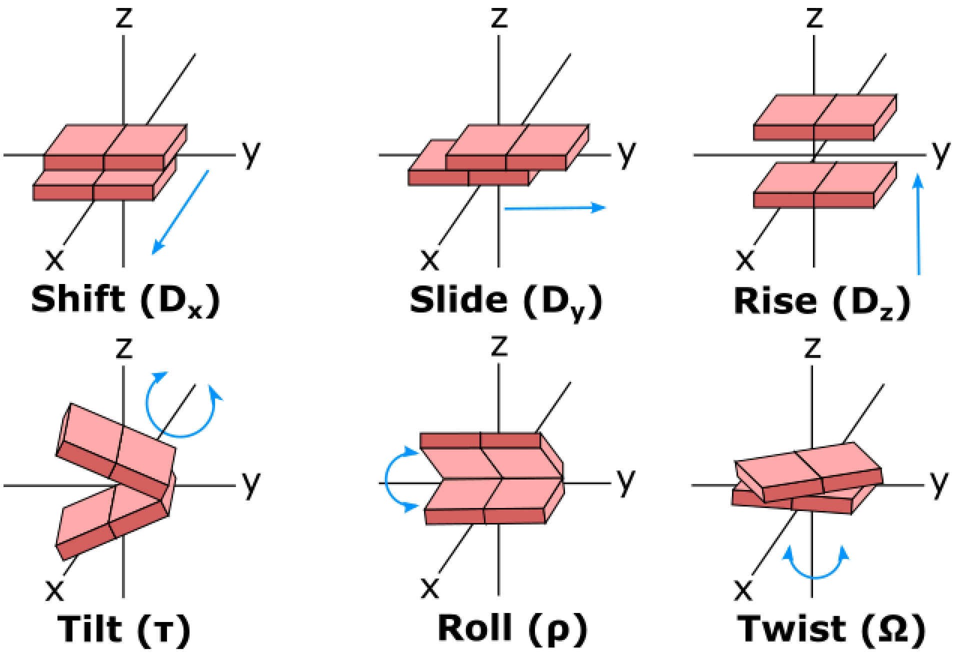


Figure 2.3: Representation of the six base-step parameters (BSP). Each red square represents a nucleic acid base, where joined squares form a bp. Shaded areas indicate minor grooves and blue arrows indicate the direction of the movements.

Note that these two angles are components of the angle δ.

Lastly, the three translations are calculated as matrix matrix multiplication of the displacement of the displacement between the two bases and the MBT:

$$[S_x S_y S_z] = (\vec{O}_I - \vec{O}_{II})T_{mbt} \tag{2.16}$$

With this process, the six BPP (see figure 2.2) that describe the geometry between two bases that compose a bp are calculated.

Before moving to the next subsection, the position of the MBT is calculated as the average position of the two bases:

$$\vec{O}_{mbt} = \frac{1}{2}\left(\vec{O}_I + \vec{O}_{II}\right) \tag{2.17}$$

This position vector describes the average position of a given bp, and will be needed for the calculation of the BSP of a double stranded structure.

2.1.3 CEHS scheme: base-step parameters (BSP)

Similarly to the BPP, the base-step parameters (BSP) are composed of six parameters that give a complete description of the geometry between two consecutive bp. Figure 2.3 shows a visual representation of these six parameters, where the parameters shift (D_x), slide (D_y) and rise (D_z) describe the displacements in 3 directions (x, y, z) and the angles tilt (τ), roll (ρ) and twist (Ω) describe rotations around the same 3 axes. Similar to the BPP, a mid-step triad (MST) is constructed for each pair of consecutive bp with their respective axes (x, y, z).

In contrast of BPP, BSP can be calculated for single and double stranded structures and, either way, a vector O_i and a matrix R_i are needed for describing the position and orientation of each base, respectively. In case of single stranded structures, O_i and R_i correspond to the position vector and orientation matrix obtained from the base fitting process (see subsection 2.1.1 and equations 2.4 & 2.3), and in case of double stranded structures, the positions O_i and orientations R_i are described by the position and orientation of the calculated MBT (see subsection 2.1.3 and equations 2.11 & 2.17). For this subsection, we will consider the case of double-stranded DNA and will be referring to the x, y, z axes of bp i as $\hat{x}_i, \hat{y}_i, \hat{z}_i$ (column vectors of R_i) respectively.

The first step for calculating the BSP is to obtain the bending angle θ, which is defined as the angle between the vectors that are tangent to the curve at position i:

$$\theta_i = \arccos(\hat{z}_i \cdot \hat{z}_{i+1}) \tag{2.18}$$

Then, the roll-tilt axis ($\hat{rt}$) which is perpendicular to these tangent vectors is calculated:

$$\hat{rt} = \hat{z}_i \times \hat{z}_{i+1} \tag{2.19}$$

Then, R_i and R_{i+1} are rotated around $\hat{rt}$ by $+\theta/2$ and $-\theta/2$:

$$T_i = R_{rt}\left(+\frac{\theta}{2}\right) R_i \tag{2.20}$$

$$T_{i+1} = R_{rt}\left(-\frac{\theta}{2}\right) R_{i+1} \tag{2.21}$$

where $R_{rt}(+\theta/2)$ and $R_{rt}(-\theta/2)$ are rotation matrices. The new rotated triads (T_i & T_{i+1}) are formed by the $\hat{x}'_i, \hat{y}'_i, \hat{z}'_i$ axes, being the latter aligned.

Then, the MST is calculated as the average of the two triads:

$$T_{mst} = \frac{1}{2}(T_i + T_{i+1}) \tag{2.22}$$

The x axis ($\hat{x}_{mst}$) of the MST points towards the major groove, while the y axis ($\hat{y}_{mst}$) points towards the backbone of the first strand and the z axis ($\hat{z}_{mst}$) points towards the molecular axis in the particular mid-step.

With the MST obtained, we can now calculate twist Ω as the angle between the y axes (or x axes) of the triads T:

$$\begin{aligned} \Omega &= \arccos(\hat{y}'_i \cdot \hat{y}'_{i+1}) \\ if\ \ (\hat{y}'_i \times \hat{y}'_{i+1}) \cdot \hat{z}_{mst} &< 0, \quad \Omega < 0 \\ if\ \ (\hat{y}'_i \times \hat{y}'_{i+1}) \cdot \hat{z}_{mst} &> 0, \quad \Omega > 0 \end{aligned} \tag{2.23}$$

To calculate the remaining roll and tilt angles, we need to first calculate the angle φ between the roll-tilt axis $\hat{rt}$ and the $\hat{y}_{mst}$:

$$\begin{aligned} \varphi &= \arccos(\hat{rt} \cdot \hat{y}_{mst}) \\ if \ \ (\hat{rt} \times \hat{y}_{mst}) \cdot \hat{z}_{mst} &< 0, \quad \varphi < 0 \\ if \ \ (\hat{rt} \times \hat{y}_{mst}) \cdot \hat{z}_{mst} &> 0, \quad \varphi > 0 \end{aligned} \tag{2.24}$$

Then, roll ρ and tilt τ angles are calculated as components of the bending angle θ:

$$\tau = \theta \sin(\varphi) \tag{2.25}$$

$$\rho = \theta \cos(\varphi) \tag{2.26}$$

Here, roll indicates the bending component that affects the major groove while tilt indicates the bending component towards the backbone.

Similarly to BPP, the three translation parameters are calculated by a matrix multiplication of the displacement vector between the two consecutive base-pairs and the MST:

$$[D_x D_y D_z] = (\vec{O}_{i+1} - \vec{O}_i) T_{mst} \tag{2.27}$$

Lastly, the position of the MST is calculated as an average of positions of the bp i and bp $i+1$:

$$\vec{O}_{mst} = \frac{1}{2}\left(\vec{O}_i + \vec{O}_{i+1}\right) \tag{2.28}$$

This is the process followed by the 3DNA program for calculating the six BSP that describe the geometry of two consecutive bases/bp.

2.1.4 CEHS scheme: rebuilding algorithm from the BSP

The mathematical process for obtaining the BSP is completely reversible. Thus, given a set of $(N-1)$ BSP that describes the geometry of a molecule made with N bp, the following "rebuilding algorithm" is applied to obtain the N vectors $\vec{O}_i$ and N orientation matrices R_i that represent the respective position and orientation of each base-pair i.

To rebuild the structure, the origin and orientation of the first bp needs to be initialized [90, 92]. Usually, $\vec{O}_1$ is placed at the origin and R_1 is made parallel to the global reference frame (x, y, z):

$$\begin{aligned} \vec{O}_1 &= 0 \\ R_1 = \mathbb{I} &= [\hat{x}, \hat{y}, \hat{z}] \end{aligned} \tag{2.29}$$

Then, following with bp $i = 1, 2, 3, ..., N-1$, the procedure is to iteratively describe the position $\vec{O}_{i+1}$ and orientation R_{i+1} of bp $i+1$ with respect the current bp i. Thus, given the BSP $[D_x, D_y, D_z, \tau, \rho, \Omega]$ that describe the geometry between bp i and bp $i+1$, the bending angle is calculated as:

$$\theta = \sqrt{\rho^2 + \tau^2} \tag{2.30}$$

Then, the RollTilt axis is equal to:

$$\hat{rt} = \frac{\rho}{\theta}\hat{y}_{mst} + \frac{\tau}{\theta}\hat{x}_{mst} \tag{2.31}$$

Here, the MST between i and $i+1$ is unknown but we can suppose it is equal to the identity matrix $T_{mst} = \mathbb{I} = [\hat{x}, \hat{y}, \hat{z}]$. This supposition is valid because at the moment we are only interested in the magnitude of the bending angle θ and the angle φ between the RollTilt axis $\hat{rt}$ and the MST y-axis $\hat{y}_{mst} = \hat{y}$:

$$\begin{aligned} \varphi &= \arccos(\hat{rt} \cdot \hat{y}_{mst}) \\ if\ \ (\hat{rt} \times \hat{y}_{mst}) \cdot \hat{z}_{mst} &< 0, \quad \varphi < 0 \\ if\ \ (\hat{rt} \times \hat{y}_{mst}) \cdot \hat{z}_{mst} &> 0, \quad \varphi > 0 \end{aligned} \tag{2.32}$$

Then the MST, the orientation and the position of bp $i+1$ with respect bp i are calculated as:

$$\begin{aligned} T'_{mst} &= R_z\left(\frac{\Omega}{2} - \varphi\right) R_y\left(\frac{\theta}{2}\right) R_z\left(\varphi\right) \\ R'_{i+1} &= R_z\left(\frac{\Omega}{2} - \varphi\right) R_y\left(\theta\right) R_z\left(\frac{\Omega}{2} + \varphi\right) \\ \vec{O}'_{i+1} &= \left[D_x D_y D_z\right]\left[T'_{mst}\right]^T \end{aligned} \tag{2.33}$$

Where $R_u(\alpha)$ are general rotation matrices around the unit vector u with angle α. Lastly, the triad and position of bp $i+1$ with respect the global reference frame are calculated as:

$$\begin{aligned} R_{i+1} &= R_i R'_{i+1} \\ \vec{O}_{i+1} &= \vec{O}_i + \vec{O}_{i+1} R_i^T \end{aligned} \tag{2.34}$$

This process can be applied to either double or single stranded structures. In case of double stranded structures, the calculated positions and triads are equivalent to the position and orientation of the MBT (see subsection 2.1.2 and equations 2.17 & 2.11), whereas, in case of single stranded structures, the resulted vectors and matrices correspond to the fitted bases (see subsection 2.1.1 and equations 2.4 & 2.3). Once the positions and triads of the MBTs are obtained, an analogous process can be implemented to obtain the positions and orientations of the nucleotides that compose each base-pair (see [90] & [92] for more details of the process).

2.2 The Length-Dependence Elastic Model (LDEM)

The Length-Dependence Elastic Model (LDEM) [114] further extends the CEHS scheme to describe the structure of NA at lengths beyond the dinucleotide level (beyond the nearest neighbor description). Its methodology is particularly useful for describing the average structural and elastic properties of DNA, and can translate these properties from a local to a global perspective, where bulk parameters are obtained for evaluating the molecule overall flexibility. The LDEM is also useful for visualizing how the structural/elastic parameters evolve as the length increases. The flexibility of DNA is evaluated in terms of the elastic parameters, which correspond to the stretch modulus, twist elastic constant and the persistence length. These elastic properties emerge from

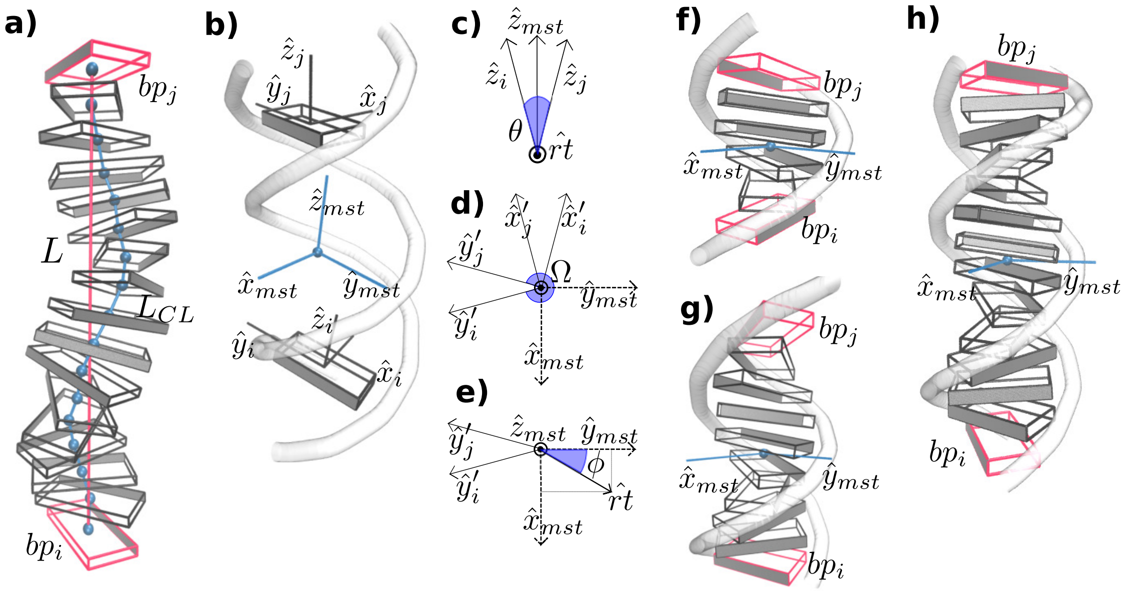


Figure 2.4: Visual representation of the implementation of the LDEM by the *SerraNA* program to calculate the structural parameters at different length scales. Here, transparent rectangular blocks represent base-pairs, and their shaded sides represent the minor groove. (a) The displacement between bp i and j, is characterised by the end-to-end distance (red) and the contour length (blue). (b) A mid-base triad (T$_{mst}$) positioned between bp i and bp j, is calculated to measure the twist and bend angles between both bp. (c) The unit vectors $\hat{z}_i$ and $\hat{z}_j$ define the bending angle θ (blue) and the RollTilt axis $\hat{rt}$. (d) The $\hat{x}'$ and $\hat{y}'$ directions lie in the $\hat{x}_{mst} - \hat{y}_{mst}$ plane, where the twist angle Ω (blue) is defined as the angle between $\hat{y}'_j$ and $\hat{y}'_i$. and $\hat{y}_{mst}$ points towards half Ω. (e) The RollTilt axis $\hat{rt}$ also lies in the $\hat{x}_{mst} - \hat{y}_{mst}$ plane, and the angle φ (blue) is used to define the bending components, tilt and roll. (f)-(h) Mid-base triads T$_{mst}$ calculated from bp (red) at the legths of $l = 5, 9, 13$, respectively. Note that the roll component points towards the grooves, while tilt points towards the backbone a the mid-point. This image was taken from [157].

bp fluctuations sampled by numerical simulations of DNA. While the LDEM can also calculate the elastic couplings, it does not provide a methodology for estimating their bulk value.

In this section, the methodology of the LDEM is provided in detail. We begin for describing the process to obtain the structural variables that quantify the geometry of NA molecules at longer lengths. Then, we move on to explain how these structural variables are used to calculate elastic parameters. We then describe the process for evaluating the global flexibility in terms of the stretch modulus, twist modulus and bending persistence lengths.

2.2.1 Geometry between bp at longer lengths

Similar to the CEHS scheme, the LDEM calculates mid-step triads to describe the geometry between bp separated by an increasing number of nucleotides (see figure 2.4b). To this end, the LDEM uses a set of 9 structural parameters: added shift, added slide, added rise, tilt, roll, twist, bending angle, contour length and the end-to-end distance [157]. The LDEM adapts the twist parameter to measure angles bigger than a helical DNA turn as the length increases (figure 2.4d). This directly affects the bending angle components of roll and tilt, since their relative orientation is now dependant of the twist parameter (figure 2.4e). The end-to-end distance is mainly affected by vertical displacements and is associated with stretching deformations, while the contour length is included to provide a more complete structural description of the molecule (see figure 2.4a). The added parameters can be interpreted as pseudo-components of the contour length, and at the dinucleotide level they are equivalent to the translation parameters of the BSP. One of the limitations of the algorithm is that, in contrast to the twist angle, the bending angle cannot capture angles greater than 180 degrees. This limitation will be discussed in detail later in this sub-section.

The LDEM describes the geometry between base-pair i and base-pair j by using the position vectors $\vec{O}_i$ and $\vec{O}_j$ and the orientation matrices R_i and R_j with column vectors $[\hat{x}_i, \hat{y}_i, \hat{z}_i]$ & $[\hat{x}_j, \hat{y}_j, \hat{z}_j]$ respectively. Here, $j > i$ and j is associated with the length (l) between both bp as $j = i + l$. When $l = 1$ bp, the process is reduced to the dinucleotide level which is equivalent to the BSP process (see sub-section 2.1.3). Notice that the LDEM can be applied to either double- or single-stranded structures. For double-stranded structures, the process starts after the BPP process and $\vec{O}_i/R_i$ correspond to the MBT position and orientation respectively (see sub-section 2.1.2). For single-stranded structures, the process begins right after the bases are fitted (see section 2.1.1).

Then, the first structural parameter to be calculated is the bending angle $\theta_{i,j}$ between bp i and j (see figure 2.4b):

$$\theta_{i,j} = \arccos(\hat{z}_i \cdot \hat{z}_j) \tag{2.35}$$

And similar to the BSP process, the RollTilt axis $\hat{rt}$ is then calculated as:

$$\hat{rt} = \hat{z}_i \times \hat{z}_j \tag{2.36}$$

With this, we can rotate the triads R that indicate the orientation of bp i and j, around the RollTilt axis by half of the bending angle:

$$T_i = R_{rt}\left(+\frac{\theta_{i,j}}{2}\right) R_i \tag{2.37}$$

$$T_j = R_{rt}\left(-\frac{\theta_{i,j}}{2}\right) R_j \tag{2.38}$$

We denote the column vectors of the new triad T_i as $[\hat{x}'_i, \hat{y}'_i, \hat{z}'_i]$, which indicates its respective orientation. If we were to follow the CEHS scheme [91], we would obtain the MST by averaging both triads. However, the LDEM first needs to estimate a provisional twist angle (Ω') that will support the evaluation of the correct twist (Ω) when it is greater than 180 degrees. To this end, Ω' is obtained as:

$$\begin{aligned} \Omega'_{i,j} &= \arccos(\hat{y}'_i \cdot \hat{y}'_j) \\ if\ \ (\hat{y}'_i \times \hat{y}'_j) \cdot \hat{z}_{mst} &< 0, \ \ \Omega'_{i,j} < 0 \\ if\ \ (\hat{y}'_i \times \hat{y}'_j) \cdot \hat{z}_{mst} &> 0, \ \ \Omega'_{i,j} > 0 \end{aligned} \tag{2.39}$$

The twist angle $\Omega_{i,j}$ is calculated by summing the provisional twist angle with the number of turns N that were covered by the previous twist angle $\Omega_{i,j-1}$:

$$\Omega_{i,j} = \Omega'_{i,j} + N2\pi \tag{2.40}$$

After, we can calculate how many half turns N' the molecule has covered from bp i to bp j:

$$N' = \frac{\Omega_{i,j} + \pi}{2\pi} \tag{2.41}$$

Notice that N' is an integer, and now we can use it to calculate the rest of the MST components:

$$\hat{x}_{mst} = \frac{1}{2}\left(\hat{x}_i + \hat{x}_j\right)(-1)^{N'} \tag{2.42}$$

$$\hat{y}_{mst} = \frac{1}{2}\left(\hat{y}_i + \hat{y}_j\right)(-1)^{N'} \tag{2.43}$$

This operation is equivalent to rotating the x and y axes by half of the twist angle. Figure 2.4d shows a visual representation of the twist angle calculation, where the number of half turns between bp i and j is being considered, and the $\hat{x}_{mst}/\hat{y}_{mst}$ point towards half of the twist angle. Notice that the column vectors of the T_{mst} are $[\hat{x}_{mst}, \hat{y}_{mst}, \hat{z}_{mst}]$.

The position of the MST is then calculated as an average of the triads of bp i and j:

$$\vec{O}_{mst} = \frac{1}{2}\left(\vec{O}_i + \vec{O}_j\right) \tag{2.44}$$

And similarly to the BSP process, the angle φ between the RollTilt axis and the y component of the MST (see figure 2.4e), is calculated as:

$$\begin{gathered} \varphi = \arccos(\hat{rt} \cdot \hat{y}_{mst}) \\ if\ \ (\hat{rt} \times \hat{y}_{mst}) \cdot \hat{z}_{mst}\ < 0, \quad \varphi < 0 \\ if\ \ (\hat{rt} \times \hat{y}_{mst}) \cdot \hat{z}_{mst}\ > 0, \quad \varphi > 0 \end{gathered} \tag{2.45}$$

This angle φ is indirectly connected to the twist angle due to the fact that the direction of $\hat{y}_{mst}$ depends on the twist angle, and the components of the bending angle, tilt and roll are calculated through φ:

$$\tau_{i,j} = \theta_{i,j} \sin(\varphi) \tag{2.46}$$

$$\rho_{i,j} = \theta_{i,j} \cos(\varphi) \tag{2.47}$$

Roll and tilt angles point towards the grooves and backbone, respectively at the position of the MST which is the molecular mid-point. Panels (f)-(h) of figure 2.4 show examples of the MST $x - y$ components as well as positions, calculated from bp separated by 4, 8 and 12 bp, respectively.

The rest of the translation variables in the BSP, are extended at longer lengths as:

$$[X_{i,j}\ Y_{i,j}\ Z_{i,j}] = \sum_{i}^{j-1}[X_i\ Y_i\ Z_i] \tag{2.48}$$

We refer to these three parameters as the added shift, added slide and added rise, respectively. In the case of a double-stranded structure, the parameters $[X_i\, Y_i\, Z_i] = [D_x\, D_y\, D_z]$ correspond to shift, slide and rise parameters of bp i (see equation 2.27), while in the case of a single stranded structure $[X_i\, Y_i\, Z_i] = [S_x\, S_y\, S_z]$ correspond to shear, stretch and stagger parameters of base i (see equation 2.16).

Then, the end-to-end distance $L_{i,j}$ between bp i and j is simply calculated as the distance between bp i and j:

$$L = \left|\vec{O}_j - \vec{O}_i\right| \tag{2.49}$$

While the contour length is calculated as the sum of all the consecutive distances that connect bp i to j:

$$L_{i,j}^{CL} = \sum_{k=i}^{j-1}\left|\vec{O}_{k+1} - \vec{O}_k\right| \tag{2.50}$$

The added parameters $[X_{i,j}\ Y_{i,j}\ Z_{i,j}]$ can be interpreted as the three pseudo components of the contour length. Panel a) of figure 2.4 shows an ilustration of the end-to-end distance (red line) and contour length (curved blue line) calculated between the two bp represented as red blocks.

One of the limitations of the LDEM is that it cannot accurately measure rotation parameters (ρ, τ & Ω) at lengths in which $\theta_{i,j} \geq 180°$. In these cases, the RollTilt axis $\hat{rt}$ points towards the wrong direction, which results in an incorrect calculation of twist, roll and tilt.

2.2.2 The length-dependent elasticity model of DNA

Given a structural variable X, uncorrelated to other variables and which distribution of values spawns a Gaussian distribution, the corresponding elastic constant K can be calculated from its variance $Var(X)$ [114, 116]:

$$K = k_B T b l \frac{1}{Var(X)} \tag{2.51}$$

Here, $b = 0.34$ nm and corresponds to the average bp rise in B-DNA [140], and l specifies the length of the oligomer in base-steps, which ranges from 1 to N-1 bp, with N being the number of bp in the molecule. Notice that bl indicates the length of the oligomer in nm.

In the LDEM, the flexibility of NA is evaluated by four elastic parameters. The stretch modulus measures the resistance to stretching deformations and is evaluated through the variance of the end-to-end distance L (see figure 2.5b). The twist modulus quantifies the resistance to deformations along the twist angle Ω (figure 2.5c), while the bending deformations are quantified through the two bending components, tilt τ and

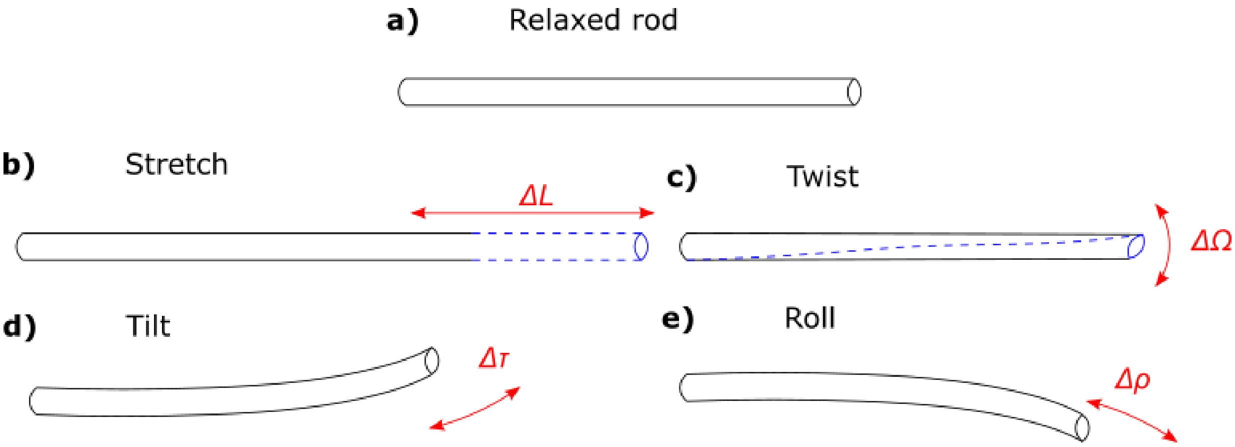


Figure 2.5: An elastic rod representing the four type of deformations in the LDEM. The stretching, twist, and two bending deformations of the rod, are characterised by changes in the structural parameters of the end-to-end distance ΔL, twist angle $\Delta\Omega$, tilt angle $\Delta\tau$ and roll angle $\Delta\rho$, respectively. Structural changes of the rod are colored in blue, while the direction of the deformations are drawn as red arrows.

roll ρ to account for the bending anisotropy [98] (see figure 2.5d-e). However, these four structural variables (L, Ω, ρ, τ) are non-orthogonal. The LDEM addresses this issue by calculating the covariance matrix V of the four structural parameters, then replacing the fraction in equation 2.51 with the inverse of the covariance matrix V^{-1} to obtain the elastic matrix F [118]:

$$F = k_B T b l V^{-1} \tag{2.52}$$

The reasoning behind this approach, is that the diagonal components of V^{-1} are equivalent to the reciprocal of the partial variances $1/Var_p(X)$, and the partial variances $Var_p(X)$ measure the residual variance of X after removing the linear effects caused by the other variables in V [114, 166]. Hence, the diagonal elements of F correspond to the four elastic constants without the contributions of other variables, while the non-diagonal elements correspond to the coupling terms between the elastic variables.

Therefore, for every sub-fragment k composed of $l+1$ bp, an elastic matrix $F_{k,l}$ is calculated:

$$F_{k,l} = \begin{bmatrix} B_k & D_k & H_k & A_{\tau,k}/B_k \\ D_k & C_k & G_k & A_{\tau,k}/C_k \\ H_k & G_k & A_{\rho,k} & A_{\tau,k}/A_{\rho,k} \\ A_{\tau,k}/B_k & A_{\tau,k}/C_k & A_{\tau,k}/A_{\rho,k} & A_{\tau,k} \end{bmatrix} \tag{2.53}$$

Notice that each element of matrix $F_{k,l}$ is calculated with equation 2.52 and derived from thermal fluctuations of bp (the variance and covariances of structural variables).

Here, the diagonal components correspond to the stretch modulus B, twist elastic constant C, roll and tilt elastic constants A_ρ & A_τ; and the off-diagonal components correspond to the elastic couplings being twist-stretch D, roll-stretch H, twist-roll G,

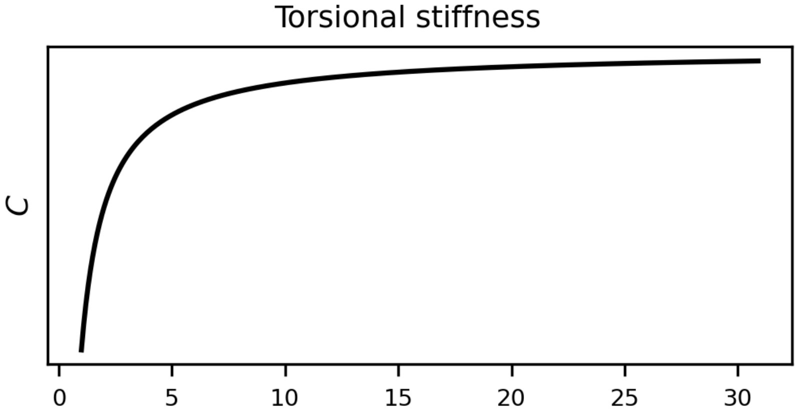


Figure 2.6: Representation of the twist modulus C as a function of length as observed in [114]. The curve reaches a plateau around 11 bp (oligomer composed of 12 bp) which is approximately the length in which a DNA turn is completed.

tilt-stretch A_τ/B, tilt-twist A_τ/C and tilt-roll A_τ/A_ρ.

Each of these elastic parameters highly depend on the temperature at which the simulation production time was executed (see equation 2.52), and in the MD simulation conditions such as ion atmosphere [9,52,169], force fields [63] (parametrization of MD simulation) and the overall quality of the MD simulation.

One of the limitations of the LDEM is that the validity of the model can only be achieved when the analysed DNA is weakly deformed as analysing systems in which the DNA is severely deformed might yield untrustworthy flexibility estimations as the system might not comply with the harmonic approximation [118] (see equations 2.51 & 2.52).

Lastly, notice that the matrices V and F are real symmetric matrices.

2.2.3 Estimation of the twist elastic constant

The twist elastic constant $C_{k,l}$ of a particular sub-fragment k, corresponds to a diagonal element of matrix $F_{k,l}$. The LDEM calculates the twist elastic modulus C_l as a function of length l as an average of all sub-fragments k composed of the same number of bp $l+1$:

$$C_l = \frac{1}{n_k} \sum_{k=1}^{n_k} C_k \tag{2.54}$$

where n_k is the number of sub-fragments composed of l+1 bp. Then, the bulk elastic constant C that globally quantifies the molecule resistance to torsional deformations is calculated as:

$$C = \sum_{l=N_a}^{N_b} \frac{C_l}{N_b - N_a} \tag{2.55}$$

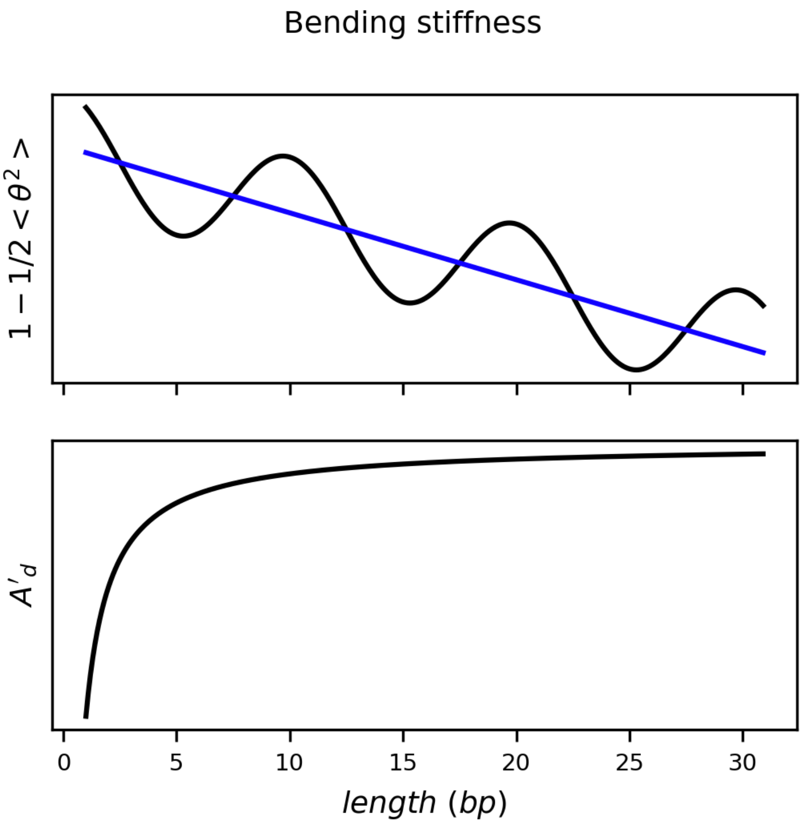


Figure 2.7: Ideal representations of the curves observed in [114] to obtain estimations of the persistence length. (Top panel) Linear fits of equations 2.57 2.58 2.59, where the black line corresponds to the directional decay of tangent-tangent correlations, and the blue line represents the linear fit. (Bottom panel) Second estimation of the dynamic persistence length A'_d as a function of length. The curve reaches a plateau around 11 bp (oligomer composed of 12bp).

where $N_a < N_b \leq N$. The LDEM suggests that the length N_a should be of at least 11bp ($N_a = 11$), which corresponds to the length of oligomers composed of 12bp. The reasoning behind this length selection is because in the study [114], they observed that the transition from local to global elastic behaviour occurs within one helix turn, then for longer lengths the molecule would behave as an elastic rod with the twist stiffness constant. Hence only oligomers of at least 12bp of length should be considered to capture most of the bulk behaviour (see figure 2.6). Notice that the analysis in [114] was performed for naked DNA molecules, hence equation 2.55 with $N_a = 11$ is ensured to work for free DNA.

2.2.4 Estimation of the bending persistence length

The LDEM estimates the persistence length using two methods. One of the methods consists in obtaining the bending persistence length A as a linear fit of the directional decay. According to the worm-like chain model (WLC) [153], the orientation of the polymer is quantified through the directional correlation of two tangent vectors separated by a discrete length bl, which decays exponentially with decay constant $1/A$:

$$< \hat{z}_i \cdot \hat{z}_j >=< \cos\theta_{i,j} >= \exp\left(\frac{-bl}{A}\right) \tag{2.56}$$

Where $b = 0.34$ nm is the average bp rise in B-DNA, and l is the oligomer length in bp units. Then, expanding the cosine component of the equation into a Maclauren series of degree two and expanding the exponential component as a power series of degree one, we obtain the following equivalency [114]:

$$1 - \frac{1}{2}\left\langle\theta_{i,j}^2\right\rangle \equiv 1 - \frac{bl}{A} \tag{2.57}$$

This equation is a suitable approximation when applied to sub-fragments shorter than the DNA persistence length, since at greater lengths, the tangent-tangent correlations would decay in a exponential form rather than linear.

It was previously observed that local tangent correlations do not decay uniformly as the length increases, and that they present oscillations with a period of approximately one helical turn (see top panel of figure 2.7) [114]. In fact, it was observed that the directional memory is lost faster for oligomers with lengths between n and $n + \frac{1}{2}$ turns than oligomers with lengths between $n + \frac{1}{2}$ and $n + 1$ where they behave as stiffer polymers. The LDEM is capable of obtaining the persistence length A through a linear fit of equation 2.57 where the periodicity is filtered out (see figure 2.7).

The persistence length A from equation 2.56 is usually referred to as the "apparent persistence length", as it can be partitioned into two quantities: the static persistence length A_s originated from the DNA intrinsic shape, and the dynamic persistence length A_d caused by the DNA stiffness [104]. Following this idea, the quantity $< \theta^2 >$ has a static and dynamic contribution, which are linked through $< \theta^2 >=< \theta_s^2 > + < \theta_d^2 >$, being $< \theta_s^2 >$ caused from random sequence-dependant static bends and $< \theta_d^2 >$ is originated from thermal fluctuations. The static bends $< \theta_s^2 >$ can be obtained through the average structure while the thermal components $< \theta_d^2 >$ can be obtained after subtracting the static component $< \theta_d^2 >=< \theta_s^2 > - < \theta^2 >$. Putting these terms in equation 2.57, the static and dynamic components of the bending persistence length are obtained through linear fits of the following expressions:

$$1 - \frac{1}{2}\left\langle\theta_s^2\right\rangle \equiv 1 - \frac{bl}{A_s} \tag{2.58}$$

$$1 - \frac{1}{2}\left\langle\theta_d^2\right\rangle \equiv 1 - \frac{bl}{A_d} \tag{2.59}$$

The top panel of figure 2.7 shows an ideal representation of the linear fits of equations 2.57 2.58 2.59 as observed in [114], where the black line represents the left hand sides of the equations, and the blue line represents the fitted line.

Once that the static and dynamic persistence lengths have been obtained, they can be mixed into A with [104]:

$$\frac{1}{A} = \frac{1}{A_s} + \frac{1}{A_d} \tag{2.60}$$

The resulting A, should be compatible with the persistence length obtained through the linear fit of equation 2.57.

The second method for estimating the persistence length relies on the inverse-covariance analysis, which provides the opportunity of calculating a second estimation of the dynamic persistence length A'_d. The tilt ($A_{\tau,k}$) and roll ($A_{\rho,k}$) elastic constants can be combined to obtain the dynamic persistence length of fragment k:

$$\frac{1}{A'_{d,k}} = \frac{1}{2}\left(\frac{1}{A_{\tau,k}} + \frac{1}{A_{\rho,k}}\right) \tag{2.61}$$

Then, the dynamic persistence length at length l can be obtained through an average of all oligomers k with same length l:

$$A'_{d,l} = \frac{1}{n_k}\sum_{k=1}^{n_k} A'_{d,k} \tag{2.62}$$

Finally, the second estimation of the dynamic persistence length is obtained through an average from length N_a to N_b:

$$A'_d = \sum_{l=N_a}^{N_b} \frac{A'_{d,l}}{N_b - N_a} \tag{2.63}$$

The bottom panel of figure 2.7 represents the ideal curve that spawns A'_d as a function of length. Similar to the twist curve (figure 2.6), A'_d presents a plateau at $l = 11$bp indicating that the bulk behaviour has been reached. Due to this fact, the LDEM suggest to set $N_a = 11$bp. Since this method only takes into account the partial variances of tilt and roll ($1/Var_p(\tau)$ and $1/Var_p(\rho)$), the linear effects of twist and the end-to-end distance (Ω and L) are removed resulting in a stiffer estimation of the persistence length compared to the one obtained through the linear fit ($A'_d > A_d$).

The persistence length A, and both its static A_s and dynamic A_d contributions can be easily estimated from ensembles obtained by numerical simulations and using equations 2.56, 2.60 & 2.63. The persistence length A can be extracted experimentally with the WLC model (from equation 2.56) using single molecule techniques such as cryo-EM [10], AFM [103] and optical/magnetic tweezers [9, 56, 85, 165], and is estimated to be around 50nm. However, it is experimentally challenging to estimate the dynamic A_d/A'_d and the static A_s contributions as in vitro experiments suffer from limited sampling. The dynamic persistence length have been estimated experimentally from intrinsically straight DNA sequences, where the apparent persistence length A should be approximately equal to the dynamic persistence length A_d (because $1/A_s$ from equation 2.60 is small) [10]. Nonetheless, cryo-EM experiments in combination with MC simulations have managed to estimate the three quantities of a natural DNA sequence (phage lambda DNA), with $A = 45nm$, $A_s = 130nm$ and $A_d = 80nm$ [10].

Lastly, bulk elastic constants of tilt A_τ and roll A_ρ can be similarly obtained through averages:

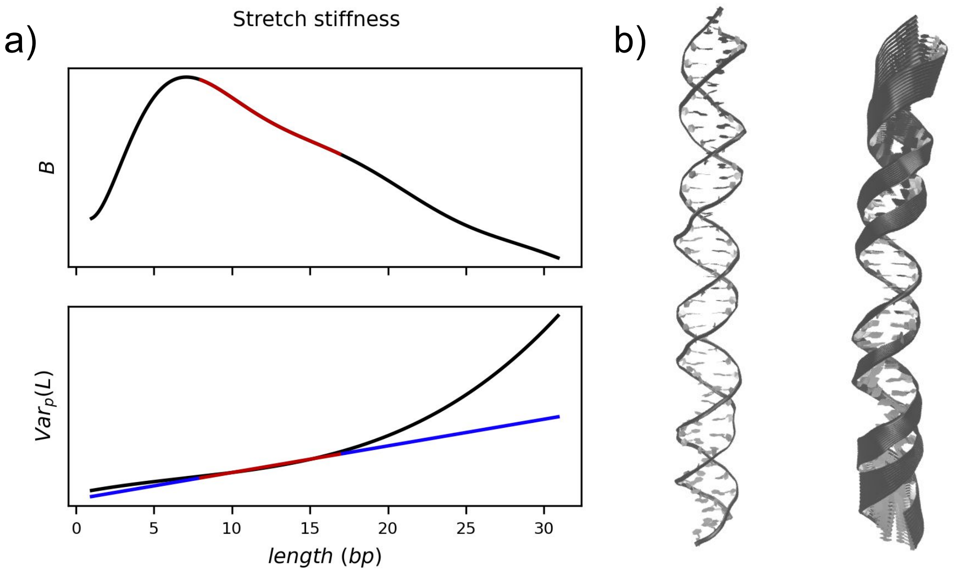


Figure 2.8: (a) The top graph shows the stretch modulus B as a function of length, where the red region indicates the length of oligomers that are included in the global stretch calculation obtained through the linear fit of the partial variance of the end-to-end distance $Var_p(L)$ shown in the bottom graph. (b) On the left the average structure of a 42bp DNA and on the right the stretching mode that causes end-effects resulting in an apparent softening in the stretch modulus curve.

$$A_\tau = \sum_{l=N_a}^{N_b} \frac{A_{\tau,l}}{N_b - N_a} \tag{2.64}$$

$$A_\rho = \sum_{l=N_a}^{N_b} \frac{A_{\rho,l}}{N_b - N_a} \tag{2.65}$$

Where $A_{\tau,l}$ and $A_{\rho,l}$ have previously been calculated by an analogous of equation 2.62.

As stated by Marko and Siggia [98], we can only consider the asymmetry between the minor and major grooves in the calculation of the dynamic persistence length through the introduction of the twist-roll coupling G in equation 2.61 [111, 142]:

$$\frac{1}{A''_{d,k}} = \frac{1}{2}\left(\frac{1}{A_{\tau,k}} + \frac{1}{A_{\rho,k} - G_k^2/C_k}\right) \tag{2.66}$$

where $A''_{d,k}$ is a estimation of the persistence length of oligomer k where the effect of twist has been specifically removed.

2.2.5 Estimation of stretch modulus

Similar to the twist, roll and tilt elastic constants, the stretch modulus B at length l is calculated as an average over oligomers made of $l + 1$ bp:

$$B_l = \frac{1}{n_k} \sum_{k=1}^{n_k} B_k \tag{2.67}$$

In contrast to the twist elastic constant C and the second estimation of the dynamic persistence length A'_d (see figures 2.6 & 2.7), the stretch modulus B does not tend to a plateau. Instead, the stretch modulus follows a complex behaviour (see the top graph of figure 2.8a): it first hardens due to the prevalence of strong base stacking interactions on oligomers shorter than 9bp, then it presents a first softening due to the coordinated motion of base pairs in oligomers of 6-11bp of length, and then it presents a second softening due to extended end-effects that act on at least one helical turn of each end (see figure 2.8b) [114]. To overcome these difficulties, the LDEM relies on the partial variance of the end-to-end distance ($Var_p(L)$) as a function of length. To filter stacking interactions and end-effects, central oligomers longer than 9bp are considered in the linear fit of $Var_p(L)$ (see bottom panel of figure 2.8a). $Var_p(L)$ increases exponentially as the length increases, indicating that the stretch modulus would be close to zero at longer lengths, but the fit forces a plateau in which the stretch modulus would converge after filtering the stacking interactions and removing the end-effects. Thus, through the linear fit of the partial variance, the global stretch modulus is obtained.

2.2.6 Error estimation and confidence levels of elastic constants

The processes described in this section are not part of the LDEM [114] but they are implemented in the SerraNA program (see results 4) to estimate errors and establish confidence intervals of the calculated DNA elastic constants. Given a random variable x, with multiple n observations $x_1, x_2, x_3, ...x_n$, the standard error on the mean s_x is calculated as [3]:

$$s_x = \frac{\sigma_x}{\sqrt{n}} \tag{2.68}$$

where σ_x is the standard deviation and $\bar{x}$ is the mean of variable x. The standard error on the mean s_x also known as the standard error, which tells us how accurate is the estimation of the mean $\bar{x}$.

Returning to the LDEM model, the global elastic constants of twist C, tilt A_τ, roll A_ρ and the dynamic persistence length A'_d & A''_d are calculated through averages over a range of lengths (from equations 2.55, 2.64, 2.65, 2.63 & 2.66, respectively), hence to quantify the accuracy of these estimations we use equation 2.68 to obtain the corresponding standard errors, where the sample size is the range of lengths in which the means were obtained $n = N_b - N_a$. Notice that these ranges might differ between elastic constants.

It is worth mentioning that the cause of randomness considered in the structural parameters are due to thermal fluctuations, however, the number of random observations taken for the estimation of bulk elastic constants are elastic constants measured at distinct lengths l. Hence, the estimated bulk elastic constants are elastic constants averaged across multiple length-scales determined by the range $[N_a, N_b]$. These ranges must be carefully chosen as the elastic constants mainly present two distinct behaviours:

one at short length scales $l < 11bp$ and the other one when bulk flexibility is reached at $l \geq 11bp$ (see the torsional stiffness C as an example 2.6). Because the magnitude of elastic constants at lengths in which bulk flexibility is reached is so similar ($l \geq 11bp$), it is recommended that the range $[N_a, N_b]$ is set within this behaviour. Calculating elastic constants over lengths which cover both length scales would result in a greater standard error s_x, which would indicate a low accuracy in the estimated elastic constant.

For the parameters obtained through a linear fit, we need to follow a different procedure to estimate their accuracy. Given n data pairs of observations (x_i, y_i), their relationship can be described by the following simple model [18]:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2.69}$$

where α and β correspond to the intercept and slope of a line $y = \alpha + \beta x$ and ϵ_i corresponds to the error/residual component, which in other words is the distance between the line and the coordinate (x_i, y_i). In a simple linear regression or linear fit, the objective is to estimate the parameters α and β that provide the best fit, so the fitted line minimizes the sum of squared residuals. Implementing a least-squares approach, we construct the following objective function [128]:

$$Z(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{2.70}$$

Then, the objective is to minimize Z by varying both two parameters ($\partial Z/\partial\alpha = 0$ & $\partial Z/\partial\beta = 0$). The solutions to this minimization problem is:

$$\alpha = \bar{y} - \beta\bar{x} \tag{2.71}$$

$$\beta = \frac{Cov(x, y)}{V(x)} \tag{2.72}$$

where $\bar{x}$ & $\bar{y}$ are averages, $V(x)$ is the variance of x and $Cov(x, y)$ is the co-variance of x and y.

To measure the accuracy of both estimated parameters, we can use confidence intervals [18]. Assuming that the residuals ϵ_i are normally distributed, we can construct the t-value t_{n-2} which has a Student's t-distribution with $n - 2$ degrees of freedom. This t-value is used to construct the following confidence intervals at confidence level $(1 - \gamma)$:

$$\alpha \in [\alpha - \Delta\alpha, \alpha + \Delta\alpha] \tag{2.73}$$

$$\beta \in [\beta - \Delta\beta, \beta + \Delta\beta] \tag{2.74}$$

where $\Delta\alpha = s_\alpha t^*_{n-2}$ & $\Delta\beta = s_\beta t^*_{n-2}$. Here, t^*_{n-2} is the $(1 - \gamma/2)$-th quantile of the Student's t-distribution t_{n-2}, and parameters s_α & s_β are the standard errors of the intercept α and slope β:

$$s_\alpha = s_\beta \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i} \tag{2.75}$$

$$s_\beta = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\epsilon_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{2.76}$$

Returning to the LDEM, the stretch modulus B is obtained through a linear fit of the partial variance of the end-to-end distance $Var_p(L)$, hence the partial variance as a function of length is approximated by:

$$Var_{p,l}(L) = \alpha_B + \beta_B l \tag{2.77}$$

where α_B is the intercept and β_B the slope. Assuming that the intercept is very close to zero ($\alpha_B \approx 0$) and using the diagonal component of equation 2.52 that corresponds to the end-to-end distance, we get the global stretch modulus:

$$B = \frac{k_B T b}{\beta_B} \tag{2.78}$$

To quantify the accuracy of B, the confidence interval of β_B cannot be directly used since the relation between B and β_B is not linear ($y = f(x)$). In other words the error does not propagate linearly and the uncertainty/error Δy can be approximated by [70]:

$$\Delta y = \frac{\partial f(x)}{\partial x}\Delta x \tag{2.79}$$

and then, the uncertainty of the stretch modulus is measured by the confidence interval:

$$\Delta B = \frac{k_B T b}{\beta_B^2}\Delta\beta_B \tag{2.80}$$

In the case of the estimation of the bending persistence length A as well as its static A_s and dynamic A_d contributions, the linear fit has an additional condition. At length zero ($l = 0$) the tangent vectors $\hat{z}$ are parallel, hence the directional correlation is exactly one $\hat{z} \cdot \hat{z} = 1$. Thus, the fit is forced to pass through the y-axis at 1, or in other words the intercept is set to $\alpha = 1$. Consequently, when minimizing the function Z the slope β of the linear model becomes:

$$\beta = \frac{\langle xy\rangle - \langle x\rangle}{\langle x^2\rangle} \tag{2.81}$$

where the symbols $\langle\rangle$ indicate averages. Then, equations 2.57, 2.58. 2.59 are approximated by the linear model:

$$1 - \frac{1}{2}\left\langle \theta_l^2 \right\rangle = 1 + \beta_A l = 1 - \frac{bl}{A} \tag{2.82}$$

$$1 - \frac{1}{2}\left\langle \theta_{s,l}^2 \right\rangle = 1 + \beta_{A_s} l = 1 - \frac{bl}{A_s} \tag{2.83}$$

$$1 - \frac{1}{2}\left\langle \theta_{d,l}^2 \right\rangle = 1 + \beta_{A_d} l = 1 - \frac{bl}{A_d} \tag{2.84}$$

Then, isolating the persistence lengths A, A_s, A_d we get:

$$A = \frac{-b}{\beta_A} \tag{2.85}$$

$$A_s = \frac{-b}{\beta_{A_s}} \tag{2.86}$$

$$A_d = \frac{-b}{\beta_{A_d}} \tag{2.87}$$

Similarly to the stretch modulus, once obtained the confidence intervals of the slopes $\Delta\beta$, the expression 2.79 is used to obtain the uncertainty in the estimations of the persistence lengths in terms of the confidence intervals:

$$\Delta A = \frac{b}{\beta_A^2}\Delta\beta_A \tag{2.88}$$

$$\Delta A_s = \frac{b}{\beta_{A_s^2}}\Delta\beta_{A_s} \tag{2.89}$$

$$\Delta A_d = \frac{b}{\beta_{A_d^2}}\Delta\beta_{A_d} \tag{2.90}$$

These are the methods for performing simple linear regressions and obtaining the corresponding confidence intervals to quantify the accuracy of the estimated parameters.

2.3 WrLINE Molecular Contour

The molecular contour of a NA molecule is described by a set of points that each represents a bp position. There are multiple definitions of the molecular contour, and softwares such as 3DNA [90, 91] and CURVES+ [76] are able to calculate the molecular contour from MD simulations of DNA. However, these molecular contours fail to capture the global shape of the molecule as they present a local periodicity that impairs measurements of writhe. The WrLINE method/software [150] offers a solution to this problem, where it filters out local irregularities by implementing a sliding-window averaged over individual DNA turns, resulting in a smooth molecular contour without the helical periodicity providing more accurate estimations of Writhe. WrLINE was developed by Sutthibutpong, Harris, and Noy published in [150], and it was first developed with the objective to provide better estimations of writhe. In this thesis project, the WrLINE molecular contour is utilized to calculate global measurements of DNA that can be compared with images from AFM experiments. In this section we briefly review the WrLINE method with its original nomenclature which is independent of the rest of the content of this thesis.

A schematic description of the WrLINE method is shown in figure 2.9, where the DNA molecular contour is defined as a set of position vectors $\vec{h}_i$ associated with the i bp step, where the main idea is to smooth the local irregularities from $\vec{h}_i$ by averaging a helical turn around bp step i. The method begins by defining a midpoint $\vec{r}_i$ that is obtained by averaging the positions of C1' atoms of the four bases (A, B, C, D) that compose the bp step i (see figure 2.9c):

$$\vec{r}_i = \frac{\vec{r}_{C1',A} + \vec{r}_{C1',B} + \vec{r}_{C1',C} + \vec{r}_{C1',D}}{4} \tag{2.91}$$

where $\vec{r}_{C1',A}$ is the coordinate of the C1' atom in base A. A-B forms a bp and its position is at the middle of the line that connects the C1' atoms $\vec{r}_{A,B} = (\vec{r}_{C1',A} + \vec{r}_{C1',B})/2$. Both $\vec{r}_{A,B}$ and $\vec{r}_{C,D}$ are used to define a local helical axis $\vec{z}_i$ (see figure 2.9a):

$$\vec{z}_i = \vec{r}_{A,B} - \vec{r}_{C,D} \tag{2.92}$$

The helical axis $\vec{z}_i$ defines a perpendicular plane Z (see figure 2.9b). The vectors $\vec{y}_{A,B}$ and $\vec{y}_{C,D}$ point to the C1' atoms of bases B and D from the midpoint $\vec{r}_i$ (respectively). These two vectors are then projected to the plane Z ($\vec{y}_{A,B;Z}$ & $\vec{y}_{C,D;Z}$) and the angle between them is precisely the twist angle θ_i (see figure 2.9b):

$$\theta_i = \arccos(\vec{y}_{A,B;Z} \cdot \vec{y}_{C,D;Z}) \tag{2.93}$$

Then, for smoothing the helical periodicity these twist angles θ_i are used to construct the Θ_m parameter defined as a sum of $(2m)$ twists around bp step i:

$$\Theta_m = \theta_i \sum_{k=1}^{m} (\theta_{i+k} + \theta_{i-k}) \tag{2.94}$$

where the integer m is chosen so that Θ_m is approximately a complete turn:

$$\Theta_m > 2\pi > \Theta_{m-1} \tag{2.95}$$

Here, Θ_m can be written in terms of Θ_{m-1}:

$$\Theta_m = \Theta_{m-1} + (\theta_{i+m} + \theta_{i-m}) \tag{2.96}$$

A weighting factor w is then introduced to force the condition $\Theta_m = 2\pi$, where w acts on the two ending bp steps:

$$2\pi = \Theta_{m-1} + w(\theta_{i+1} + \theta_{i-m}) \tag{2.97}$$

And solving this last equation we get the value of w:

$$w = \frac{2\pi - \Theta_{m-1}}{\theta_{i+m} - \theta_{i-m}} = \frac{2\pi - \Theta_{m-1}}{\Theta_m - \Theta_{m-1}} \tag{2.98}$$

Finally, each bp step i is associated with the position vector $\vec{h}_i$ which is obtained as an average of the $(2m+1)$ midpoint positions $\vec{r}_i$ around bp step i, where the weighting factor w acts on the ending bp steps ($\vec{r}_{i+m}$ & $\vec{r}_{i+m}$) (see figure 2.9d):

$$\vec{h}_i = \frac{1}{2(m+w)-1} \left(\vec{r}_i + w(\vec{r}_{i+m} + \vec{r}_{i-m}) + \sum_{k=1}^{m-1} (\vec{r}_{i+k} + \vec{r}_{i-k}) \right) \tag{2.99}$$

The complete set of points $\vec{h}_i$ represents the DNA molecular contour. This molecular contour passes through the center line of double stranded DNA, where the local irregularities caused by the helical periodicity have been filtered.

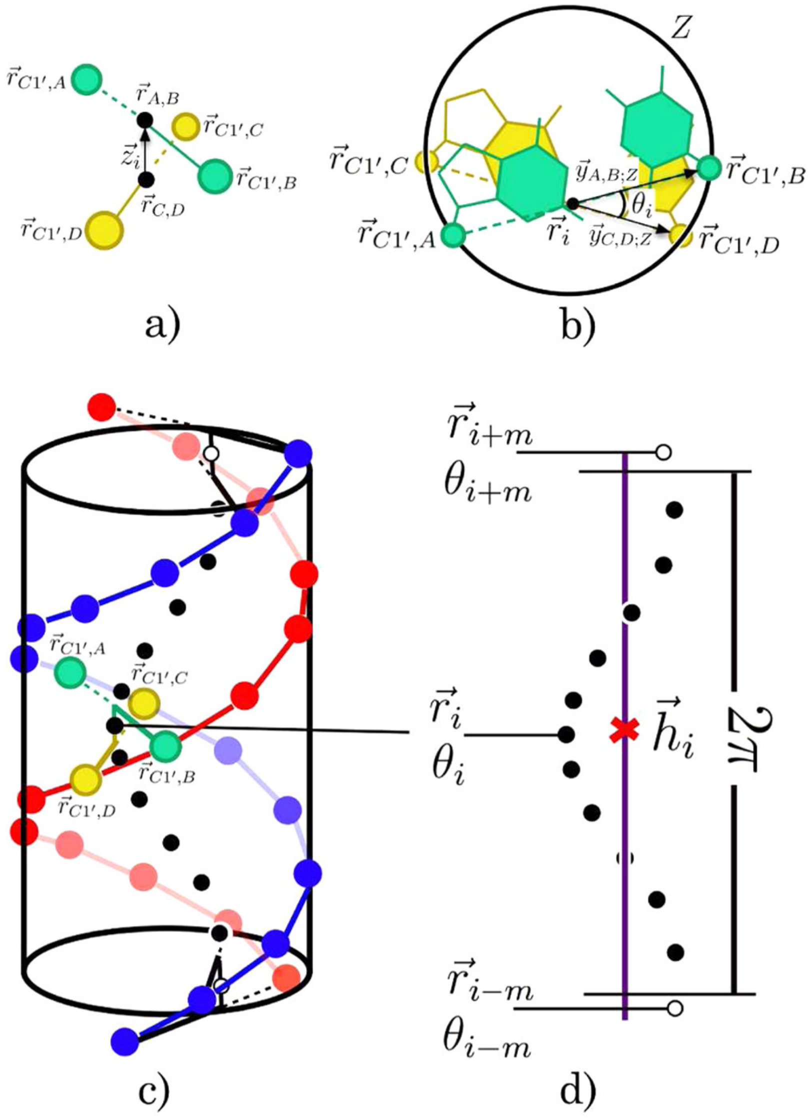


Figure 2.9: Visual representation of the WrLINE molecular contour. (a,b) Bases A, B, C, D with their respective C1' atom coordinates $\vec{r}_{C1'}$. Base-pair coordinates $\vec{r}_{A,B}$ & $\vec{r}_{C,D}$ define the local helical axis $\vec{z}_i$ which is normal to the plane Z. The twist angle θ_i is the angle between the vectors $\vec{y}_{A,B;Z}$ & $\vec{y}_{C,D;Z}$ that point to C1' atoms from the midpoint $\vec{r}_i$ on the plane Z. (c,d) Side views of a helical turn represented by a cylinder, with all the midpoints (black dots), bases on each strand (red and blue), bases in current iteration (green and yellow) and flanking midpoints $\vec{r}_{i+m}$ & $\vec{r}_{i-m}$ which are weighted to correct the excess of twist in θ_{i+m} & θ_{i-m}. The molecular contour $\vec{h}_i$ (red cross) is obtained by averaging the midpoints over a complete turn.

2.4 Plane fitting

In this section, the mathematical tools for fitting a plane to a set of coordinates is provided. To find the plane that best fits a set of $3D$ coordinates composed of N elements and expressed in matrix form X with dimensions $3 \times N$ centered at the origin, it is necessary to minimize the squared sum of orthogonal distances $\varphi(\hat{n})$ from each element i to the plane [2]:

$$\varphi(\hat{n}) = \sum_{i=1}^{N} (\vec{X}_i \cdot \hat{n})^2 = \left\|X^T \hat{n}\right\|^2 \tag{2.100}$$

where the unit vector $\hat{n}$ is normal to the plane. Matrix X can be factorised by implementing singular value decomposition (SVD) [46]:

$$X = USV^T \tag{2.101}$$

Here, U and V are unitary matrices of dimension 3×3 and $N \times N$ respectively. Given that X is a real matrix, U and V are real orthogonal matrices, which have the property of $UU^T = U^T U = I$, where I is an identity matrix with same dimension as U (the same case applies to V). S is a $3 \times N$ rectangular matrix, where all of its off-diagonal elements are zero, and some of its diagonal terms are non-zero and are called 'singular values' λ_i of matrix X. Using the SVD form of X 2.101 in equation 2.100 we get:

$$\varphi(\hat{n}) = \left\|VS^T U^T \hat{n}\right\|^2 \tag{2.102}$$

And $S^T U^T \hat{n}$ can be written in the form of:

$$\vec{P} = S^T U^T \hat{n} = \begin{vmatrix} \lambda_1 \hat{u}_1 \cdot \hat{n} \\ \lambda_2 \hat{u}_2 \cdot \hat{n} \\ \lambda_3 \hat{u}_3 \cdot \hat{n} \\ 0 \\ \vdots \\ 0 \end{vmatrix} \tag{2.103}$$

The resulting vector $\vec{P}$ has a dimension of N, with the particularity that only three elements are non-zero. The unit vectors $\hat{u}$ are the column vectors that compose matrix U. Taking advantage that V is an orthogonal matrix with the property of $\|V\vec{z}\| = \|\vec{z}\|$ $\forall\ \vec{z} \in \mathbb{R}^N$ and plugging P in equation 2.100, we get:

$$\varphi(\hat{n}) = \left\|V\vec{P}\right\|^2 = \vec{P} = \lambda_1^2(\hat{u}_1' \cdot \hat{n})^2 + \lambda_1^2(\hat{u}_2' \cdot \hat{n})^2 + \lambda_3^2(\hat{u}_3' \cdot \hat{n})^2 + 0 + \cdots + 0 \tag{2.104}$$

Given that vectors $\hat{u}_i$ form an orthonormal basis for $\mathbb{R}$, the normal vector $\hat{n}$ can be expressed as a linear combination of $\hat{u}_i$:

$$\hat{n} = a_1 \hat{u}_1 + a_2 \hat{u}_2 + a_3 \hat{u}_3 \tag{2.105}$$

And given that $\hat{n}$ is unitary, then the sum of projections is equal to one ($\sum_j a_j = 1$). Since all singular values are positive $\lambda_j > 0$, then the following equation must be true:

$$\varphi(\hat{n}) = \left\|X^T\hat{n}\right\|^2 \geq \lambda_j^2 \tag{2.106}$$

where λ_j corresponds to the minimum singular value of X. In other words, the distance $\varphi(\hat{n})$ is minimized when $\hat{n} = \hat{u}_j$, being u_j the eigenvector associated with eigenvalue λ_j.

These set of equations show that for finding the plane that best fits a set of coordinates X, it is enough to implement a SVD decomposition and find the eigenvalue matrix S and eigenvector matrix U. In practice, it turns to be difficult to find the SVD form of X since it is a non-square matrix. The covariance matrix XX^T is a real symmetric matrix and therefore diagonalizable which is also associated with S and U:

$$XX^T = (USV^T)(VS^TU^T) = USS^TU^T = US^2U^T = US'U^T \tag{2.107}$$

Note that the identity $V^TV = I$ was used and that $S' = SS^T = S^2$ is a 3×3 square matrix where its diagonal is composed by the squared eigenvalues λ_i^2. From this last expression, matrices U and S' can be approximated by implementing the Jacobi algorithm, which is used for calculating the eigenvalue W and eigenvector G matrices of a real symmetric matrix M with the form $M = GWG^T$ (see section 2.5). Then, equation 2.107 can be solved by implementing the Jacobi method, where U corresponds to a 3×3 orthonormal matrix calculated from a series of rotations U (see equation 2.112) and S' would be approximately diagonal. Once U and S' are calculated, the plane that best fits X with normal vector $\hat{n}$ would correspond to the column vector of U with the corresponding minimum eigenvalue $S_{i,i}$.

2.5 Jacobi algorithm

The Jacobi algorithm is used for the diagonalization of matrix M [124]. The Jacobi algorithm consists of an iterative process that aims to factorise matrix M into the form:

$$M = GWG^T \tag{2.108}$$

where W is a real symmetric matrix and G corresponds to a Givens rotation matrix with the form:

$$G(i,j,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & cos\theta & \cdots & -sin\theta & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & sin\theta & \cdots & cos\theta & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \tag{2.109}$$

Here both indices i & j are fixed and $i > j$. Elements $G_{i,i} = G_{j,j} = \cos(\theta)$ and $G_{j,i} = -G_{i,j} = -sin\theta$. The rest of all other diagonal elements are equal to one and all off-diagonal elements are equal to zero. Then, the Jacobi algorithm consists of performing several rotations with the aim of eliminating the off-diagonal elements of

W until it is approximately diagonal. The algorithm is initialized with $W = M$, and before constructing matrix G, the angle θ is defined as:

$$\theta = \frac{1}{2}\arctan\left(\frac{2W_{i,j}}{W_{j,j} - W_{i,i}}\right) \tag{2.110}$$

where $W_{i,j}$ corresponds to the off-diagonal element with the largest absolute value. In the case that $W_{j,j} = W_{i,i}$, then:

$$\theta = \frac{\pi}{4} \tag{2.111}$$

The process finishes when W becomes approximately diagonal and the diagonal elements of W are precisely the eigenvalues of matrix M. The eigenvectors of M correspond to the column vectors of matrix U, which is defined as the m rotations performed in order to diagonalize W:

$$U = \prod_{s=1}^{m} G_s(i_s, j_s, \theta_s) \tag{2.112}$$

Notice that each G_s is a Givens rotation matrix.

2.6 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a popular technique for capturing the essential modes of molecular systems typically obtained by MD simulations. The technique focuses on obtaining and selecting the minimum number of eigenvectors that explain the maximum amount of system variance. This allows to greatly compress the trajectory size, where a few essential modes recreate most of the original trajectory without affecting the overall behaviour.

Given a MD simulation of N atoms and K frames, the $3N \times 3N$ covariance matrix V is calculated. Then, V is diagonalized in order to obtain the corresponding eigenvalues and eigenvectors. The eigenvectors are the essential modes, and their associated eigenvalues are the quantities that explain the amount of system variance. The total amount of variance is obtained by summing all the eigenvalues. Once the eigenvectors and eigenvalues have been obtained, the coordinates $\vec{X}_k$ at frame k can be written in terms of their principal components as a linear combination:

$$\vec{X}_k = p_k^1\vec{e}_1 + p_k^2\vec{e}_2 + p_k^3\vec{e}_3 + \cdots + p_k^{3n}\vec{e}_{3N} + \vec{A} \tag{2.113}$$

where $\vec{e}_j$ is the eigenvector of mode j, p_k^j is the projection of mode j on frame k and $\vec{A}$ is the average structure. Note that each frame of trajectory X is described as a vector of dimension $3N$ and therefore, there exist $3N$ principal modes.

We use PCAsuite [138] to perform PCA to a given trajectory in which it then sorts the essential modes according to the amount of variance they contribute to the system. Usually, just a few essential modes M are responsible for most of the trajectory's dynamics. Then the trajectory can be rebuilt using any mode j ($1 \leq j \leq M$) as:

$$\vec{X}_k^j = p_k^j\vec{e}_j + \vec{A} \tag{2.114}$$

Similarly, we can rebuild the coordinates $\vec{X}_k$ using a combination of M modes $(i_1, i_2, i_3, ..., i_M)$:

$$\vec{X}_k = p_k^{i_1}\vec{e}_{i_1} + p_k^{i_2}\vec{e}_{i_2} + p_k^{i_3}\vec{e}_{i_3} + \cdots + p_k^{i_M}\vec{e}_{i_M} + \vec{A} \tag{2.115}$$

Lastly, the percentage proportion of system variance ν_i explained by eigenvector i can be computed through the related eigenvalue:

$$\nu_i = \frac{\lambda_i}{\sum_{j=1}^{3N}\lambda_j} \tag{2.116}$$

where $3N$ is the number of eigenvectors. Notice that $\sum_{j=1}^{3n}\lambda_j$ corresponds to the total system variance.

2.6.1 Comparison of principal components

One of the main applications of PCA is to compress a trajectory to a certain fraction (percentage) of its original size, by calculating the essential modes that capture most of the system variance. However, the technique has also been used in the past to study the nature of the principal modes [39,116]. Following the procedures of Noy et al. [116], the principal modes of two structures can be compared between them through their dot product (γ):

$$\gamma_{U,V} = \vec{e}_i^{\,U} \cdot \vec{e}_j^{\,V} \tag{2.117}$$

where $\vec{e}_i^{\,U}$ corresponds to the eigenvector i of structure U .

One of the main flaws of the previous equation is that both structures require to have the same number of atoms. Notice that even if A and B have the same number of bp, their number of atoms could still be different due to their sequences. To allow the comparison between two structures U and V with same number of bp and not necessarily same sequence, we propose the following pre-process:

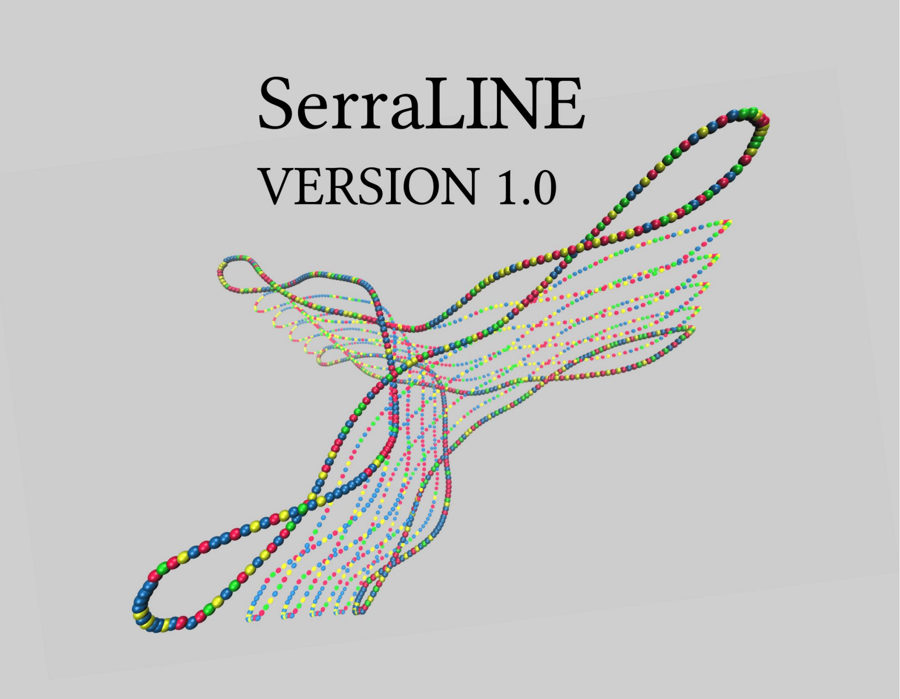
- For both U and V, atoms that do not form part of the DNA backbone are removed. This filtering is performed to both average structures ($\vec{A}^U$ & $\vec{A}^V$) and eigenvectors. This pre-process ensures that all mathematical objects have the same dimension.
- A 3x3 rotation matrix (R_U^V) is calculated from the average structure. This rotation matrix minimises the root-mean square (RMSD) in 3D space when taking the average structure of U as reference and rotating around the average structure of V.
- Each eigenvector $\vec{e}_j^{\,V}$ is then rotated in the 3D space by applying the rotation matrix R_i^j to it.
- Because elements were removed from each eigenvector, they are no longer normalised. Therefore, we re-normalise each of the eigenvectors once more ($||\vec{e}|| = 1$).

After this pre-process, the essential modes of two structures with the same number of bp can be compared using equation 2.117.

It is worth pointing out that if we were to rebuild the trajectory using this method, we would lose valuable structural information as only the backbone atoms would remain, and it would not be possible to calculate parameters such as BPP, BSP and LDEM related variables. However, the primary objective of this pre-process is not to reconstruct trajectories or extract the aforementioned parameters. Instead, it serves as an approximation that facilitates the comparison of two principal modes calculated from DNA structures with the same number of base-pairs but different sequences.

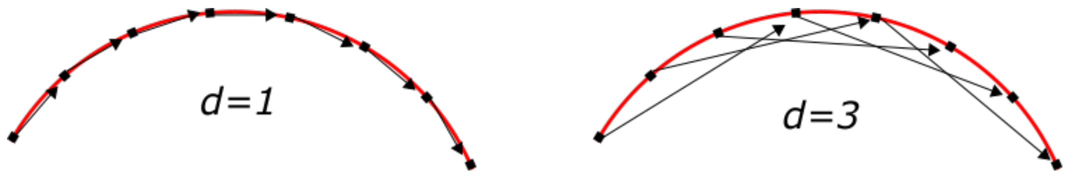
Chapter 3

SerraLINE



Synopsis

SerraLINE is a program we developed for the analysis of MD simulations. SerraLINE uses the molecular contour calculated by the WrLINE program to calculate bending angle distributions at different lengths. Additionally, the program can project the molecule to a plane and calculate global quantities like the aspect ratio or deviation from planarity that can be used to directly compare with single molecule experiments such as AFM, where the molecules are visualised in a 2D plane. In this chapter, we present the proceedings and general workflow followed by SerraLINE as well as results that have been obtained with the program and have been proven to be extremely useful in the analysis of supercoiled DNA. By comparing the results of distinct levels of supercoiling in DNA minicircles with corresponding AFM experiments, we successfully studied the response of DNA to torsional stress. The remarkable agreement between the computational and experimental approaches underscores the significance of developing computational tools for the comprehensive analysis of DNA simulations and experiments. Furthermore, our multi-approach analysis using SerraLINE has provided novel insights into the structural response of DNA when subjected to DNA supercoiling, which has great biological relevance in essential processes such as DNA packaging.



(a) Tangent vectors at resolution $d = 1$. (b) Tangent vectors at resolution $d = 3$.

Figure 3.1: Representation of the direction of tangent vectors (black arrows) assigned to each bp (black squares) constructed with 2 different resolutions d, where the red curve represents the molecular contour. The magnitude of tangent vectors are exaggerated in (b) for visualisation purposes, but they are normalised in SerraLINE (see equation 3.1).

3.1 Tangent vectors

Given a snapshot of a molecular contour X represented as a $3 \times N$ matrix and composed of N bp where each element i represents a bp with coordinates $\vec{x_i}$, SerraLINE assigns tangent vectors $\hat{t_i}$ to each bp to describe their direction (see figure 3.1). These tangent vectors are parallel to the line that connects bp i with the bp at index $i + d$:

$$\hat{t_i} = \frac{(\vec{x}_{i+d} - \vec{x}_i)}{|\vec{x}_{i+d} - \vec{x}_i|} \tag{3.1}$$

where, d is the resolution of the tangent vectors (in bp steps), which can be adjusted in order to mimic experimental techniques such as AFM. Notice that when $d = 1$, the vectors are approximately tangent to the molecular contour as shown in figure 3.1a.

3.2 Bending angles in 3D space

Once tangent vectors are assigned to each bp, SerraLINE proceeds to calculate the bending angle of every possible oligomer whose length is composed of $l + 1$ bp, where l ranges from 1 to $N - 1$ bp. SerraLINE calculates the bending angle between bp i and j in 3D space as:

$$\theta_{i,j} = \arccos\left(\hat{t_i} \cdot \hat{t_j}\right) \tag{3.2}$$

Figure 3.2 shows a 2D representation of two bending angles being calculated at two different length-scales l. SerraLINE calculates bending angles at each time frame and, then, it computes averages and standard deviations for each oligomer formed. Finally, these quantities are written in an output file.

3.3 Plane projection

Similar to single molecule techniques where molecules are projected and visualised in a 2D space, SerraLINE can project each trajectory frame X to the plane that best fits the whole structure, or to a substructure Y indicated by the user and composed of N_Y bp. The coordinates of the substructure Y are made out of X, and its number of points needs to be higher or equal than 3 bp ($N \geq N_Y \geq 3$) in order to perform the

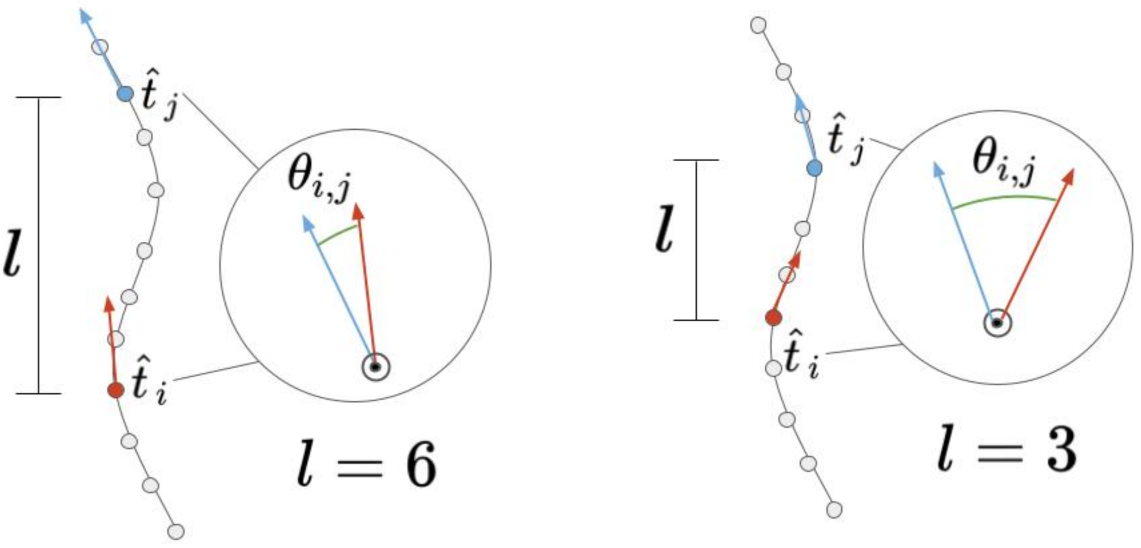


Figure 3.2: Representation of bending angles calculated for a structure at two different length scales l, one at 6 bp and the other one at 3 bp. In this example, $d = 1$.

fitting process.

The fitting procedure consists of two main steps: a global plane is first calculated, and then the structure is projected onto it. To calculate the global plane G, SerraLINE first translates each coordinate $\vec{x}_i$ of the input trajectory, so the substructure Y is centred at the origin:

$$\vec{x}_i{}' = \vec{x}_i - \vec{c} \tag{3.3}$$

$$\vec{y}_i{}' = \vec{y}_i - \vec{c} \tag{3.4}$$

where X' is the translated structure with coordinates $\vec{x}_i{}'$, Y' the translated substructure with coordinates $\vec{y}_i{}'$ and $\vec{c} = \sum_i \vec{y}_i / N_Y$.

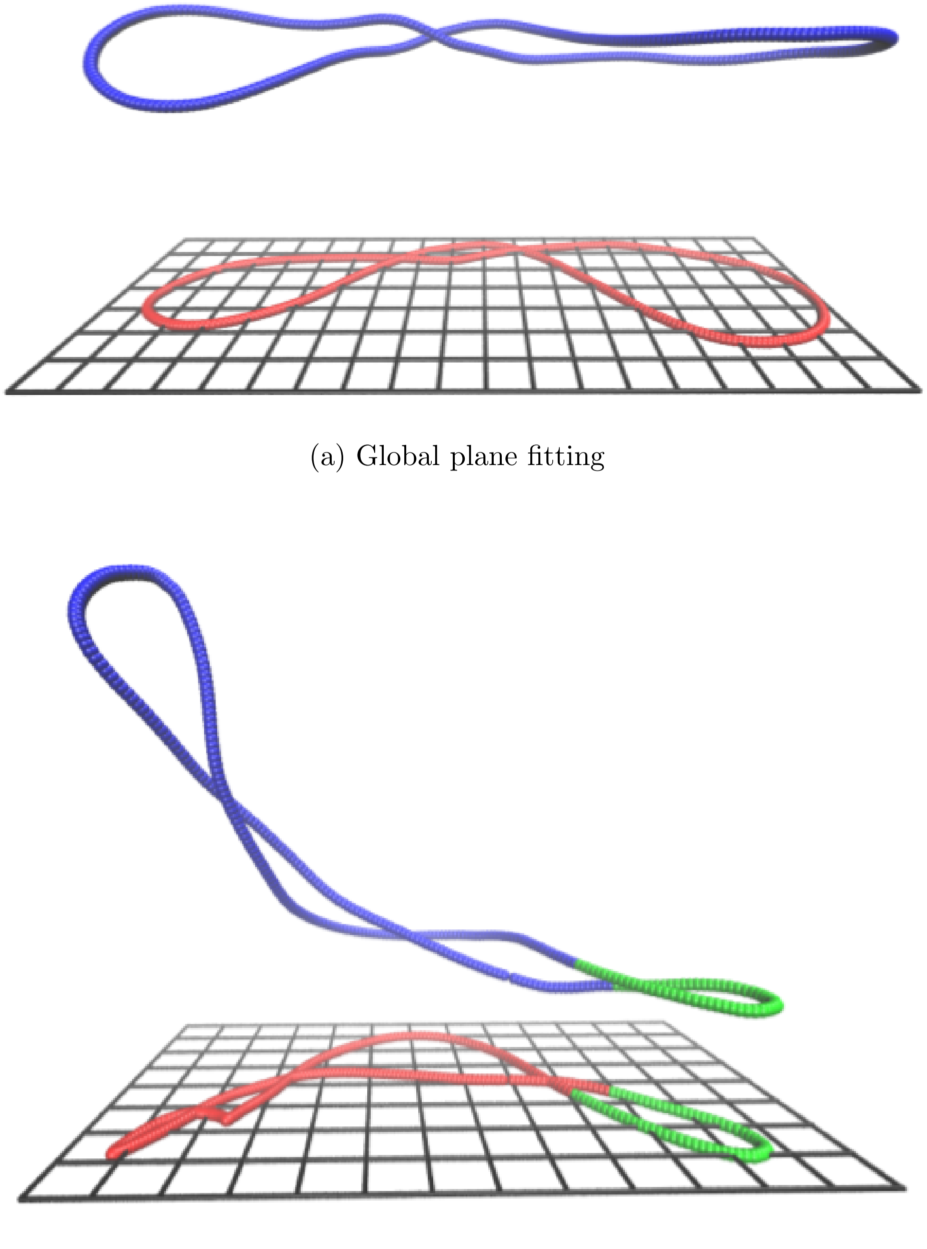
Once the whole structure X' has been translated, by following the methods in sections 2.4-2.5, SerraLINE fits a plane to the selected substructure Y', which can be the whole structure $Y = X$ or a part of $Y \subseteq X$. As a result, the normal vector $\hat{n}_G$ that defines the plane G that best fits coordinates Y' is calculated.

The final step of the fitting process consists in projecting the whole structure X' to the just calculated plane G:

$$\vec{z}_i = \vec{x}_i{}' - \frac{\vec{x}_i{}' \cdot \hat{n}_G}{\|\hat{n}_G\|^2} \hat{n}_G \tag{3.5}$$

where the structure Z is the projected structure with coordinates $\vec{z}_i$. Notice that in equation 3.5, a component normal to the plane is being subtracted.

Figure 3.3 shows two examples of plane fittings with $Y = X$ (top panel) and with $Y \subset X$ (bottom panel). For both cases, the same fitting process is applied and in either case $Y \subseteq X$.



(a) Global plane fitting

(b) Specific region plane fitting

Figure 3.3: The top figure shows a global fitting, where a plane is fitted to the whole 3D structure (in blue) and projected to a plane. The bottom figure shows the second type of fitting, where the plane is fitted to a particular region (green) from the 3D structure.

3.4 Compaction measurements

Single molecule experiments such as AFM, are abl6e to visualise molecules in 2D and often describe the observed shapes in terms of compaction parameters such as the aspect ratio or the radius of gyration. The radius of gyration is a great tool for determining the spread or mass concentration in a polymer, and can be calculated with programs such as cpptraj which is a software for analysing MD simulations part of AmberTools [16]. However, even though the radius of gyration can provide information regarding the compaction of structures, we consider it is not the adequate technique for our case of study as we are interested in investigating how the shape of the DNA changes at distinct levels of torsional stress. For this reason, we choose to use the aspect ratio as the suitable technique for this thesis, as this parameter can be used to describe the proportion of sizes of a structure. Nonetheless, we recognize that the radius of gyration is a useful technique that will be worth adding as a new feature in future versions of SerraLINE, as this will allow our software to provide a more complete descriptions for the analysis of MD simulations.

Continuing with SerraLINE procedures, in the case that the structure X has been projected forming Z, SerraLINE calculates compaction measurements such as the aspect ratio, width, height and deviation from planarity.

Being calculated the projected structure Z, SerraLINE first finds the height axis as:

$$\hat{h} = \frac{\vec{z_j} - \vec{z_i}}{|\vec{z_j} - \vec{z_i}|} \tag{3.6}$$

where j and i, are the two elements whose separation is the largest in the given time frame. In other words, the height axis $\hat{h}$ is parallel to the largest distance in X. In combination with the normal vector $\hat{n}_G$, the height axis is used to define the width axis $\hat{w}$:

$$\hat{w} = \hat{n}_G \times \hat{h} \tag{3.7}$$

The axis $\hat{w}$ is perpendicular to $\hat{h}$ and lies in the plane G as well. Once these axes have been calculated, height H and width parameters W, which describe the sizes of the box that would surround structure Z, are calculated as:

$$H = \frac{(\vec{z_j} - \vec{z_i}) \cdot \hat{h}}{\left\|\hat{h}\right\|^2} \tag{3.8}$$

$$W = \frac{(\vec{z_p} - \vec{z_q}) \cdot \hat{w}}{\left\|\hat{w}\right\|^2} \tag{3.9}$$

Here, i and j form the largest distance when projected in $\hat{h}$ and p and q form the largest distance projected in $\hat{w}$ (see figure 3.4). Height and width have the same distance units as the trajectory coordinates, which is usually angstroms Å. These distances, are also the sizes of a rectangle with the minimum area that can enclose the

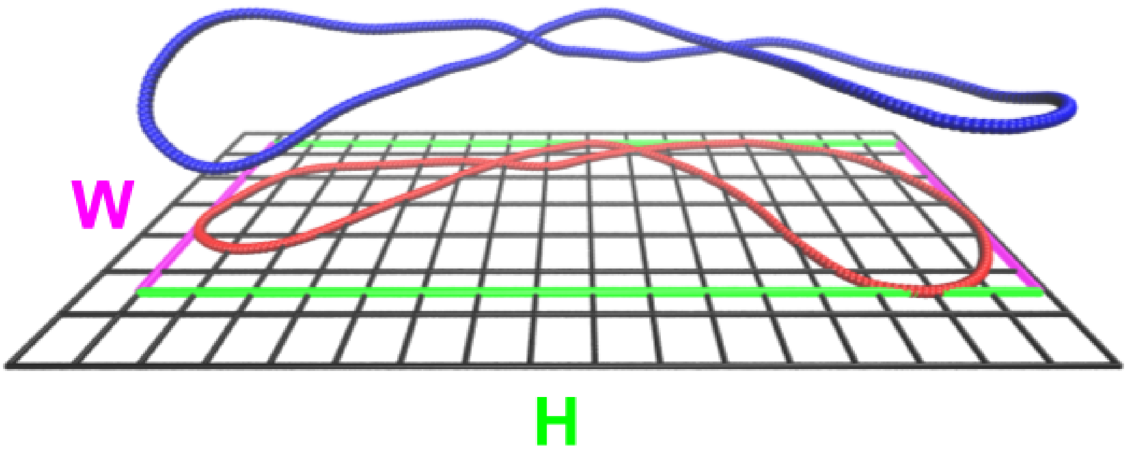


Figure 3.4: Representation of width W (pink) and height H (green) calculated from a structure (blue) projected (red) into the best fitted plane (black grid).

projected structure Z.

Once calculated H and W, the aspect ratio A is simple calculated as:

$$A = W/H \tag{3.10}$$

This global geometric parameter is dimensionless and is particular useful since it explains how rectangular a structure is, and can be compared with experimental results.

Besides, SerraLINE can calculate the deviation from planarity D which we define as the distance to the plane:

$$D = \frac{1}{N}\sum_{i=1}^{N} \frac{(\vec{x}_i{}') \cdot \hat{n}_G}{\|\hat{n}_G\|^2} \tag{3.11}$$

Notice that the translated structure X' is used rather than Z, and each distance to the plane is precisely the normal component of each bp coordinate $\vec{x}_i{}'$. Another very useful parameter to measure the planarity is the distance between the plane and the farthest point in X':

$$D_{max} = \max_{i}\left(\frac{(\vec{x}_i{}') \cdot \hat{n}_G}{\|\hat{n}\|^2}\right) \tag{3.12}$$

where $\vec{x}_i{}'$ is the farthest element along the perpendicular component $\hat{n}_G$.

Lastly, SerraLINE can calculate the relatives of distances D and D_{max} with respect the height parameter H:

$$\dot{D} = 100\% \frac{D}{H} \tag{3.13}$$

$$\dot{D}_{max} = 100\% \frac{D_{max}}{H} \tag{3.14}$$

This measure is very important to support AFM experiments because it indicates if the molecule will strongly be deformed from the fact of being in a 2D surface.

In general, these set of parameters allow us to globally quantify how compact and planar a structure is, and are suitable for comparison with experiments.

3.5 Bending angles in 2D space

When the structure is projected into a plane, SerraLINE calculates the tangent vectors $\hat{t}_i$ using the same equation 3.1 but using the projected structure Z rather than X, and in contrast of the 3D structure the bending angle is calculated differently. Basically, having the structure Z lying on a plane allow us to associate a directionality with the bending angle. To define the direction of the bending angle between the consecutive elements i and $j = i + 1$, we first introduce the following vector $\vec{b}$:

$$\vec{b} = \hat{t}_i \times \hat{t}_j \tag{3.15}$$

Then, with the help of normal vector $\hat{n}_G$, we define the scalar d:

$$d = \vec{b} \cdot \hat{n}_G \tag{3.16}$$

which allows us to add the directionality of bending angle θ_i:

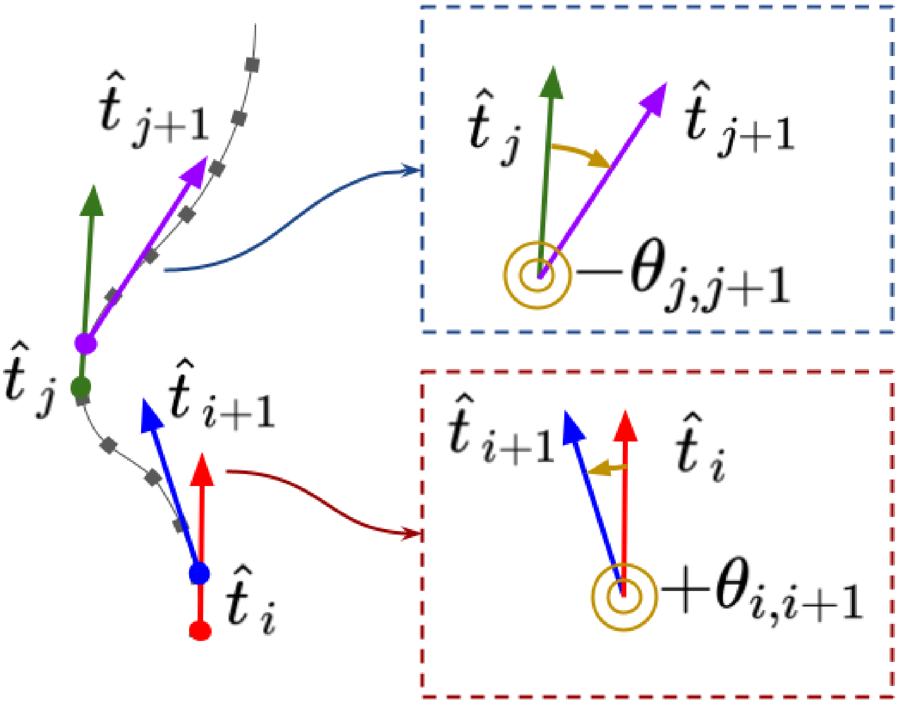
$$\theta_{i,j} = \arccos(\hat{t}_i \cdot \hat{t}_j) \quad if \quad d \geq 0 \tag{3.17}$$

$$\theta_{i,j} = -\arccos(\hat{t}_i \cdot \hat{t}_j) \quad if \quad d < 0 \tag{3.18}$$

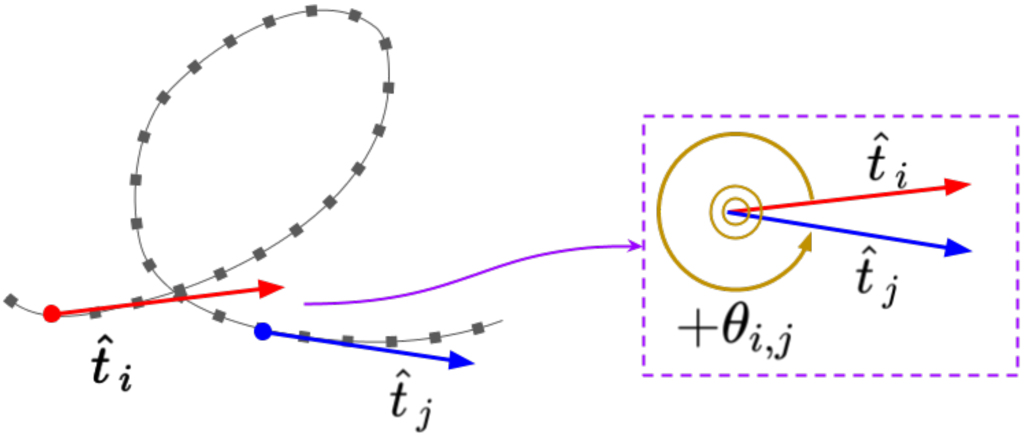
In general, d indicates if bp $i+1$ is bent to the left or right of i, which is indicated by the sign of θ_i. Once all consecutive bending angles are calculated ($l = 1$), SerraLINE proceeds to calculate the bending angle for higher lengths $l > 1$:

$$\theta_{i,j} = \sum_{k=i}^{j} \theta_{k,k+1} \tag{3.19}$$

Notice that $j = i + l$ and that bending angles $\theta_{i,j}$ have direction and can be greater than 180°. Figure 3.5 shows an example of 2D bending calculations, where bending angles are first given direction/sign at the length $l = 1$ bp (figure 3.5a), and then at longer lengths $l > 1$ bp (figure 3.5b). This is particularly useful for keeping track of how many turns the contour travels, which can later be used to characterise important events such as the effect of supercoiling, protein binding, etc...



(a) Example of a positive bending angle ($\theta_{i,i+1}$) and negative angle ($\theta_{j,j+1}$) taken from consecutive tangent lengths $\hat{t}_i/\hat{t}_{i+1}$ and $\hat{t}_j/\hat{t}_{j+1}$, respectively.



(b) Example of bending angles at longer lengths $l > 1$ bp, where the angle between bp i and j, is calculated as a sum of consecutive angles (see equation 3.19).

Figure 3.5: Visual representation of positive and negative bending angles in a 2D plane at the length $l = 1$ bp (3.5a) and at longer lengths $l > 1$ bp (3.5b), where the contour almost completes a turn. Notice that for both examples, a bending to the left corresponds to a positive angle, whereas a bending to the right to a negative angle.

3.6 General workflow

Figure 3.6a) shows the general workflow, where the required input trajectory must have an Amber crd format [16] or the format outputted by the WrLINE program [150]. SerraLINE can work without a topology file, but the sequence specificity is lost if this file is not given, hence it is treated as an optional input. SerraLINE is separated into two programs: a main program called **SerraLINE** and a secondary program called **Extract**. **SerraLINE** processes the inputs and calculates the 3D bending parameters. In case that the projection method is selected, **SerraLINE** calculates the 2D bending angles as well as the compaction measurements and the projected trajectory. The main program **SerraLINE** creates the output file *SerraLINE.out*, which contains information regarding the structure, the selected method, compaction measurements if it is the case, and bending angles sorted by length l. The secondary program **Extract** is used to process the bending angles, filtering them at a specific length l given by the user. A file called *subfragment_$l.out* is then created whose name depends on the input length l and which is ready to be plotted.

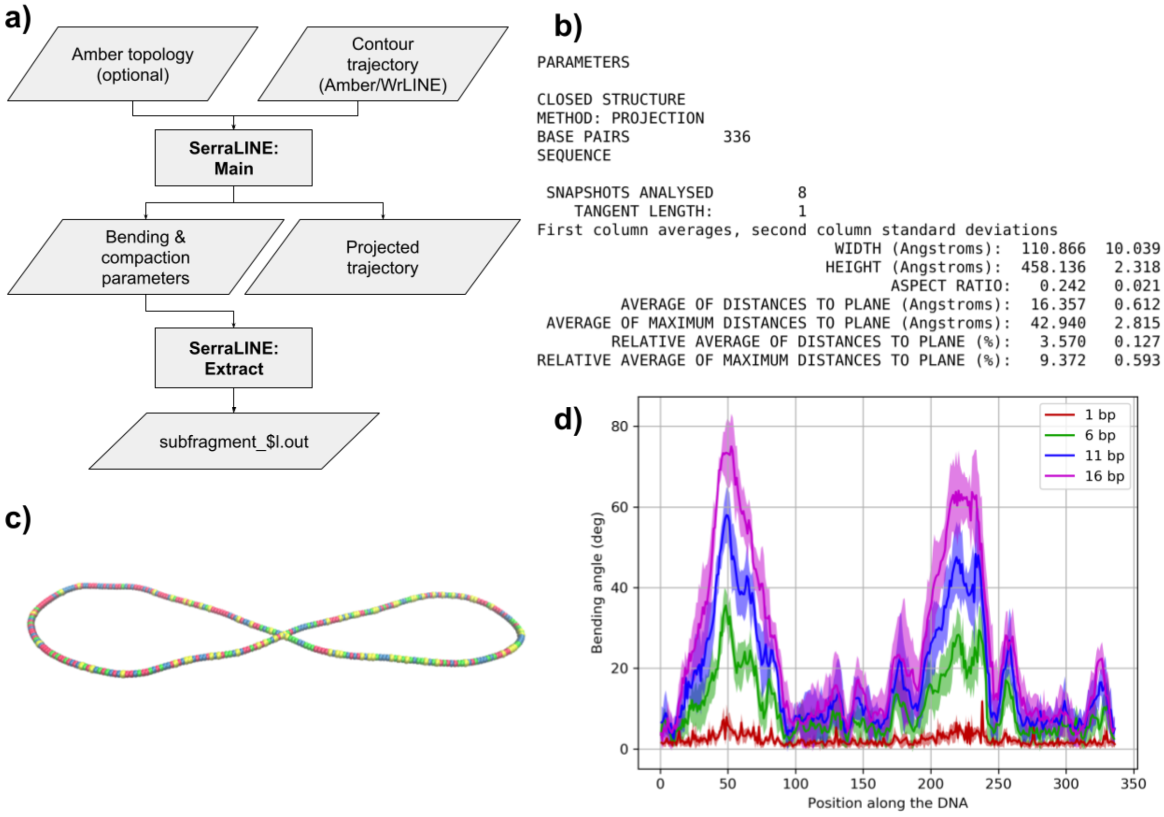
The SerraLINE program comes with an example trajectory contour created by WrLINE that corresponds to a 336 bp supercoiled DNA minicircle with 8 frames. Panel **b)** of figure 3.6 shows the top lines of the output file *SerraLINE.out*, where information regarding the system and user options (in this case, correspond to the projection method and tangent lengths constructed from consecutive bp with resolution $d = 1$) is printed together with the averages and standard deviations of compaction and bending parameters. The width, height and aspect ratio shown in panel **b)** agree with the shape of the projected structure (panel **c**), where height is about 4 times larger than width. The relative average of distances to the plane tells how planar is the 3D structure compared to height, which for this trajectory has a value of 3.57% indicating that the 3D structure is mostly planar throughout the simulation. Finally, panel **d)** of figure 3.6 shows the plot of the bending angles at the lengths $l = 1, 6, 11, 16$, where it can be seen that at the positions of 50 and 230 along the DNA, the bending angles reach values higher than 60°. These two highly bent regions correspond to the U-turns shown in the projected structure of panel **c)**.

In general, the workflow of SerraLINE is really simple and easy to use, and provides the necessary tools for data analysis and visualisation which greatly facilitates the interpretation of results. SerraLINE is written in Fortran and is a free access program, which is available under version 3.0 of the GNU Lesser General Public Licence[*] at agnesnoy/SerraLINE GitHub repository[†]. We chose this particular licence as it has little restrictions and allows users as well as developers, to add and integrate software components to their own libraries, and in the case of modifying SerraLINE, they are required to publish their own modifications under the same licence.

For detailed information on accessing and installing SerraLINE, please refer to section E.1. Furthermore, specific instructions detailing the usage of SerraLINE are provided within its accompanying manual (see section F).

[*] `https://www.gnu.org/licenses/gpl-3.0.en.html`

[†] `https://github.com/agnesnoy/SerraLINE`



```
PARAMETERS

CLOSED STRUCTURE
METHOD: PROJECTION
BASE PAIRS          336
SEQUENCE

 SNAPSHOTS ANALYSED         8
    TANGENT LENGTH:         1
First column averages, second column standard deviations
                                WIDTH (Angstroms):  110.866  10.039
                               HEIGHT (Angstroms):  458.136   2.318
                                      ASPECT RATIO:   0.242   0.021
         AVERAGE OF DISTANCES TO PLANE (Angstroms):  16.357   0.612
 AVERAGE OF MAXIMUM DISTANCES TO PLANE (Angstroms):  42.940   2.815
        RELATIVE AVERAGE OF DISTANCES TO PLANE (%):   3.570   0.127
RELATIVE AVERAGE OF MAXIMUM DISTANCES TO PLANE (%):   9.372   0.593
```

Figure 3.6: (a) General SerraLINE workflow. (b) Information from the output file *SerraLINE.out* created by the **SerraLINE** main program. (c) Projected molecular contour. (d) Bending profiles at the lengths of 1, 6, 11 and 16 bp.

3.7 Results and discussion

We now proceed to demonstrate the valuable that SerraLINE offers by analysing a set of MD simulations of DNA minicircles with different levels of supercoiling as well as different lengths. This set consists of two DNA minicircles composed of 260 and 339 bp, and the level of supercoiling is indicated by the linking difference ΔLk being $0, -1, -2$ and $0, -1, -2, -2, -3, -3, -6$, respectively, where repeated numbers indicate replica simulations. The results are compared with high-resolution AFM images of DNA minicircles of 251 and 339 bp. Figure B.1 shows a visual comparison of the observed structures from AFM and MD simulations, which demonstrates good agreement between the two approaches and the variety of conformations adopted by supercoiled DNA. More information regarding the sequences and structures can be found in the sections A.1 & B.

In this results section we aim to study the effect of negative supercoiling on the structure/flexibility of DNA minicircles. We use SerraLINE to characterise bending angles that correspond to B-form DNA as well as bends associated with DNA defects. Compaction and planarity parameters were also obtained to study the DNA minicircles overall response to negative torsional stress. This analysis supported the comparison between AFM and MD and is published in [125].

3.7.1 Bending angles in negative supercoiled DNA minicircles

For measuring bending angles, we selected the maximum resolution ($d = 1$) for defining the tangent vectors (see figure 3.1a), and structures were projected onto the best fit plane to mimic the AFM 2D perspective to allow comparison between both approaches.

From analysing the curvature of both high-resolution AFM images and atomistic MD simulations of DNA, we were able to determine the maximum bent that B-DNA can sustain, being larger bends associated with disrupted DNA such as kinks, melting bubbles and other types of DNA defects. Regarding AFM, figure B.2 shows the process for measuring bending angles, where red triangles indicate detected defects in the double helix. Regarding simulations, bending profiles allow us to observe strong bent regions and by visual inspection, we identify whether they are caused by the rupture of the B-DNA structure, which are characterised by the lost of base pairing or by a disruption to the base stacking (see figure 3.7). By analysing bending profiles, we classify bend angles as defective DNA (red triangles) or B-DNA (blue circles), and by comparing both AFM images and simulations, we see a clear cutoff between the two types, deducing that B-DNA can sustain bending deformations up to around 70 degrees (74 in the case of simulations and 76 in the case of AFM, see figure 3.8) in regions of approximately one and a half DNA turns of length (16 bp $\approx$ 5.3 nm). Assuming that this is the maximum bend B-DNA can sustain, to loop a DNA (bend 180°) without causing defects, the minimum length required is $\sim$ 39 bp or $\sim$ 4 turns, which highly agrees with results from coarse-grained simulations [100].

Figure 3.7 shows the bending profiles at the critical length of $l = 16$ bp as well as the average structures of three DNA minicircles of 339 bp with different levels of supercoiling, where the zoomed regions provide visual representation of DNA defects. The relaxed minicircle (see figure 3.7a) presents two high bends that are less than the critical angle ($< 75°$) by preserving B-form [125]. The bending profile of the minicircle with 3 turns removed (figure 3.7b) presents three main peaks which are associated with defects. The first defect corresponds to a type II kink [75] (zoomed in red around position 50), which is characterised by the breaking of the hydrogen bonds of two consecutive bp and which bases are stacked on the 5' neighbour bases; the second defect located around position 215 corresponds to a denaturation bubble with a size of two bp (zoomed in red), in which the hydrogen bonds of the two bp are broken and each strand becomes single stranded; and the third defect corresponds to a type I kink [75] located around position 240 (zoomed in blue), where due to the strong bent the stacking interaction is loss but the base-pairing remains (unbroken hydrogen bonds). Similarly, the bending profile of the minicircle with $\Delta Lk = -6$ presents multiple peaks (figure 3.7c), where the high bends correspond to three defects, which are associated with denaturation bubbles in which the base-pairs are flipped out of the duplex. All the defects contain averaged bend angles that exceed the 75° cutoff demonstrating that they behave as flexible hinges where the DNA can relax the accumulated bend and torsional stress. These hinges/defects allow single helical turns to sustain bends of 180° agreeing with previous coarse-grained results [100].

Figure 3.8 shows bending angles at $l = 16$ bp for the whole set of 339 bp minicircle simulations, where 10 high bent regions correspond to defective DNA while 23 to B-DNA (31 and 5, respectively, in case of AFM). The highest averaged bend angle without associated defects is 74° (purple dotted line) and is observed for the -3 topoisomer, which is similar to the highest bend angle associated with B-DNA observed by AFM being 76° (black dotted line). The averaged bend angle classified as B-DNA is $57 \pm 9°$, while the averaged bend angle associated with DNA defects is $120 \pm 32°$, where in case of AFM measurements correspond to $69 \pm 5°$ and $106 \pm 15°$, respectively. A tendency observed is that increasing the level of negative supercoiling increases the magnitude of bending angles, which then, combined with the untwisting of the double helix can provoke DNA defects. These defects have been previously observed in small supercoiled minicircles (60-100 bp) by MD simulations [75, 105, 151] and by experimental approaches such cryo-electron microscopy [27, 83] and biochemical analysis [35], where enzymes were used to selectively cut disrupted DNA regions. More recently, supercoiled minicircles of 336 bp were analysed by cryo-electron tomography [62], where enzymes also probed the structures and detected defects in minicircles with $\Delta Lk = -2, -3, -6$, which agrees with our observations (see figures 3.7 & 3.8). Furthermore, we observe that the onset of defects is in $\Delta Lk = -1, -2$ ($\sigma \approx -0.03, -0.06$), which is in the range of the superhelical density in vivo ($\sigma \approx -0.06$) [57]. This indicates that in vivo DNA is found in a negative superhelical state, which makes it relatively common to find defects.

It is important to emphasise that in the case of MD simulations, the shape of structures as well as bend angles of highly curved regions were quite stable, and defects remained in the same locations. Nevertheless, fluctuations in bend angles were observed throughout the simulation, and it was observed that some bend angles classified as B-DNA may overlap with the angles classified as DNA defects in the range of $[69°, 86°]$ (indicated by the shaded area in figure 3.8). An interesting observation was made regarding the topoisomerase -3, where it was noticed that the bend between positions 200 and 250 in replica 2 of topoisomerase -3 was classified as B-DNA, which coincided with the position of two defects in replica 1 of the same topoisomerase (as

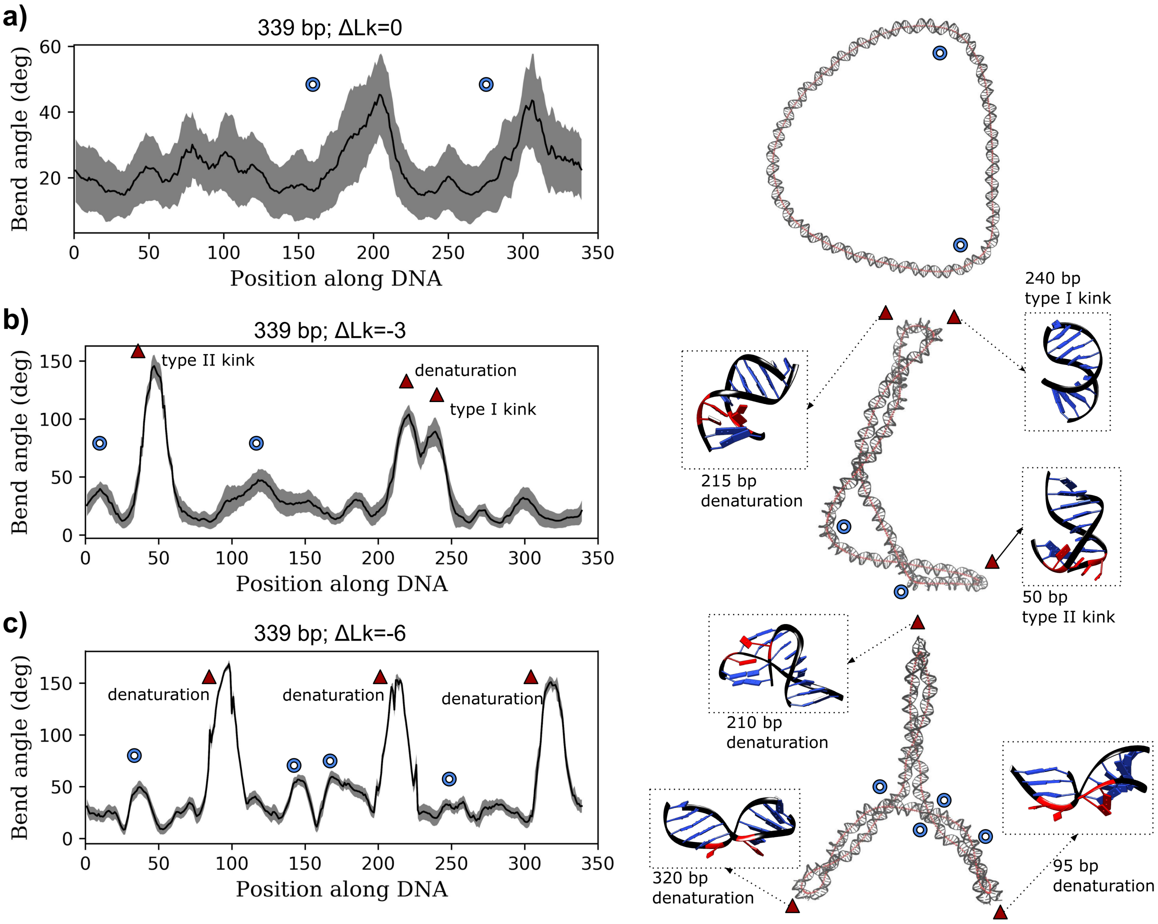


Figure 3.7: Bending angle profiles at the length of 16 bp along with the structures in atomistic (grey scales) and molecular contour (red) representations for the 339 minicircles with a liking difference (ΔLk) of 0/relaxed (a), -3 (b) and -6 (c). Blue circles indicate bent regions classified as B-DNA and red triangles as bent regions classified as defects. Black colour in zoomed structures represent DNA backbones, base-pairs with preserved hydrogen bonds are coloured as blue while base-pairs with broken hydrogen bonds are coloured as red. Shaded areas represent standard deviations. Defects are accompanied by text labels that indicate the type of defect (type I kink, type II kink or denaturation) as well as the bp position in the case of zoomed structures.

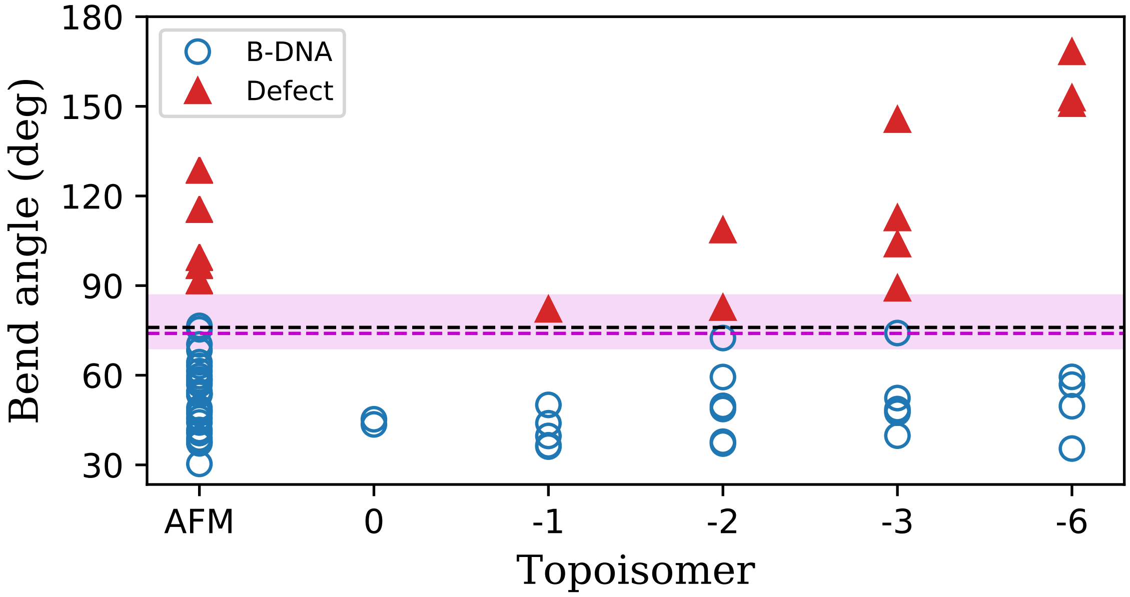


Figure 3.8: Analysis of bent DNA regions of one and a half DNA turns in size (16 bp), taken by AFM (first column) and MD approaches. Blue circles correspond to bent regions classified as B-DNA, while red triangles to regions associated with defects. Dotted lines indicate the maximum bend angles associated with B-DNA in AFM (black) and MD simulations (purple) being approximately 76° and 75° respectively. The shaded area $[69°, 86°]$ in purple represents the range in which bend angles associated with B-DNA and defects overlapped in MD simulations.

depicted in Figures B.4e-f). The bending angles in these two replicas are in fact within the overlapping range, which leads us to deduce that the transition from B-DNA to defects may occur within this angle range. However, it must be acknowledged that to directly measure this transition, experiments or simulations in which torsion is gradually increased would be needed to be performed. Unfortunately, such experiments are beyond the scope of this study, as the superhelical density remained constant throughout the simulations/experiments.

Lastly, when sufficient superhelical stress is imposed to B-DNA, it becomes unstable and is capable to transition to a wide range of sequence-dependant conformations such as Z-DNA (left handed helix), unwound strands or cruciforms that may absorb superhelical stress and prevent defect formation at other sites [117]. The minicircles analysed in this study do not present this type of sequences, and the defects detected are much smaller than those observed in 2-5 kbp negatively supercoiled plasmids, where unwound strands extend from 40 to 60 bp [133]. However, the size of the bubbles detected in our simulations are of 1-2 bp (see figure 3.7b-c), which are of similar size to those observed at the tip of plectonemic loops [100]. We can then conclude that DNA defects and strong bends are frequently found in nature and are incredibly important in DNA recognition processes such as DNA damage detection [64] or transcription regulation, where they can reduce the distance between enhancer and promoter [88].

3.7.2 Planarity and aspect ratio measurements of negatively supercoiled DNA minicircles

Distributions of deviation from planarity show a low deviation from planarity for both 260 bp and 339 bp minicircles, which is less than 15% (5 nm) on average (see figure 3.9). These planar conformations do not present great fluctuations, although some regions along the DNA may present a high deviation from planarity being around 4 nm, which is still less than 20 % of the longest distance in the molecule (see equation 3.14). The 339 bp minicircle with $\Delta Lk = -1$ is an exemption as it is the least planar structure, having an average deviation of planarity around 4 nm, which is still less than 15 % compared to the longest distance. For this system, the introduced negative supercoiling induces a kink (see figure 3.8) that only allows a partial relaxation of the structure, which then adopts a less planar conformation. Higher degrees of negative supercoiling allow the 339 bp minicircle to adopt more planar conformations due to the presence of defects that enable the relief of higher levels of superhelical stress. The low deviation from planarity obtained by simulations is a favourable result, as structural analysis performed by AFM imaging in planar molecules present less distortions from surface immobilisation.

Regarding the aspect ratio distributions, relaxed minicircles appear as open rings with high aspect ratio (0.8) throughout the simulations (see figure 3.9). From ΔLk =0 to -2, we observe a trend in which increasing the level of negative supercoiling decreases the aspect ratio. This is because at $\Delta Lk = -2$, the molecule tends to adopt plectonemes, which consist of elongated intertwined helices. If the minicircles are further supercoiled ($\Delta Lk = -3$), the aspect ratio increases because the induced kinks allow the molecule to adopt more planar and less rectangular structures. Further increasing the torsional stress ($\Delta Lk = -6$), the kinks can sustain high bends ($\theta > 150°$,

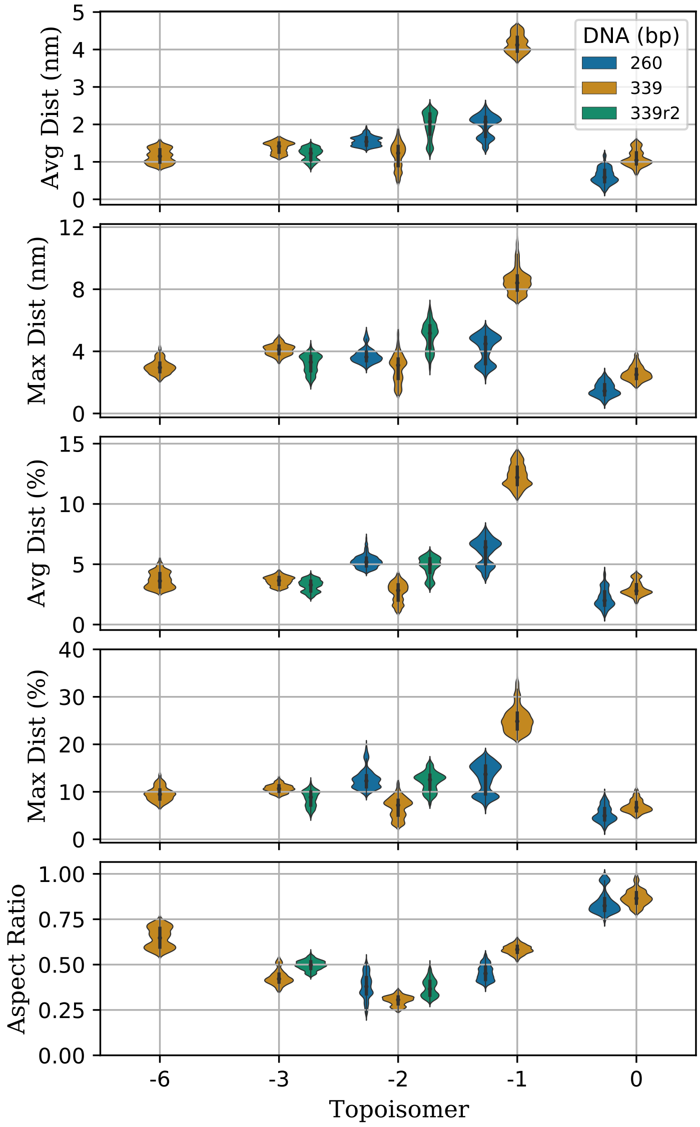


Figure 3.9: Distributions of deviation from planarity and aspect ratios in both 260 bp and the two replicas of the 339 bp minicircles (339 and 339r2) as a function of negative supercoiling. For every system, the average deviation from planarity is less than 3 nm (8%) except for the 339 bp minicircle with 1 turn undertwisted, which approximately increases to 4 nm ($\sim$ 15%). Similarly, for every system the maximum deviation from planarity at any region along the molecule is less than 8 nm ($<$ 20%), except for the 339 bp minicircle with $\Delta Lk = -2$, which has a value of $\sim$ 8 nm ($\sim$ 30%). The aspect ratio is reduced almost by half its value when transitioning from relaxed structures to -2 topoisomers. Further increasing negative supercoiling gradually restores the aspect ratio.

see figure 3.8) allowing the molecule to adopt a trefoil shape (see figure 3.7c), which has higher aspect ratio (~ 0.6) than previous supercoiled structures. These results agree with the aspect ratios observed in the experimental counterpart (see figure B.3c), where similar measurements are provided by both approaches.

However, in the experimental approach they observe what would seem to be a counter-intuitive behaviour when the DNA is being untwisted by approximately two turns (see figure B.3c), in which the aspect ratio increases. This is because at this level of superhelical stress ($\Delta k = -2$), one or two kinks can form and allow the structure to relax and adopt conformations with higher aspect ratios; while conformations in which kinks do not form, the resulting structures tend to be more coiled and with lower aspect ratios (see figure B.4). In the MD case, the 339r2 replica has a higher aspect ratio since it is a structure with no defects, while the first replica presents defects that allow the structure to relax and acquire a smaller aspect ratio (see figure 3.9 & figures B.4-B.5). This feature can be better visualised by analysing the time-series of aspect ratios in figure 1B.5, where the two replicas present opposite behaviours, where the aspect ratio of the less stable structure (cyan line: 339r2 bp; $\Delta k = -2$) tends to fluctuate more and towards higher aspect ratios, while the more stable structure (brown line: 339 bp; $\Delta k = -2$) tends fluctuate less and with lower aspect ratios due to the relaxation provided by the defects. These observations demonstrate that superhelical stress globally compacts DNA, which in combination with defects allow to relieve the imposed torsion.

Lastly, we also analysed time-series of the deviation from planarity parameters as well as aspect ratios (see figure B.5). In general, the deviation of planarity parameters present relatively low fluctuations, indicating a well equilibration and convergence. Aspect ratios do not greatly fluctuate for most structures as well, however, in structures with no torsional stress ($\Delta k = 0$) or where defects might not form ($\Delta k = -2$), the aspect ratio tends to vary through the simulation. In case of relaxed structures with $\Delta k = 0$, the minicircles are free to acquire conformations with high aspect ratios from 0.8 to 1.0. In case of structures with $\Delta k = -2$, defects might not form and the stress cannot be relaxed. This unrelieved stress is capable of inducing shapes with aspect ratios between 0.2 and 0.4. We consider this variety of conformations an indication of good sampling, which allows us to explore multiple behaviours of supercoiled DNA structures, the role of defects in relaxing the torsional stress, and their impact on the overall shape of DNA minicircles.

3.8 Conclusion

In this chapter we have presented SerraLINE, which is a simple to use open software that calculates bending profiles as well as global structural parameters of DNA that are suitable for comparison with experimental measurements. These global measurements correspond to compaction parameters such as the aspect ratio and its components, as well as deviation from planarity.

Bending profiles at the critical length of 16 bp, allowed us to observe regions in negative supercoiled DNA minicircles in which the DNA may present defects that disrupts the B-DNA structure. We determined that bending angles higher than 75 degrees correspond to DNA defects, which can be melting bubbles or type I/II kinks. These defects act as flexible hinges that allow short DNA loops to be highly bent. Negative supercoiling is the driving mechanism of the defects formation, where we discovered that the onset of these defects is in the range of physiological superhelical levels between -0.03 and -0.06 ($\sigma = -0.06$ in vivo).

Our deviation from planarity measurements indicate that the circular structures are approximately planar. On average, the deviation from planarity was less than 15% for all structures, even in the less planar minicircle. This result is advantageous for the experimental counterpart, since planar molecules make AFM imaging less distorted from surface immobilisation.

By analysing the aspect ratios of minicircles, we discovered that the DNA decreases its size by half from ΔLk =0 to ΔLk =-2. On the other hand, further increasing the supercoiling level ($\Delta Lk < -2$) causes the aspect ratio to increase. The AFM counterpart measured similar magnitudes of aspect ratios, however, they obtained higher aspect ratios for the -2 topoisomers. The second simulation replica indicates that, at this superhelical level, the torsional stress can induce kinks that allow the relaxation of the molecule, which results in an decrease of the aspect ratio; however, when kinks do not form, the molecule adopts a less stable structure with higher aspect ratio. These results further corroborate that torsional stress can induce defects formation that allow the DNA to relax and reduce its aspect ratio. This property is particularly important in DNA packaging, as defect formation caused by torsional stress may be the underlying mechanism for reducing the size of the DNA molecule.

Lastly, results in this chapter demonstrate that the features that SerraLINE offers are greatly exploited when mutually complemented with experimental approaches. It is worth mentioning that one of the potential applications of SerraLINE is that it can be also used to characterise systems where proteins are interacting with the DNA molecule. Particularly, the program has been used in our group to characterise the DNA binding modes induced by a protein through the comparison of simulations with AFM experiments [170]. Thanks to the collaborations that we have made [125,170], we have been able to further expand the program by introducing features such as 'plane projection' and 'resolution of tangent lengths'. SerraLINE in combination of AFM experiments, has allowed us to further expand our understanding regarding the impact of DNA supercoiling on the structure of double-stranded DNA. In the next chapter, we will introduce a program that we also developed for the analysis of not only structural properties of DNA but also elastic properties.

Chapter 4

SerraNA

Synopsis

SerraNA is an open software that we developed for the calculation of structural and elastic properties of NA at different length-scales, from the nucleotide level to the whole molecule using ensembles of numerical simulations. The program is a direct implementation of the LDEM, which allows the analysis and visualization of local properties that can be used to evaluate the molecule overall flexibility.

In this chapter we first describe the methodology to obtain local structural and elastic parameters and how to translate them into measures of overall flexibility. Afterwards, a general workflow is presented to show the general usage of the program and how it connects different inputs/outputs. Finally, results obtained with the program are presented to demonstrate its utility and suitability for comparison with experimental results, where it then is used to analyse sequence-dependant features as well as to visualize the emergence of bulk flexibility from local fluctuations using bendability as an example.

4.1 Program structure and usage

SerraNA is a program based in the LDEM [114], where the CEHS scheme [90, 91] is applied to calculate the BPP and BSP parameters (see subsections 2.2 & 2.1.3) and then it switches to the LDEM to consider pairs of bp separated by an increasing number of nucleotides (see subsection 2.2.1 & 2.2.2). The program is versatile, and can handle single or double-stranded structures of either DNA or RNA molecules, as well as linear or circular trajectories (minicircles).

The program SerraNA is an auto-contained program, where all the necessary elements for the proper functioning of the software are found within the program. Here we describe the mathematical procedures that SerraNA implements along with the components that compose the software and how these components are connected to produce the structural and elastic analysis from an input simulation.

The software workflow is managed by two main files: 'SerraNA.f90', which is in charge of calculating all the local parameters including the BPP, BSP, the structural and the elastic parameters; and the file 'Analysis.f90', which performs the analysis to calculate the global elastic constants. Additionally, a supportive file named 'Extract.f90' helps to process and filter the results for data visualisation. For the proper functioning of these files, a module named 'functions_mod.f90' contains all the necessary mathematical tools, while the module 'io_mod.f90' contains all the input/output functions and subroutines. Lastly, a file called 'parms.f90' contains the parameters used by the programs in the SerraNA software. With the help of the modules and parameters file, the main program 'SerraNA.f90' compiles the executable **SerraNA**, the analysis program 'Analysis.f90' compiles **Analysis** and the supportive program 'Extract.f90' compiles **Extract**.

Similarly to SerraLINE, SerraNA is a free access program written in Fortran 90 and available under version 3.0 of the GNU Lesser General Public Licence[*] at agnesnoy/SerraNA GitHub repository[†]. This licence was selected as it has few restrictions and is suitable for users/developers to integrate the software to external libraries and in case of any modifications, they must publish them using the same licence. Finally, through our GitHub page we provide SerraNA along with an example simulation composed of 250 frames and 32 bp to help the user familiarise with the software implementation. The documentation includes instructions for analysing the short simulation and provides a python script that uses the matplotlib library [60] to process the outputs for visualisation (see figure 4.1b,c,d).

Finally, for comprehensive guidance on accessing and installing SerraNA, please refer to the appendix section E.2. In addition, for specific program usage instructions, consult its accompanying manual in section F.

[*] https://www.gnu.org/licenses/gpl-3.0.en.html

[†] https://github.com/agnesnoy/SerraNA

4.1.1 SerraNA

Figure 4.1a) shows the workflow followed by SerraNA, where the main program takes as input Amber topology and trajectory files of a NA molecule. The main program **SerraNA** follows the procedures described in the methods subsection 2.1.1 to identify the ring atoms in each base and then fit a standard base to obtain the reference point $\vec{O}$ with orientation matrix R (see equations 2.4 & 2.3). This vector and matrix describe the position and orientation of each base conforming the molecule. The program considers N bases in case of single stranded DNA and N bp in case of double stranded DNA after being discarded the two bases/bp at each end in order to avoid end effects and temporal loss of base pairing.

For double-stranded structures, the set of six BPP parameters (shear S_x, stretch S_y, stagger S_z, buckle κ, propeller twist ω and opening σ) are calculated by following the procedures of subsection 2.1.2 using the positions and orientations of each bp. Then, for either single or double stranded DNA, SerraNA follows the procedures of the CEHS scheme (see subsection 2.1.3) to obtain the six BSP parameters (shift D_x, slide D_y, rise D_z, tilt τ, roll ρ and twist Ω) plus the bending angle θ. The program **SerraNA** then outputs the averages and standard deviations of the BPP and BSP parameters, which are compatible with the 3DNA program [90] and are saved in separate files in human readable format (see figure 4.1a).

SerraNA then switches to the LDEM [114] for obtaining the structural and elastic parameters beyond the dinucleotide length. Independently if the input molecule is single or double-stranded, SerraNA follows the methods described in section 2.2.1 to obtain geometric variables at lengths beyond the dinucleotide level. Figure 2.4 is a representation of how SerraNA implements the LDEM, where geometric variables are calculated for every possible oligomer ranging from 2 to N bp in length. For every oligomer a set of 11 structural parameters is calculated: twist Ω, roll ρ, tilt τ, added-shift X, added-slide Y, added-rise Z, the end-to-end distance L, contour length L_{CL}, bending angle θ, square of bending angle θ^2 and the directional correlation decay $cos(\theta)$. Once these parameters are calculated at every frame, SerraNA calculates the averages for each oligomer formed. To obtain an estimation of the curvature, SerraNA implements the rebuilding algorithm (see section 2.1.4) using the six averaged BSP to obtain an averaged structure. From this, the directional correlation $\cos(\theta_s)$ and the bending angle θ_s are calculated at every length-scale:

$$\begin{aligned} \cos(\theta_s) &= \hat{z}_{i,s} \cdot \hat{z}_{j,s} \\ \theta_s &= \arccos(\hat{z}_{i,s} \cdot \hat{z}_{j,s}) \end{aligned} \tag{4.1}$$

where the unit vectors $\hat{z}_{i,s}$ & $\hat{z}_{j,s}$ are tangent to the curve at bp i and bp j. The average structure bending angle θ_s, square of bending angle θ_s^2 and directional correlation $cos\theta_s$ are added to the set of 11 structural parameters and saved in a single human readable file, where the parameters are sorted by length and sequence.

If the simulation is longer than one frame, SerraNA proceeds to calculate the elastic parameters via the inverse covariance method detailed in section 2.2.2, where the end-to-end distance L, twist Ω, tilt τ and roll ρ, are used to calculate the elastic matrix F (see equation 2.53) which is conformed by four elastic parameters as well as the 6 respective elastic couplings. Additionally to the 10 elastic variables obtained from the inverse covariance method, SerraNA obtains the dynamic persistence length (A'_d) by combining tilt A_τ and roll A_ρ using equation 2.61. We refer to A'_d as the second estimation of the dynamic persistence length. Finally, the set of elastic variables is composed by: stretch B, twist C, roll A_ρ, tilt A_τ, stretch-twist D, stretch-roll H, stretch-tilt, twist-roll G, twist-tilt, tilt-roll, the dynamic persistence length A'_d, variance of the end-to-end distance $Var(L)$ and the partial variance of the end-to-end distance $Var_p(L)$. Similar to the structural parameters, the elastic parameters are calculated for each oligomer. Finally, the SerraNA program proceeds to write the elastic parameters in a human readable file.

One of the limitations of the program is that it can only handle full single-stranded or double-stranded molecules. An important fact to mention is that for linear/open structures, $N(N-1)/2$ set of structural/elastic parameters are calculated, while for circular/closed structures $N(N-1)$. Finally, the run time scales as $O\left(KN(N-1)/2\right)$ in case of linear structures and $O\left(KN(N-1)\right)$ in case of closed structures, where K is the total number of time frames.

4.1.2 Analysis

One of the powerful features that the SerraNA software offers lies in the second main program **Analysis**. Having calculated the structural and elastic parameters, the **Analysis** program can proceed to calculate the bulk elastic constants (see figure 4.1e), which evaluate the molecule overall response to stretching, twisting and bending deformations.

To calculate the bulk elastic constants, SerraNA requires two intervals:

1.- Interval $[a, b]$ indicates the section of the molecule in which the program will focus to measure the flexibility, where a and b are the bp indices and $a < b$.

2.- Interval $[N_a, N_b]$ which indicates the lengths to be considered in the calculation of elastic constants, where the lengths are in base-steps and $N_a < N_b$. The allowed lengths must fall within the $[a : b]$ interval, hence $N_a > 1$ and $N_b <= b - a$.

A different type of intervals can be established for each elastic constant.

For the torsional modulus, SerraNA sets $N_a = 11$ and $N_b = N - 10$ as default to capture the bulk behaviour of twist and to avoid averaging C over less than ten oligomers using equation 2.55. However, if these are not the desired lengths/intervals by the user, the program has the option to modify the defaults lengths/intervals (see equation 2.55). The uncertainty of the estimated twist elastic constant is measured through the standard error, which is calculated using equation 2.68. Similarly to the twist elastic constant C, for tilt A_τ, roll A_ρ, the dynamic persistence lengths A'_d and A''_d are calculated using $N_a = 11$ and $N_b = N - 10$ by default to capture bulk behaviour and have statistics of at least 10 oligomers per length. To measure the accuracy of these four elastic constants, the standard errors are calculated with equation 2.68.

The global stretch modulus cannot be obtained via an average as a function of length because it follows a complex behaviour as it was explained in the methods section 2.2.5 (see the top graph of figure 2.8a). Instead, the stretch modulus is obtained through a linear fit of the partial variance of the end-to-end distance ($Var_p(L)$) (see bottom graph of figure 2.8 and equation 2.77) that considers the central 18mer of the molecule and goes through the lengths $N_a = 8$ to $N_b = 17$ base-steps. As stated in the methods subsection 2.2.6, the accuracy of the global stretch (B) is measured through the confidence interval of the linear fitting (see equation 2.80) set to 70%. For molecules shorter than 18 bp, SerraNA considers the whole fragment ($[a = 1, b = N]$) and sets the interval $[N_a = 8, N_b = N - 1]$ by default, and in the case that the molecule is shorter than 9 bp, then the second interval is set to $[N_a = 1, N_b = N - 1]$.

Regarding the persistence length A and its static A_s and dynamic A_d contributions (see equations 2.82-2.85), linear fits are performed through lengths $N_a = 1$ and $N_b = N - 10$. Similarly to the stretch modulus, confidence intervals are obtained following equations 2.88-2.90 at the level of 70% .

It is worth mentioning that we choose to estimate the elastic constants of stretch modulus B, persistence length A and its static A_s and dynamic A_s contributions through linear fits, because it is a simple and effective method for describing/modelling data that has linear tendencies. In case of tangent-tangent correlation decays, linear fits are suitable when the lengths considered are smaller than the DNA persistence length, as at longer lengths the decays would start to exhibit an exponential form rather than linear. For naked DNA, most trajectories would fall within this regime as the DNA persistence length is around 50 nm which is approximately 150 bp. In case of the stretch modulus, the default options of SerraNA perform the linear fit on the partial variance of the end-to-end distance ($Var_p(L)$) in the range $[N_a = 8, N_b = 17]$, which previous evidence indicates that this region is well described by a linear function [114]. In the case of the naked DNA simulations presented in this thesis, figure 4.5 shows that indeed, the simple linear fits are suitable for estimating the aforementioned elastic constants.

SerraNA combines the static A_s and dynamic A_d persistence lengths using equation 2.60 to calculate a persistence length $\tilde{A}$ which should be compatible with the persistence length A previously obtained through the linear fit ($\tilde{A} \approx A$). Analogously, SerraNA combines the second estimation of the dynamic persistence length A'_d with the static component A_s resulting in a second prediction of the persistence length A'. Since A'_d has higher values to A_d because linear contributions of twist and stretch have been removed, A' is stiffer than A.

As a summary, **Analysis** uses the structural and elastic parameters to calculate 10 global elastic constants that evaluate the molecule overall response to stretching, twisting and bending deformities, these being: tilt (A_τ), roll (A_ρ), twist (C), stretch modulus (B), persistence length length (A), static persistence length (A_s), dynamic persistence length (A_d), persistence length ($\tilde{A}$), second estimation of the dynamic persistence length (A'_d) and second estimation of the persistence length (A'), where B, A, A_s, A_d are obtained through linear fits, A_ρ, A_τ, C, A'_d through the inverse co-variance analysis (using elastic matrix F), and $\tilde{A}, A'$ combining both approaches by using equation 2.60. Figure 4.1e shows the printed information by **Analysis** of a 32bp DNA where 28 bp are analyzed instead of 32 bp as **SerraNA** discarded two bp at each end in order to avoid end-effects. Finally, it is worth mentioning that the notation of the persistence lengths outputted by the **Analysis** program changes slightly as this information is printed to the computer screen (see figure 4.1e). These changes in notation are as follows: A [a] = A, As [a] = A_s, Ad [a] = A_d, A [b] = $\tilde{A}$, Ad [c] = A'_d, A [d] = A', where quantities with [a] are calculated through linear fits, the persistence length with [b] is calculated combining A_s & A_d using equation 2.60, the dynamic persistence length with [c] is calculated through the inverse co-variance analysis, and the persistence length with [d] combines A_s with A'_d using equation 2.60 as well.

4.1.3 Extract

The four human readable files outputed by **SerraNA** can be processed with the supportive **Extract** program that can filter the information to produce simple files that are ready to be plotted and processed by external scripts.

In case of the structural and elastic files, two types of filtering can be done:

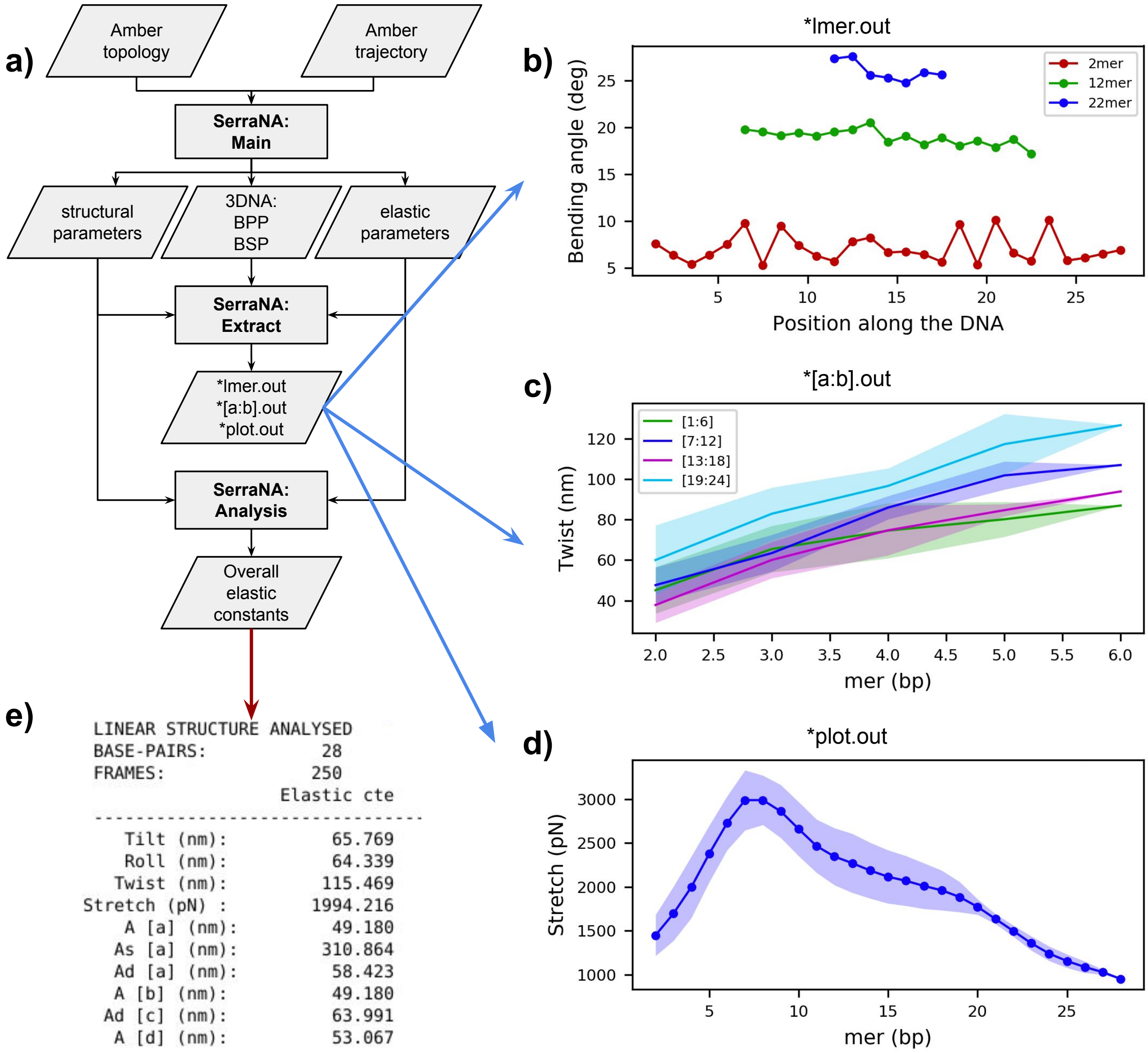
- The parameters can be filtered by lengths, where their profiles are sorted by their position along the molecule. This produces files with extension '*lmer.out'.
- Length-dependent parameters are calculated in a region $[a, b]$ specified by the user (see previous subsection 4.1.2), producing the file with extension '*[a:b].out'.
- The last type of filtering also produces length-dependant parameters profiles, where the whole molecule is considered ($[a = 1, b = N]$). This type of filtering produces the file '*plot.out'

Figure 4.1b-e) shows examples of the outputs produced by SerraNA using a short MD simulation of a 32bp long DNA composed by 250 frames, which is available at the agnesnoy/SerraNA GitHub repository. Starting with the first type of filtering, extracting parameters at particular lengths is useful for observing how the structure/elasticity evolves along the molecule without losing sequence-dependent features (see figure 4.1b). Regarding the second type of filtering, plotting length-dependent parameters at particular regions is useful for analysing how the variables behave at certain sections of interest in the molecule. The twist elastic constants is plotted at different sections of the molecule in figure 4.1c) by using multiple '*[a:b].out' files. And finally, plotting length-dependant parameters of the whole molecule using the '*plot.out' files is a fast option for analysing and visualising how the flexibility/structure of the whole molecule evolves as a function of length. Figure 4.1d) shows that the stretch modulus of this short simulation adopts the complex shape previously observed in the LDEM [114] and shown in the methods section 2.2.5 and in the top panel of figure 2.8a).

More information of specific commands and inputs/outputs is available in the program's manual located in the appendix section of this project (Appendix F) and at the GitHub repository.

4.2 Results and discussion

To demonstrate the versatility and usefulness of SerraNA, we will now proceed to show elastic and structural results obtained from distinct DNA systems which include free



```
LINEAR STRUCTURE ANALYSED
BASE-PAIRS:                28
FRAMES:                   250
                     Elastic cte
---------------------------------
    Tilt (nm):           65.769
    Roll (nm):           64.339
   Twist (nm):          115.469
Stretch (pN) :         1994.216
     A [a] (nm):         49.180
    As [a] (nm):        310.864
    Ad [a] (nm):         58.423
     A [b] (nm):         49.180
    Ad [c] (nm):         63.991
     A [d] (nm):         53.067
```

Figure 4.1: Workflow followed by SerraNA using a DNA of 32 bp as an example taken from agnesnoy/SerraNA GitHub repository. (a) Given an input simulation the main program calculates the BPP, BSP, structural and elastic parameters, which then are processed by the supportive program **Extract** to create data files for (b) plotting profiles along the molecule for specific lengths (*lmer.out) and to (c) plot length-dependant parameters in the range [a,b] (*[a:b].out) or (d) considering the whole fragment (*plot.out). (e) The Analysis program can then use the structural and elastic parameters to calculate the 10 elastic constants, where [a] indicates persistence lengths obtained through linear fits, A [b] corresponds to $\tilde{A}$, Ad [c] to A'_d and A [d] to A'. The repository includes a python script to process the outputs and to create the plots (b,c,d) using the matplotlib library [60].

DNA, protein-DNA complexes and sequence DNA mismatches. We separate these systems into two sets of simulations: one composed of canonical linear DNA, the other composed of protein-DNA complexes as well as sequence mismatches.

The set of canonical DNA simulations consists of four linear DNA fragments of 32, 42, 52 and 62 bp long fragments that were extracted from longer sequences analysed in [104, 160] (see appendix A.2) which here we refer to them as 32mer, 42mer, 52mer and 62mer. These four sequences were chosen since there are theoretical and experimental estimations of their persistence lengths (see table 4.1). Two additional 32 bp long fragments are also added to this set: one constructed with the same sequence as the 32mer but built with the parmOL15 force-field [173, 174] that we name 32ol15; and one built with random sequence taken from the BIGNASIM database [59] which here we name 32rand.

For the second set, two simulations of protein-DNA complexes were analysed: one consisting of a nucleosome with PDB ID 1kx5; and the other consisting of a DNA bound to the transcription factor GCN4 with PDB ID 2dgc. Regarding the sequence mismatches, two 13 bp long oligomers were analysed, one with an A:A mismatch on the middle and the other with G:G. Both protein-DNA complexes and sequence mismatches simulations were also obtained from the BIGNASim database [59]. More information regarding the sequences and simulation conditions are located in the appendix section A.4 .

Additionally, we analysed the sequence-dependent elastic properties of all the distinct 136 tetranucleotide sequences obtained from the set of MD simulations of the ABC consortium [119], consisting of 39 DNA fragments made of 18 bp. More information regarding simulation conditions, procedures and DNA sequences can be found in the appendix section A.

4.2.1 Gaussian test and convergence

One central assumption of the LDEM is that the structural variables that capture elasticity are assumed to be Gaussian distributed [114]. Before analysing the elastic properties of our trajectories with SerraNA, we decided to first test if the structural variables in the simulations indeed follow a Gaussian distribution. Therefore, we performed a Gaussian test using quantile-quantile (q-q) plots, which is a graphical technique for determining if two distinct data sets come from populations with similar distributions [154]. For each trajectory, we built one data set per structural variable and tested it against a Gaussian distribution. Using the q-q plots we were able to obtain q-q correlation coefficients (R^2), which quantify how similar the two distributions are, where a correlation coefficient of $R^2 > .9$ indicates a high degree of similarity between the two sets. This technique allowed us to demonstrate that the four structural variables used in the inverse covariance analysis (see methods 2.2.2) are approximately Gaussian distributed as they present high q-q correlations ($R^2 > 0.98$) (see figure 4.2), with some exceptions in the twist angle due to the bimodal behaviour in some CG base-steps [25, 26] and an asymmetry observed in distributions of the end-to-end distance at long lengths (see figure C.1a-b). However, the bimodal behaviour

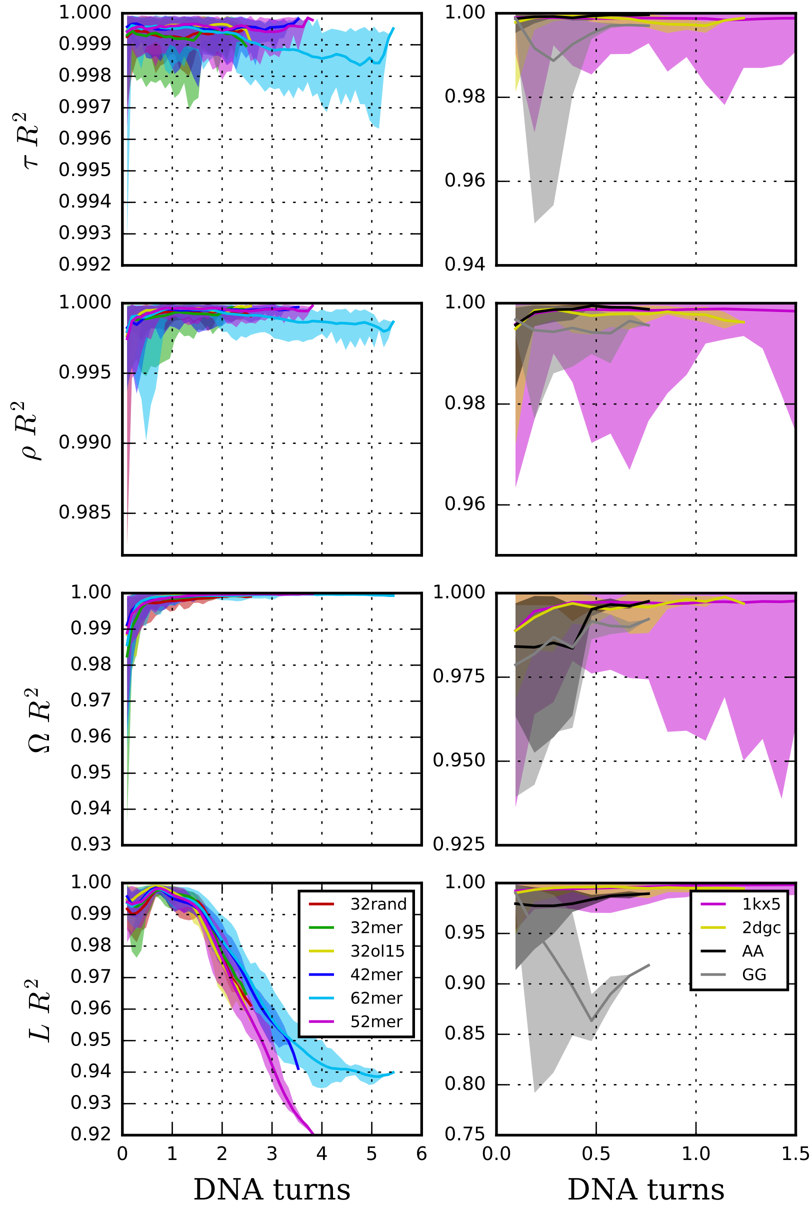


Figure 4.2: R-squared values of q-q plots' linear fits between distributions spanned by structural variables of tilt (τ), roll (ρ), twist (Ω) and the end-to-end distance (L) against the estimated Gaussian distributions. Averages are shown as solid lines while shaded areas represent the maximum and minimum values for each sub-fragment length. Graphs on the left correspond to the set of linear DNA simulations, where most of the cases pass the normality test ($R^2 > 0.95$), except for bimodal Ω at the dinucleotide level (see figure C.1a), and L at long lengths (see figure C.1b). Graphs on the right correspond to the set of perturbed DNA, where Gaussianity is affected but SerraNA can still provide valuable insight about how these deformations affect the structure and flexibility of DNA.

observed in CG base-steps is quickly suppressed and becomes Gaussian when the length is increased. We consider this to be a positive result as we discard short length-scale measurements when calculating global elastic parameters. It is still worth mentioning that non-Gaussianity at the local level is highly sequence-dependant [25,26] and might be relevant for local protein-DNA interactions as the DNA could exhibit more than one elastic behaviour. Lastly, Gaussianity was also tested for simulations with perturbed DNA, where the q-q correlation coefficients indicate that for most of the cases Gaussianity can still be assumed and described by the harmonic elastic model. The lowest R^2 value was observed for the GG mismatch trajectory at the length of $l = 4$ for the end-to-end distance L. By analysing the distributions for this trajectory we found a slight asymmetry in the distribution of the bp-step located at the GG mismatch position (see figure C.3a). Although the Gaussianity test was passed with an R^2 value higher than 97% for this case, the asymmetry resulted in skewed and slightly bimodal distributions at longer lengths (see figure C.3b).

We also checked whether simulations were sufficiently long to provide trustworthy measurements of elasticity. To answer this question we have monitored how the elastic constants evolve every 100ns of simulation time by implementing the default options in SerrraNA (see figure 4.3). Our results show that convergency is quickly reached for most of the elastic constants in our simulations, with some cases presenting a relative lack of convergence such as the A_s in the 62mer DNA. As it is explained in the following section 4.2.3, the static persistence length is one of the parameters that present larger fluctuations and therefore it is more difficult to estimate.

Lastly, for the set of simulations obtained from the ABC consortium, Gaussianity and convergency have already been tested in previous studies [119].

4.2.2 Twist elastic modulus

The twist elastic constant C as a function of length shows a transition from local to bulk behaviour within 1 DNA turn in all simulations as previously stated in the LDEM [114]. According to our results, at short length-scales, the DNA is more torsional flexible, where values of C range from 30-60 nm, and at longer length-scales the DNA becomes more rigid and then reaches a plateau that tends to be approximately 100 nm (see figure 4.4). Soft values of the twist modulus at short lengths are consistent with experiment techniques such as the analysis of crystallographic structures of DNA, fluorescence polarization anisotropy [42,55], SAXS measurements [136], and many MD studies [63,116,121,169], while the stiffer estimations at long lengths agree with many single-molecule experiments [13,85,107] as well as some modeling studies [73,82,96,169]. Applying the default analysis performed by SerraNA, we are able to filter local stiffness and average over bulk behaviour, where for all simulations the global twist elastic constant is around 96.6 nm on average (see table 4.1).

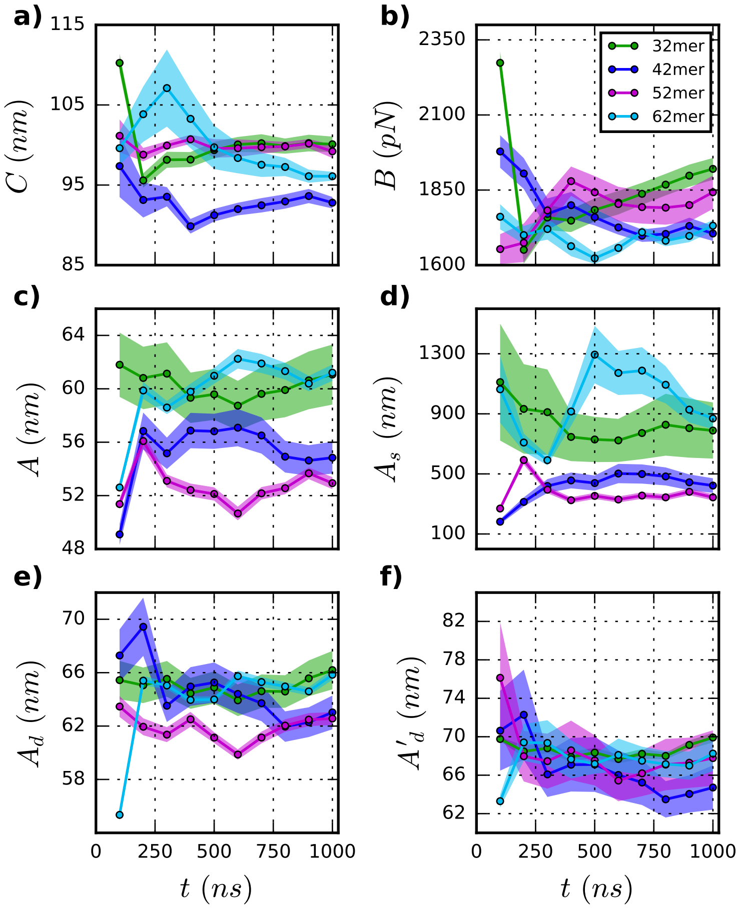


Figure 4.3: Elastic constants of twist (C), stretch modulus (B), persistence length (A), static persistence length (A_s), dynamic persistence length (A_d) and the second estimation of the dynamic persistence length (A'_d) measured at distinct accumulative simulation time for the 32mer, 42mer, 52mer and 62mer. Lines represent the predicted elastic constants, while shaded areas represent standard deviations in case of C and A'_d, and uncertainty at 70% of the linear fit in case of B, A, A_s and A_d.

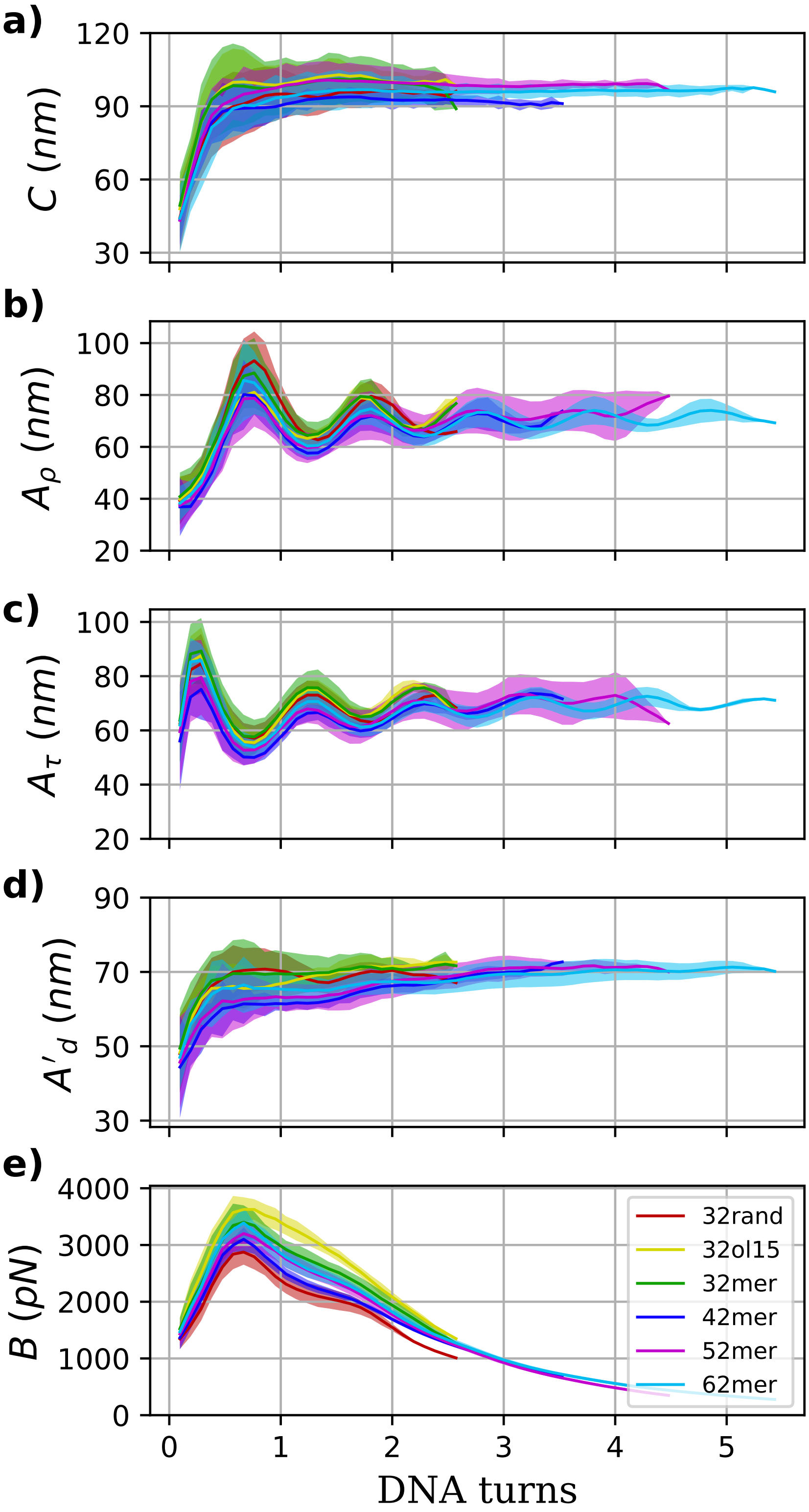


Figure 4.4: Elastic constants $(C, A_\rho, A_\tau, A'_d, B)$ as a function of length and obtained through the inverse-covariance method, where A'_d was obtained by combining A_τ and A_ρ using equation 2.61. For each simulation, averages over sub-fragments with the same length are plotted as solid lines, while their standard deviations are represented as shaded areas.

DNA	C (nm)[a]	A (nm)[b]	A_s (nm)[b]	A_d (nm)[b]	A'_d (nm)[a]	A'(nm)[a]	B (pN)[b]
32rand	94.8 ± 0.8	57.4 ± 1.6	473 ± 87	65.4 ± 0.6	68.3 ± 1.0	59.7	1696 ± 15
32mer	100.1 ± 1.0 *101.4 ± 1.2*	61.1 ± 1.1 *58.7 ± 1.4* $\underline{56.3}$ **50.5 ± 2.1**	789 ± 93 *562 ± 74*	66.2 ± 0.7 *65.5 ± 0.8*	69.9 ± 0.5 *69.0 ± 1.1*	64.2 *61.4*	1920 ± 18 *2207 ± 41*
42mer	92.8 ± 0.7	54.8 ± 0.6 $\underline{54.8}$ **45.5 ± 0.5**	422 ± 24	63.0 ± 0.6	64.7 ± 2.3	56.1	1705 ± 12
52mer	99.2 ± 0.8	52.9 ± 0.2 $\underline{51.5}$ **45.5 ± 0.8**	344 ± 10	62.6 ± 0.2	67.8 ± 2.9	56.6	1843 ± 27
62mer	96.1 ± 0.6	61.2 ± 0.3 $\underline{51.5}$ **41.7 ± 0.5**	869 ± 36	65.8 ± 0.2	68.2 ± 1.8	63.3	1731 ± 9
Average	96.6 ± 2.7	57.5 ± 3.3	579 ± 210	64.6 ± 1.5	67.8 ± 1.7	60.0 ± 3.3	1779 ± 88
Experiments	[90, 120]	[45, 55]	[1100, 1500]				

Table 4.1: Global elastic constants that evaluate the overall torsional, bending and stretching stiffness of the set of free DNA simulations. Parameters in italics correspond to elastic constants of the simulation with OL15 force-field, underlined and bold text correspond to persistence lengths calculated from MC simulations [104] and experimental [160] approaches (respectively). The last row shows ranges of values for the elastic constants of twist [13, 85, 107], persistence length [9, 56, 85, 103, 165] and stretch modulus [50, 147] reported in the literature and measured by multiple experimental setups. [a] Averages and standard deviations of elastic constants obtained through the inverse-covariance method (C, A', A'_d) were calculated using default SerraNA options (see sub-section 4.1.2). [b] Estimations and uncertainties of persistence lengths (A, A_s and A_d) and stretch modulus (B) were obtained through linear regressions with confidence levels at 70% (see sub-section 4.1.2).

4.2.3 Persistence length

The persistence length (A), its static (A_s) and dynamic (A_d) components, were obtained through linear fits of tangent-tangent correlations using the default options of the **Analysis** program (see 4.1.2). In general, our estimations of these quantities for our set of linear DNA simulations are slightly stiffer than experimental results [160] and are within the same range than simulation studies [104] (see table 4.1). More specifically, we obtain values around 57.5 ± 3.3, while experimental results get predictions between 45-55 nm [9, 56, 85, 103, 165] and simulations get predictions in the range of 51-56 nm [73, 104, 169]. There are multiple factors that could explain the discrepancies between simulation and experimental predictions. One possible reason might be due to the difference between ionic solutions. In MD simulations, counterions are highly controlled and usually consist of monovalent ions like Na and K, while the experimental buffers are formed by a variety of ionic species comprising Hepes, Tris or EDTA [56, 85, 103, 160, 165] which are known to affect the DNA overall flexibility [9, 52, 169]. Other possible factors for the discrepancies between these studies including ours, could be due to inaccuracies in the modelling methods as well as sequence-dependant features, however, it is difficult to assess them due to the limited number of oligomers and without comparing with the same exact sequences (our sequences are extracted from longer molecules, see appendix A.2).

The static directional decay have strong oscillations in phase with the DNA periodicity along the length. This is due to the DNA intrinsic curvature [114] and might be exploited in processes such as protein-DNA recognition [80, 95] and DNA loop formation [120]. It is also interesting to note that values of A_s are greater in magnitude and have more variability than values of A_d ($A_s = 579 \pm 210$ nm and $A_d = 64.6 \pm 1.5$ nm). But this is not a strange finding, previous MC simulation studies have already observed this behaviour [104] and it might explain why some experimental results differ in values of both A_s and A_d ($A_s \approx 130$ nm and $A_d \approx 80$ nm [10]; $A_s > 1000$ nm and $A_d \approx 50$ nm [162]). The variability of the static directional decay complicate the estimation of A_s which in consequence affects A. The limited molecular lengths achieved by our simulations (smaller than 100 bp), make challenging the estimations of A_s due to strong oscillations compared to small decays. This is reflected in the relative low convergence observed in some structures (see figure 4.3) and in the broad confidence intervals obtained (see table 4.1). Another important observation to consider is that different force-fields (BSC1 and OL15) might yield discrepancies in the estimations of A which are mainly caused by the A_s component rather than A_d. SerraNA can be used as a tool to validate realistic measurements of DNA flexibility when designing force-fields.

An important thing to remember is that the confidence intervals of linear fits are calculated under the assumption that the residuals ϵ_i are normally distributed (see methods section 2.2.6). We calculated q-q plots to test these assumptions in figure C.2, where we found that the residuals for A and A_s are normally distributed as their ordered values follow a linear behaviour, and for all trajectories, we obtain R^2 values higher than 90%. In the case of the dynamic persistence length (A_d), for all cases we obtained R^2 values around 90%, which indicate that the residuals can be considered to follow a normal distribution, however, in the case of the 62mer and specially the 52mer, we observe that the shape of the q-q plots slightly presents bimodal behaviour (plots

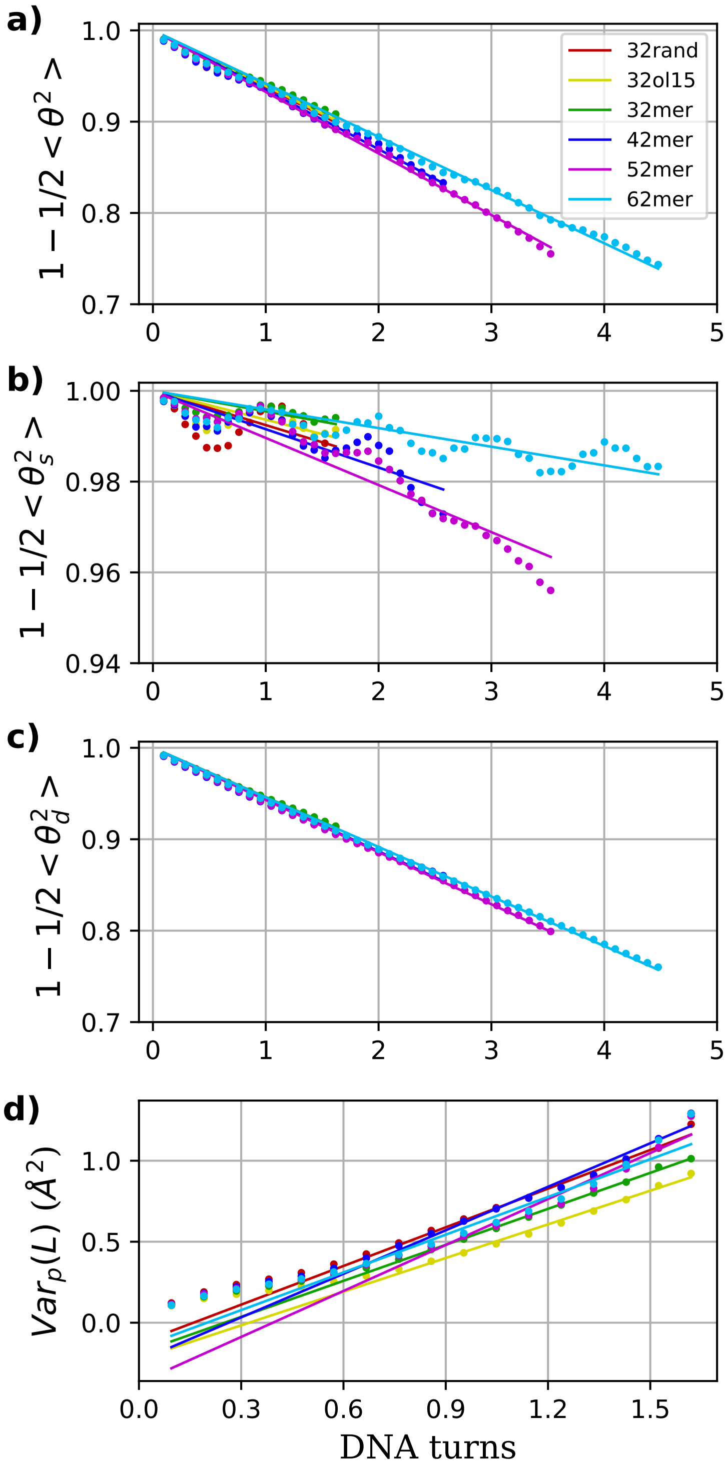


Figure 4.5: Persistence length (a) along with its static (b) and dynamic (c) components which are obtained through linear fits of the directional decays associated with $\langle\theta^2\rangle$, $\langle\theta_s^2\rangle$ and $\langle\theta_d^2\rangle$, respectively. Parameters used in the calculation of the persistence lengths are averages over all sub-fragments with the same length, where the ten longest lengths have been discarded (see sub-section 4.1.2). (d) The stretch modulus (B) is estimated through a linear fit of the partial variance of the end-to-end distance, where the fit goes through the central 18mer and from 8 to 17 base-steps (see sub-section 4.1.2).

have an S shape). According to these results, we can conclude that the residuals in the A_d elastic constant, present more deviations from a normal distribution than A and A_s. In the future, it would be interesting to check if this bimodality persists for longer DNA fragments as the number of residuals ϵ_i depend on the length of the molecule in base-pairs, which in case of our DNA molecules, this number is rather limited as these molecules have short lengths.

Focusing on the inverse-covariance analysis, the resulting dynamic persistence lengths (A'_d) have higher values compared to the first estimation ($A'_d = 67.8 \pm 1.7$ nm on average, see table 4.1), which consequently provides a higher value for the second estimation of the persistence length ($A' = 60.0 \pm 3.3$nm) when mixed with A_s. The reason behind this increase in stiffness is related to the fact that thermal fluctuations correlated to tilt and roll are filtered out via the partial variances (see methods section 2.2.2). Figures 4.4b-c shows tilt (A_τ) and roll (A_ρ elastic constants as a function of length, where they reach bulk behaviour within one DNA turn with periodic oscillations along the length. The profiles of A_τ and A_ρ are anti-symmetric due to the bending anisotropy, but at half and complete DNA turns they align because grooves and backbones face evenly towards both bending components (see figure 2.4). Combining A_τ and A_ρ produces A'_d shown in figure 4.4d, where the oscillations have disappeared and the resulting profile resembles to the twist profile.

A final thing to note about the dynamic persistence length is that we have compared the second estimation A'_d with the third estimation A''_d that is calculated from tilt, roll and the twist-roll coupling (G) as stated in equation 2.66. However, the elastic profiles of both quantities are visually indistinguishable (see appendix C.4), and their bulk constants only differ by 0.2 nm (see table C.1). Introducing the coupling G do not provoke major changes in the dynamic persistence length as the linear correlations of other variables (including twist) have already been removed when calculating the partial variances.

4.2.4 Stretch modulus

For all linear DNA simulations, the profiles of the stretch modulus (B) follow the non-monotonic behaviour described in previous studies [114], as well as reproduced by others [169] (see the methods subsection 2.2.5). At short lengths, the stretch modulus increases considerably up to the length of 7 bp where a maximum is reached. The stretch modulus obtained from the contour length (see figureC.5) has similar behaviour in this range of short lengths, which demonstrates that it is mainly caused by base-stacking interactions. At longer lengths B presents two softenings due to the coordinated motion of base-pairs. The first softening appears at the length of 13 bp (around 1 DNA turn) where an essential mode appears and causes a plateau in B, which would correspond to the stretch modulus captured by force-extension experiments [50, 147]. At longer lengths (around 2 DNA turns), the stretch modulus presents the second softening which makes it even more flexible and is caused by long-ranged end-effects [114].

As stated before in the LDEM [114], the fluctuations of the end-to-end distance are

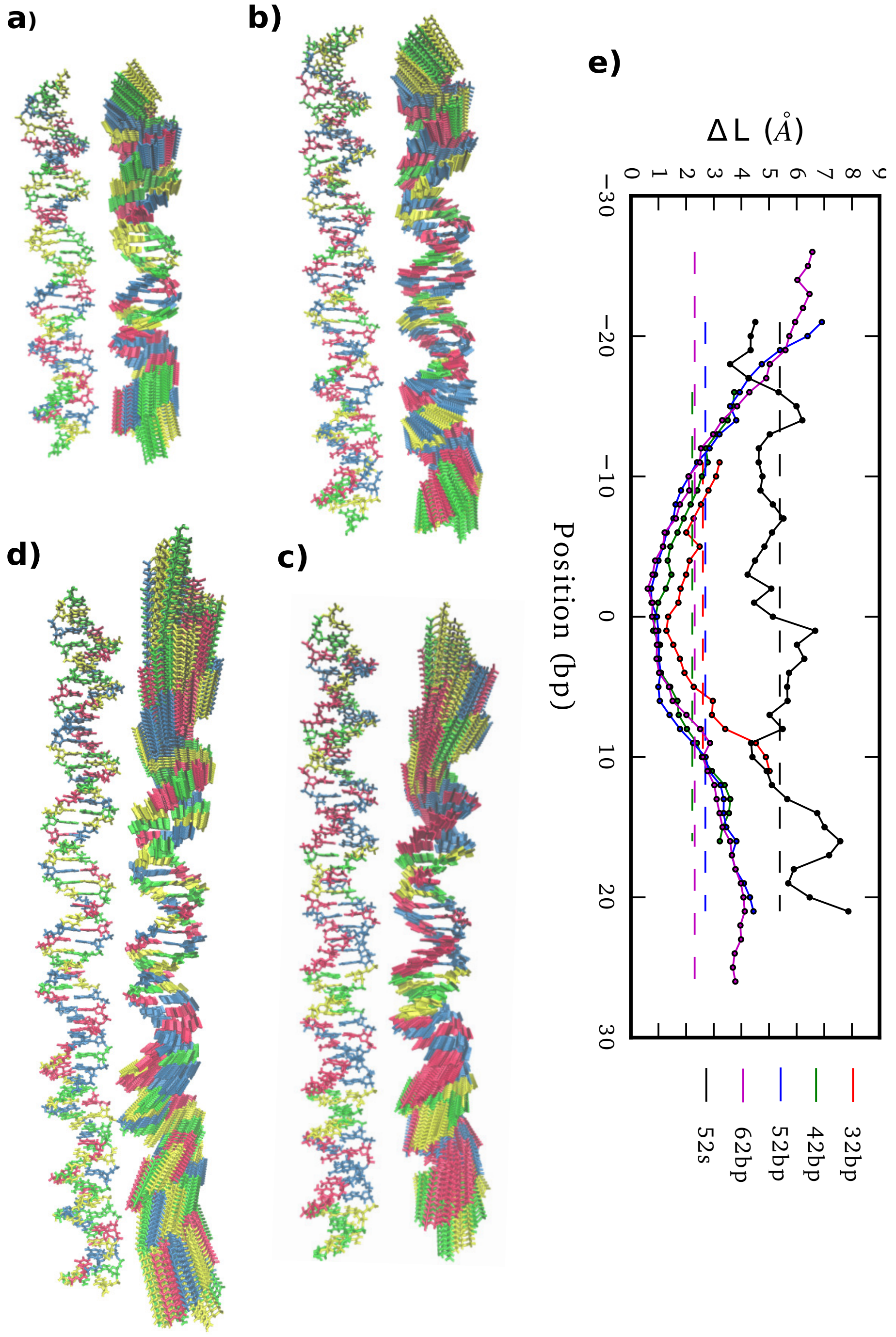


Figure 4.6: Averaged structures along with their corresponding stretching modes for the 32mer (a), 42mer (b), 52mer (c) and 62mer (d). Plot of base-pair/position dependence of local increments in the end-to-end distance (ΔL) using the sub-fragment length of 5 bp (e), caused by the stretching modes from the relaxed (averaged) structures and by the pulling simulation (52s). Dashed lines indicate the average increments.

able to capture multiple essential modes in which the base pairs move in a coordinated way. These coordinated motion of base-pairs cause a non-linear increase in the variance of the end-to-end distance, inducing the apparent softening in the stretch modulus at long lengths. Using PCA [138] (see methods section 2.6) we are able to obtain an essential mode that captures vibrations from the edges and provokes the incredibly soft B at long lengths (see figure 4.6a-d). This stretching end mode is consistent across our simulations even though they have different lengths, which suggests that the characteristic length of this mode has not been reached by the simulations analysed in this project, being larger than 5 DNA turns. As a corroboration, the increments in L (ΔL) are higher the farther from the centre of the molecule (see figure 4.6e). We performed an additional simulation where the DNA was pulled as in [135] (simulation named 52s, see the appendix section A.3 for more details). The increments ΔL in 52s are uniformly distributed along the molecule (see figure 4.6e), which further demonstrates that the coordinated motion at the ends is just due to a vibrational mode that is not relevant for the extraction of the intrinsic stretch modulus of DNA. Nonetheless, as this mode is present in all our linear simulations, it suggests that this mode is real (not an artifact) however, as stated before the coordinated motion of base-pairs characteristic of this mode causes the apparent softening in the stretch modulus estimated by the end-to-end distance.

As suggested by previous studies [114] and following the process described in the methods section 2.2.5, we are able to filter the end stretching mode and estimate the stretch modulus via the linear fit of the partial variance of end-to-end distance ($V_p(L)$) (figure 4.5d). Our estimated values of B are on average 1779 ± 88 pN (see table 4.1), which are relatively close to experimental values of 1500 pN [50].

The confidence intervals for the stretch modulus are based under the assumption that the residuals resulting from the linear fit adhere to a normal distribution. To test this assumption we calculated q-q plots (see figure C.2), where our analysis indicates that the residuals indeed follow a normal distribution as the R^2 are approximately 90%. However, in the case of the 32mer case, we observe some deviations from normality which may affect the accuracy of our stretch modulus prediction. Although the R^2 value for this case is 78%, we still consider our confidence intervals to be valid given the overall normal behaviour of the residuals.

4.2.5 Transition from local to global behaviour

One of the main features SerraNA offers is to visualize how the elastic and structural properties emerge from smaller length-scales. Here we show an example in 4.7 where we compare how the bending angle behaves along two distinct molecules (52mer and 62mer) at five length-scales ($l = 1, 7, 15, 27, 37$ base-steps). We choose these two fragments because the 52mer is the most bendable structure according the three persistence lengths (A, A_s & A_d), while the 62mer is the least bendable structure (see table 4.1). It is interesting to note that at the dinucleotide level ($l = 1$), bending angles are comparable between the two structures (7.1 ± 1.5 and 7.2 ± 1.1 degrees for the 52 and 62 mers, respectively), but at the length-scale of $l = 37$bp, they have distinct average values (35.6 ± 1.6 and 33.0 ± 0.7 degrees). At the intermediate lengths of 7, 15, and 27 base-steps, the shape in the profiles are clearly different, where a periodicity can be

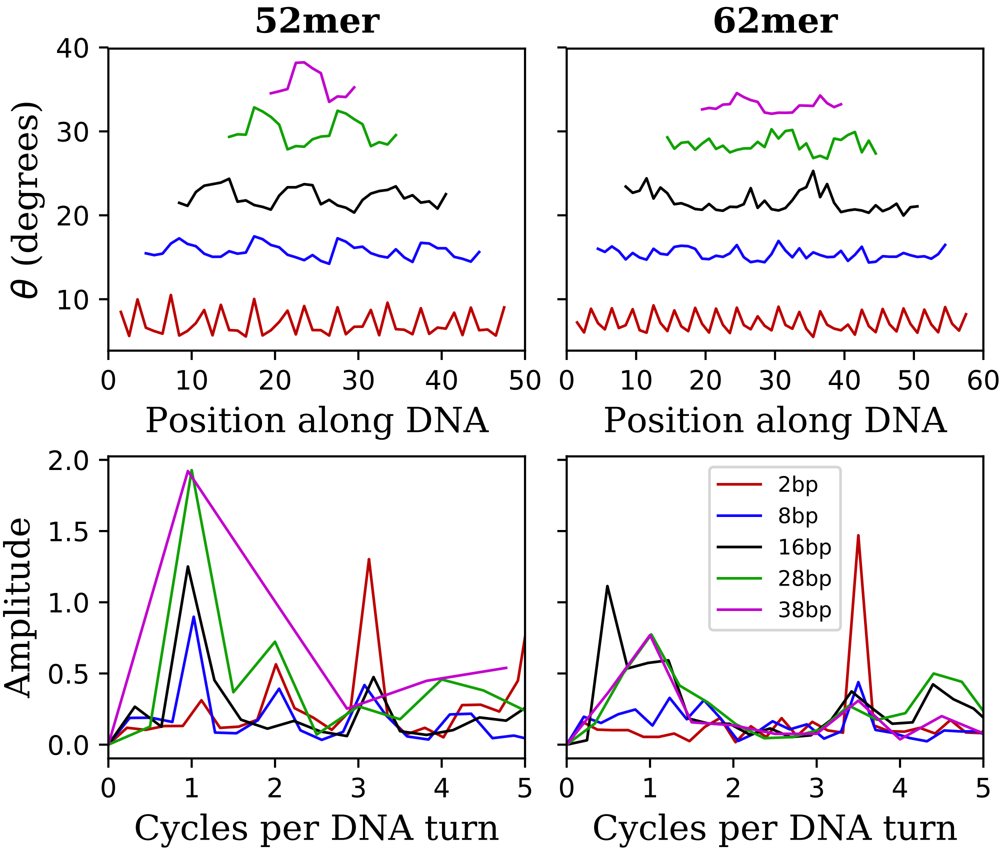


Figure 4.7: Top: Bending angle profiles for the 52mer (left) and 62mer (right) along the sequence using 5 distinct sub-fragment lengths. Bottom: The respective frequencies calculated from fast Fourier transforms of the bending angles. Each colour/line represents a sub-fragment length.

oberved in the 52mer (the more bendable oligomer) but not in the 62mer. This periodicity is in phase with the helical shape of the DNA, where bending angles complete one cycle per turn (when $l = 7, 15, 27$bp). Looking at the frequencies of the bending angles at the bp-step level ($l = 1$), approximately 3.5 cycles per turn are completed in case of the more rigid structure (62mer), whereas in the case of the more flexible structure (52mer) 3 cycles are completed per turn. This might suggest that an integer number of cycles per turn is needed at the bp-step level, so at higher lengths local bends can couple and form the global curvature.

Our results highlight the importance of periodicity for understanding special cases [95] with prominent bendability properties like A-tracts [94] or sequences positioning nucleosomes [6, 139].

4.2.6 DNA-protein complexes and DNA sequence mismatches

Another advantage that SerraNA offers is that it is not limited to work with linear unperturbed simulations of NA. The program can still process simulations where the molecule was perturbed either by sequence mismatches, protein binding or even strong supercoiling. However, there is the possibility that in these special cases, the four structural variables used in the inverse-covariance analysis might not comply with the harmonic approximations (see figure 4.2). Even so, the program can still provide some insight of how the perturbations affect the structure and flexibility of the DNA.

In the case of protein-DNA complexes, we study the case of a nucleosome (PDB 1kx5) and the transcription factor GCN4 (PDB 2dgc). Figure 4.8 shows the elastic constants, where both complexes present higher stiffness compared to free DNA. Independently if these proteins severely curved the DNA, both proteins restraint its thermal fluctuations consequently reducing its flexibility in terms of the five elastic constants. On the contrary, it seems that introducing A:A or G:G sequence mismatches at the middle of the molecule is enough to increase the overall flexibility (see figure 4.8). These results suggest that protein-DNA binding could function by restraining the DNA into a particular conformation and that sequence mismatches could be detected by the cellular machinery due to its enhanced flexibility. However, these results are preliminary and the analysis of more systems is necessary for a proper conclusion.

4.2.7 Tetranucleotide elastic constants from the ABC simulation database

As mentioned before, SerraNA gives the opportunity to study how elastic properties depend on DNA sequence. To demonstrate this feature, we analyse all 136 tetranucleotide DNA sequences extracted from the set of 39 oligomers of the ABC simulation database [119] (see figure 4.9). For the following results, each tetranucleotide sequence has the form of XXXX and we denote purine and pyrimidine bases as R and Y respectively. In general we observe a strong variability in the flexibility that depends on the sequence with differences over 200% in all elastic parameters. More specifically, tetramers with TA and CA base-steps are considerably more flexible than other sequences, while tetramers that contain AA and AT base-steps tend to be the most rigid sequences. Hence AT regions when phased properly, can generate rigid/soft mechanical

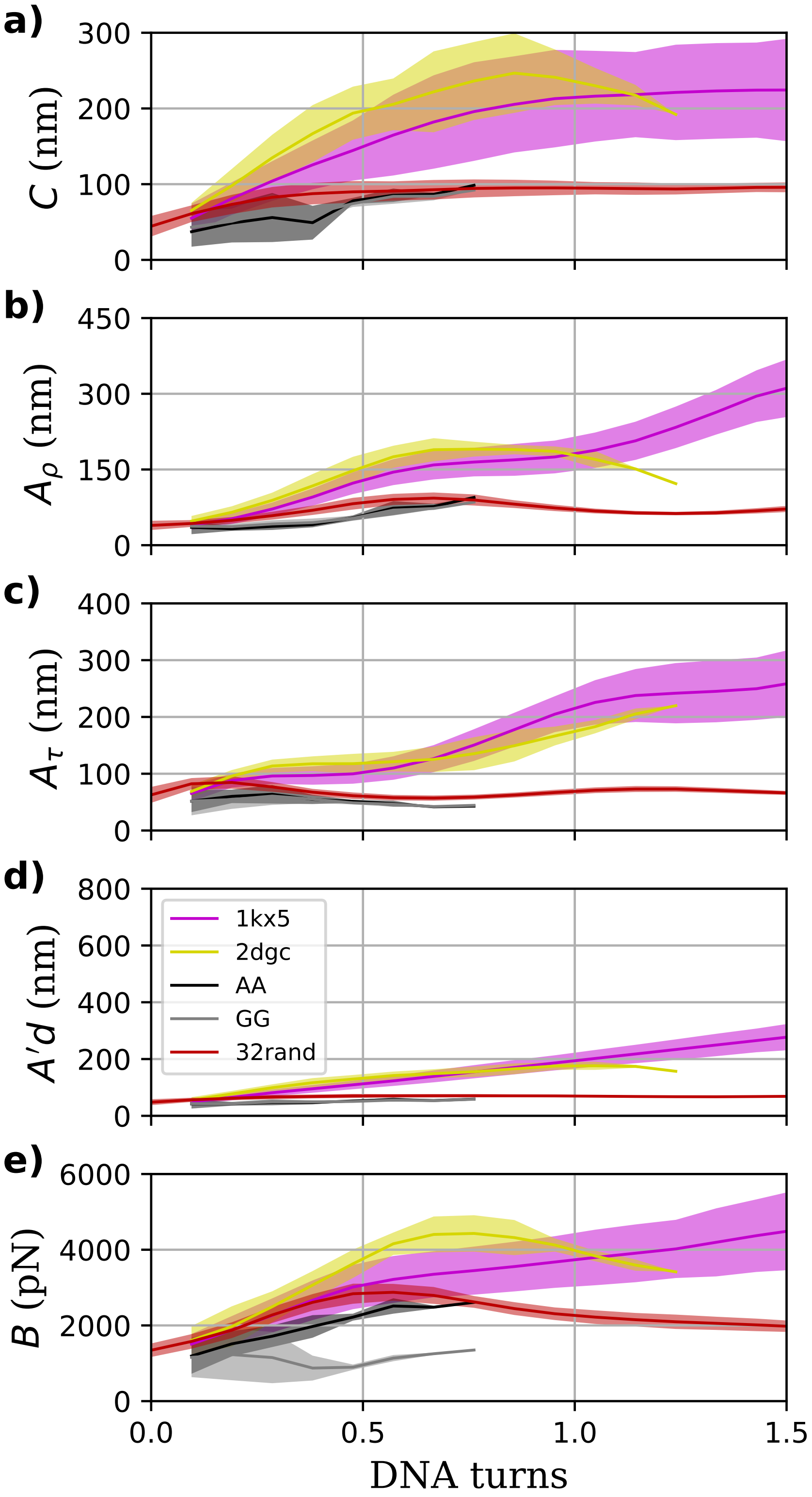


Figure 4.8: Elastic constants $(C, A_\rho, A_\tau, A'_d, B)$ as a function of length and obtained through the inverse-covariance method from the second set of DNA simulations as well as the linear DNA 32rand showed as a reference. Averages are calculated over all sub-fragments with same length which are showed as solid lines with shaded areas representing their standard deviations.

properties.

The static persistence length A_s is the parameter with highest variability (see figure 4.9), where can varies up to three orders of magnitude. Sequences like TGGG, TGCA and CATG are extremely soft having A_s less than 25 nm. However, most sequences are considerably flexible where only 13 sequences have values of A_s higher than 200 nm (see tables C.2-C.11). Most of the rigid sequences include AA or AT central base-steps, where having puring-pyrimidine (R...Y) or purine-purine (R...R) flanking bases further reduces the flexibility. This is reflected in the two most rigid sequences, AAAA (RRRR) and AATT (RRYY) with 1267 and 970 nm, respectively (see table C.2 and C.7). These measurements agree with previous studies, where it have been observed that A-tracts are incredibly stiff, preventing nucleosome formation [127], but can aid in looping and gene regulation when they are placed in phase with the helical periodicity [53, 127, 155]. One last thing to note is that the values at that particular length are specially soft, compared with experimental data, due to most bends are towards the major groove [95], which is the reflected by a decrease of tangent-tangent correlations (see figure 4.5b).

Focusing in the persistence length A (figure 4.9), we observe that the most flexible sequences have the form YRYR, in agreement with previous crystallographic and modelling studies [74, 119, 121]. This confirms that YR base-steps act as hinges and adding them together tends to soften the molecule. In contrast, RRRY and RRYY sequences are the most rigid.

Moving to the two estimations of the dynamic persistence length (A_d and A'_d), we see that both parameters yield similar patterns on the heat-maps of figure 4.9, where the major difference lies in the ranges being A'_d about 10 nm stiffer than A_d. Similar to A, RRRY and RRYY tetramers are particularly rigid, while YRYR are the most flexible sequences. These results indicate that even if values in A_s are much higher than A_d, the dynamic component is still an important contribution to the overall flexibility in terms of the persistence length A.

Twist elastic constant C (see figure 4.9) ranges from approximately 40 to 95 nm. These values are within both small and long length-scale observations, which agrees with the overall tendency observed in figure 4.4 as 4 bp is an intermediate length where the transition from local to bulk occurs. An interesting result is that tetramers with central YR base-steps (in the form XYRX), are the most rigid, as opposite to the observed at the length-scale of 2 bp, where these are the most flexible [74, 113, 121] (also see figure C.8). A last thing to note, is that central CG & AG base-steps known to have bimodal behaviour at the dinucleotide level [25, 26] do not show any particular feature. All these facts show the complex correlation between dinucleotide steps, where increasing the length together with the effect of flanking bases can suppress bimodal features and significantly change the flexibility. All these observations demonstrate the importance of the interplay of sequence-dependent and length-dependent properties of DNA, and once more, highlights the importance of interactions between dinucleotide steps [4, 5].

Finally, averaged values of the stretch modulus (2054 ± 354 pN see table C.12) fall within the length-dependency of B at the length of 4 bp (see figures 4.4 and 4.9), where many tetramer sequences present relative stiffer values compared with experiments at longer length-scales. As previously stated, the high stiffness of B at lengths less than one DNA turn is due to strong stacking interactions. This fact is reflected by the similarity in patterns between B and the stretch modulus associated with L_{CL} and Z_0, where Z_0 is the added rise and is built by stacking interactions (see figures 4.4 and C.9). The added slide component (Y_0) also contributes to the flexibility besides the added rise. We also observe strong sequence effects: some of the most rigid sequences such as CCGG, TTGC and CGAC are more than twice as stiff ($B > 2700$ pN) than the most flexible sequences like AAGG, AGGA and AGGG ($B < 1400$ pN). In general, RRYY and YYRR sequences present the highest values of B, while RRRR tetramers are the most flexible on average. Again AAAA tetramer is an exception, as it is particular stiff ($B = 2441 \pm 88$ pN, see table C.2), which is in good agreement with experimental results ($B \approx 2400$ pN) [94].

The analysis presented in this subsection, shows that the flexibility of DNA is influenced by both sequence and the length. This flexibility extends beyond the dinucleotide and the 4 bp levels, and is essential in determining the global flexibility of a given DNA fragment. We also observed that tetramer sequences with YRYR, and RRYR tend to be more flexible in contrast to RRYY, YYRR, RRRY and YRRY sequences which tend to be more rigid. However, this highly depends on the type of elastic constant that is being tested.

4.3 Conclusion

In this chapter we presented SerraNA, which is an open program that calculates structural and elastic properties of MD simulations of NA. Then, it translates these properties to evaluate the bulk flexibility of the molecule in terms of the elastic constants of persistence length, stretch modulus and twist modulus. We have explained the features and tools that SerraNA offers as well as the parameters and analysis it can output.

To demonstrate the usefulness of SerraNA, elastic profiles of a set of linear DNA fragments with distinct lengths have been calculated. In general, the elastic profiles shown in this chapter are in agreement with previous observations [114], where the crossover between local and global flexibility occurs within one DNA turn. We also find good agreement between our estimations of stretch and twist modulus (97 ± 3 nm and 1778 ± 88 pN) with experimental results (100 nm and 1500 pN), where in case of the persistence length our estimations are slightly more rigid (57 ± 3 nm) than accepted experimental values which are around 50 nm.

We have shown that SerraNA is not limited to the cases of linear and free DNA, but it can also analyse protein-DNA complexes as well as structures that do not follow the WC base-pairing rules like mismatches. We demonstrated that even if some cases do not comply with the harmonic approximation such as the GG mismatch trajectory where some distributions were skewed, SerraNA was still able to provide valuable insight regarding their structural and elastic properties. Overall, we discovered that the flexibility of DNA is apparently more rigid when proteins are bound to it. This apparent stiffness is caused by interactions with proteins, where they restraint the DNA

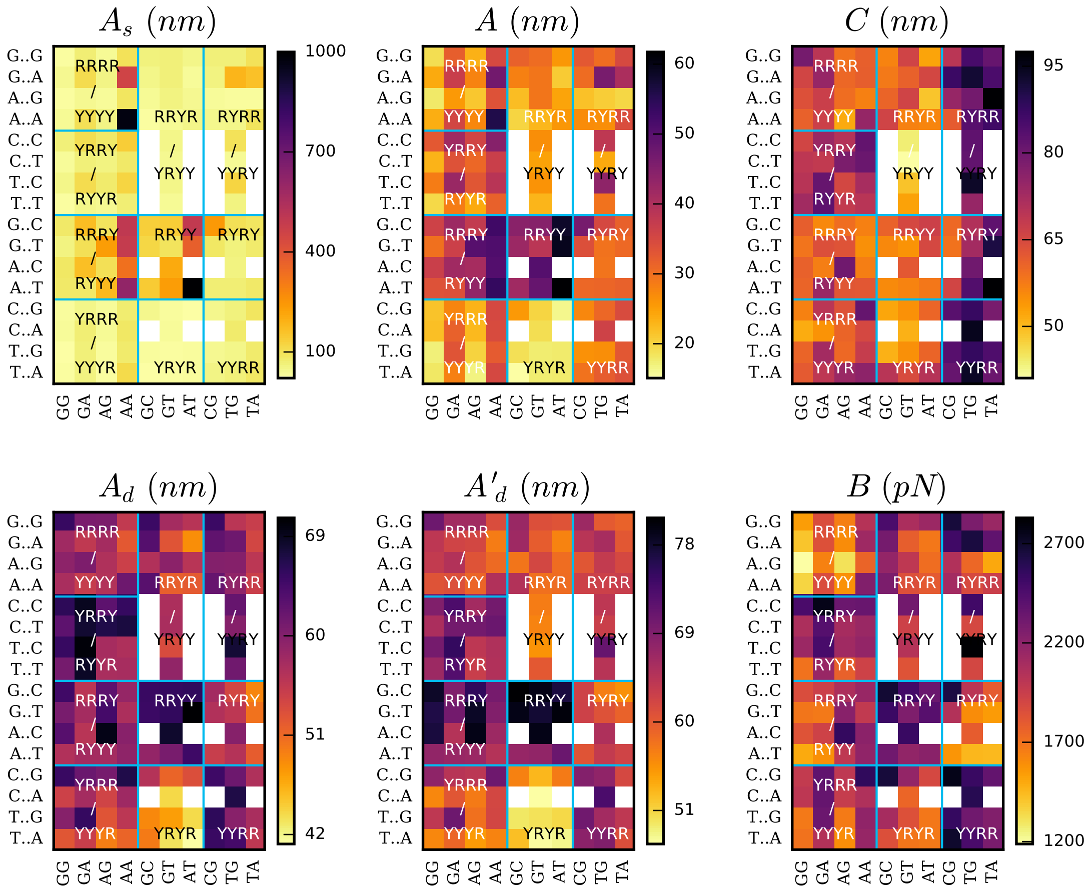


Figure 4.9: Sequence-dependent elastic constants at the length of 4 bp obtained from the set of 136 tetra-nucleotide sequences from the ABC simulation database. The persistence lengths and its contributions (A, A_s and A_d) are calculated from the directional decay at the length of l = 3bp using equations 2.57-2.59. Twist (C), stretch modulus (B) and the second estimation of the dynamic persistence length (A'_d) are calculated from the inverse-covariance method. The horizontal axes indicate the flanking bases (X..X) while the vertical axes indicate the middle steps (XX). Blue lines sort the sequences according to their purine (R) and pyrimidine (Y) type. Duplicated sequences are colored as white squares. Sequence AATT is off palette in case of A_s with a value of 1267 (see table C.7).

into a particular conformation decreasing its overall flexibility. On the other hand, we found that sequence mismatches can increase the flexibility of DNA as they act as flexible hinges. These two features could have important biological implications in DNA recognition, where proteins would bind sequences that allow the DNA to adopt particular conformations [163], and DNA repair, where cellular machinery may be able to detect sequence mismatches due to its enhanced flexibility.

Finally, we explored the sequence-dependency of elastic properties using all the 136 tetra-nucleotide sequences from the ABC simulation database [119]. We observe that the flexibility of DNA is strongly sequence-dependent, where some sequences are twice as rigid than others in all elastic parameters. We found that in general RRYY and RRRY sequences are the most rigid while YRYR are the most flexible. More specifically, we find that AT and AA base-steps are less bendable due to their structure being intrinsically straight, while TA and CA base-steps are more flexible. We also find that RRYY and RRRY tetramers containing AT and AA steps have a persistence length of 38 nm which is higher than YRYR tetramers containing TA and CA steps. This indicates the high importance of AT-rich motifs in defining extreme mechanical properties, which can then build up global flexibility on longer fragments when positioned in phase with the helical shape. Consequently this demonstrates the potential that SerraNA has in providing analysis that could uncover different mechanical properties between AT and GC-rich sequences as well as their biological function [159]. Furthermore, our analysis indicates that both sequence-dependant and length-dependant mechanical properties are of crucial importance in biological processes such as DNA-protein interactions, as the mechanical responses of DNA depend on both sequence and length, and interacting proteins might benefit from these properties.

Overall, this chapter convincingly demonstrates the versatility and applicability of SerraNA in analysing the DNA mechanical properties across various length scales. Its capability to derive global elastic parameters, suitable for comparison with experimental data, further solidifies its value in validating simulations and facilitating multi-approach investigations. We firmly believe that SerraNA will become an invaluable tool in the biophysics field, benefiting the scientific community in numerous ways.

It is worth noting that this chapter does not delve into the analysis of DNA elastic couplings, an important aspect of DNA flexibility. Even though SerraNA is programmed to calculate the elastic couplings as a function of length, the methodology for estimating these couplings remains lacking not only in the SerraNA program but also in the existing literature. In the following chapter, we will explore deeper into these crucial aspects of DNA flexibility and aim to shed more light on this subject.

Chapter 5

Investigating the DNA elastic couplings

Synopsis

For decades, the scientific community has been aware that deformations of the DNA are coupled due to the chirality of the double helix. Theoretical investigation carried out by Marko and Siggia [98] yielded the existence of the twist-bend coupling (G), while deviations from the TWLC in force-extension measurements were the first indicators of the existence of the twist-stretch coupling (D) [49]. Since then, scientists have tried to measure D and G couplings through single molecule experiments such as magnetic tweezers and optical tweezers. Most of the attention has been paid to the D coupling, where experiments have yielded a negative twist-stretch coupling raging from -6.5 to -11.9 nm [48], [86], [49], [134]. At that time, this negative value was denoted as a non-intuitive behaviour as it means that the DNA molecule overwinds when stretched. This is opposed to double stranded RNA which underwinds when stretched, having a positive twist-stretch coupling around 6.2 nm [86]. On the other hand, the twist-bend coupling has been measured through force-torque experiments with a value around 25-30 nm [106, 111, 141]. Lastly, it is known that DNA also has a bend-stretch coupling (H), which only a handful of computational studies attempted to estimate [74, 116]. However, there is no experimental setup that aims to estimate bend-stretch nor a consensus value in the literature.

SerraNA has the capability to calculate the DNA elastic couplings. Here, we aim to mathematically describe the behaviour of the DNA coupling terms as a function of length, and to uncover the movements that originate them through the use of PCA.

5.1 Elastic cross-terms

In the previous chapter, we only showed elastic profiles of the diagonal components of elastic matrix F (see equation 2.52). Here, we analyse the elastic profiles of the remaining 6 components and only focusing on the 52 and 62 mers, since they are long enough to observe features beyond two helical turns (see figure 5.1).

5.1.1 Investigating the periodic nature of the cross-terms

Upon initial examination, we notice that the elastic couplings exhibit two features. The first feature is characterised by three significant couplings, which correspond to twist-stretch (D), twist-roll (G) and roll-stretch (H), while the couplings related to tilt are close to zero on average (see figure 5.1). The second feature we observe is that all elastic couplings present periodic behaviour, which we investigate by using Fourier transforms.

Regarding the three non-zero couplings, the Fourier transforms indicate that D presents oscillations with a period of one DNA turn, while G and H have periods of two DNA turns. Similar to the stretch profile (see figure 4.4e), the twist-stretch coupling D does not follow a monotonic behaviour: it first presents a periodic behaviour for lengths shorter than 2 DNA turns and then it reaches a negative plateau around -5 nm, which qualitatively agrees with previous studies as they have measured a negative coupling ranging from -6.5 to -11.9 nm [48, 49, 86, 134]. The G and H couplings present ondulatory behaviour as well, where their profiles behave as damped waves that fade as the distance increases; however, these two couplings are out of phase by approximately 3 bp. Focusing on G, its maximum value (global maximum) is around 20 nm, which highly agrees with values reported in the literature of approximately 20 nm [106]. In spite of almost no information in the literature, we find that H is a relevant coupling as it reaches values around 50 nm within one DNA turn, which is larger in magnitude compared to D and G. Lastly, a very important thing to keep in mind for the rest of the analysis is that these couplings are effective elastic constants, which means that the effect of other variables have been removed for each coupling.

A very interesting behaviour to note is the interplay between D, G and H. At the length of half a turn, the twist-stretch coupling is near 0, which means that the twisting of the molecule would not cause variations in the end-to-end distance. However, at this length, the twist-roll coupling is at its global maximum, which means that bending would unwound the DNA. This can be related to DNA deformations caused by the binding of the family of transcription factors bzip, in which two arms bind to the major grooves of a region of half turn of distance [14]. There is also evidence that the GCN4 transcription factor (see figure 1.3a), which belongs to this family, smoothly bends and slightly underwinds the DNA structure [37, 72], which agrees with the positive twist-roll coupling that we observe at half DNA turn. At lengths between 1 and 1.5 DNA turns, D reaches minimum, G changes its sign and reaches a minimum as well, and H approaches to 0. This qualitatively agrees with the binding of the 434 repressor, which binds a region between these two lengths (see figure 1.3b) and bends and overwinds DNA [1]. IHF is known to bend the DNA through interactions of its 'arms' with two contact points separated by 9bp1.3c). Our profiles show that, at this length, twist

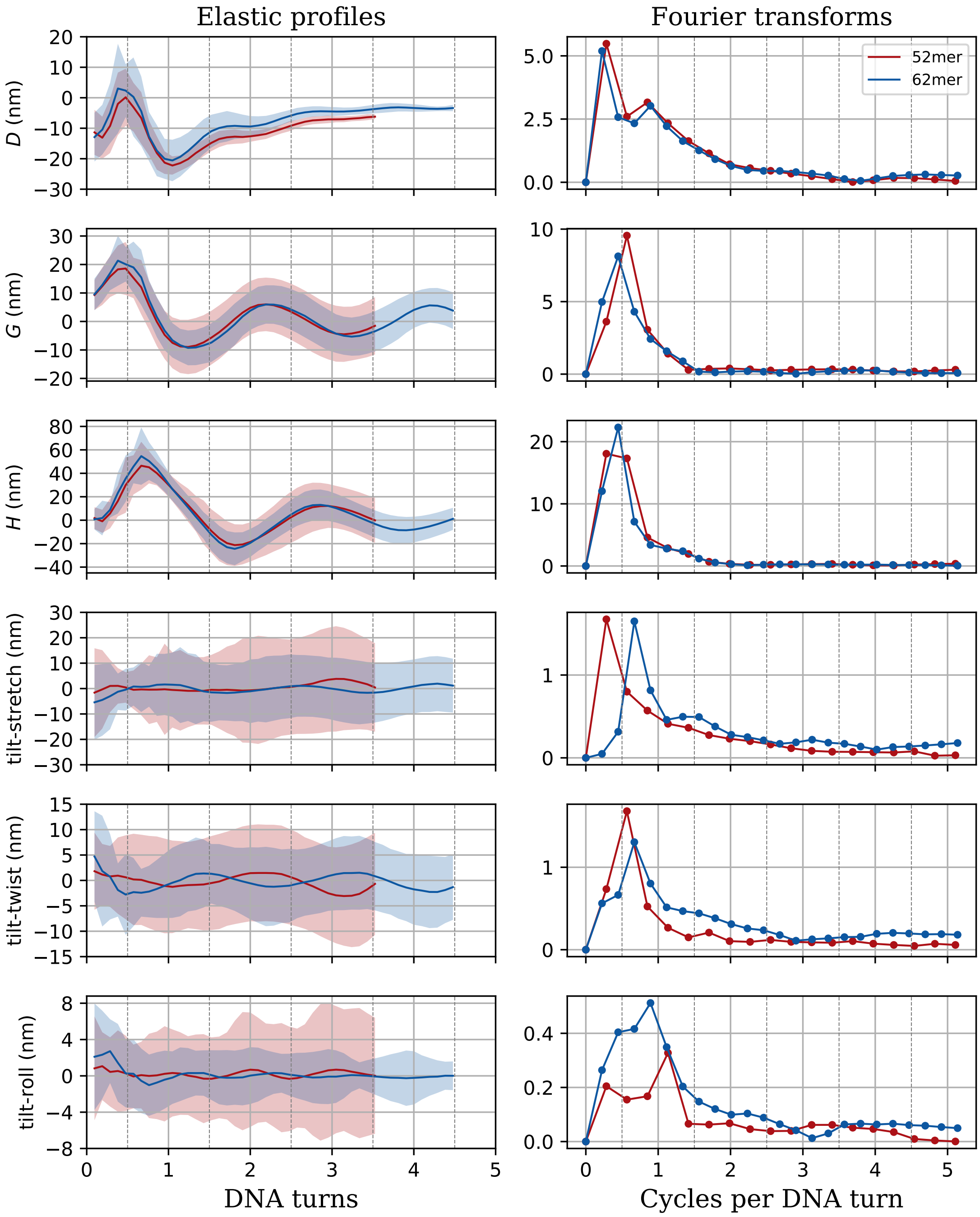


Figure 5.1: Elastic profiles of the six elastic couplings with their respective Fourier transforms. Red colour corresponds to the 52mer and blue to the 62mer. Shadowed areas represent standard deviations.

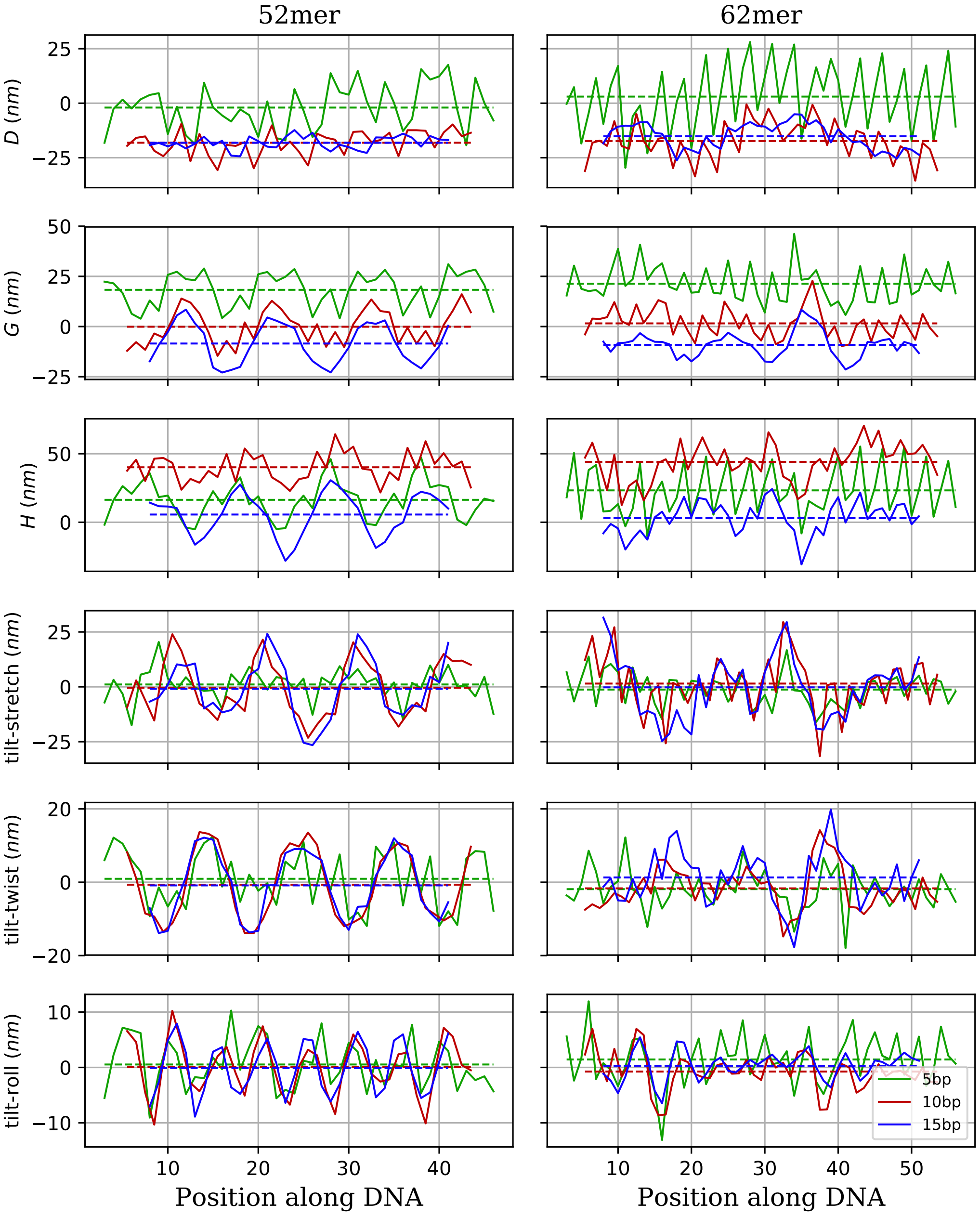


Figure 5.2: Elastic couplings along the sequence for the 52mer and 62mer. The couplings are plotted at the key lengths of 5, 10 and 15 bp, which corresponds to approximately .5, 1 and 1.5 DNA turns. Dashed represent the average at each length.

and roll are not correlated, which agrees with crystallographic data of IHF where no overtwisting of the double helix is observed [129]. However, it is worth pointing out that the sharp bend that IHF induces causes a kink in one of the DNA contact points which may cause the flexibility to deviate from the harmonic approximation. These are a few examples in which proteins might benefit from key lengths in order to induce particular structural conformations on the DNA.

An important inquiry that arises when examining the couplings associated with roll (G and H) is the reason behind their two-DNA turn periodicity. The answer to this question lies in the computation of the mid-step triad (MST) as a function of length, from which the roll and tilt angles are derived (see the methods 2.2.2). The MST is constructed between the base-pairs i and j, and while these two base-pairs will be aligned every DNA turn, the axes of the MST will be synchronised every two DNA turns, causing the roll and tilt angles to have identical periods of two DNA turns as well. To verify this hypothesis, we calculated the structural profiles for the roll (ρ) and tilt (τ) angles (see figure D.1). The Fourier transforms of these structural profiles support our hypothesis, revealing that both angles have a period of approximately 2 DNA turns (with a frequency of half a cycle). This periodic behaviour is reflected in the elastic couplings of twist-roll (G) and roll-stretch (H).

Focusing on the tilt couplings, we observe that on average they are close to zero. However, Fourier transforms indicate that they have periodic behaviour with a period around 1 DNA turn. Nonetheless the relatively high standard deviations are intriguing since they suggest that even if the couplings are zero on average, they might have considerable contributions in some cases. Hence we plotted the couplings at the key lengths of 10, 15 and 20 bp to further investigate this matter (see figure 5.2). Similar to the bending angles of figure 4.7, the couplings have periodic behaviour along the sequence, where in case of the 52mer (more bendable structure), they are characterised by a more regular pattern than the 62mer (less bendable structure). This behaviour could be originated by the more bendable nature of the 52mer, since its sequence is a better candidate for nucleosome formation than the 62mer [160]. Therefore it is remarkable to observe such a clear periodicity along the sequence, where the couplings might adjust their magnitude and sign to facilitate the wrapping around histones. These characteristic patterns also resemble the twist waves found in circular DNA [110, 143], where due to the bending coupling, the elasticity propagates through waves along the length of the molecule. In the case of couplings related to tilt, these oscillations are around 0 at any length, which is why the elastic profiles show that tilt couplings are zero on average (see figure 5.1). However, in case of couplings related to roll, they also oscillate along the DNA but around a non-zero value that oscillates as a function of length.

These results highlight the importance of the DNA elastic couplings, which exhibit a length-dependent behaviour (see figure 5.1). We observed significant variations in the elastic couplings across different lengths even as small as one DNA turn. These findings suggest that proteins acting on the DNA within these ranges may benefit from these couplings as they could exploit their flexibility and reduce the energy required to induce deformations on the DNA. Additionally, our results indicate that tilt-couplings can be neglected at the global level, as previously stated in the MS model [98]. However, as shown in figure 5.2, certain DNA sequences may not follow this general rule. These

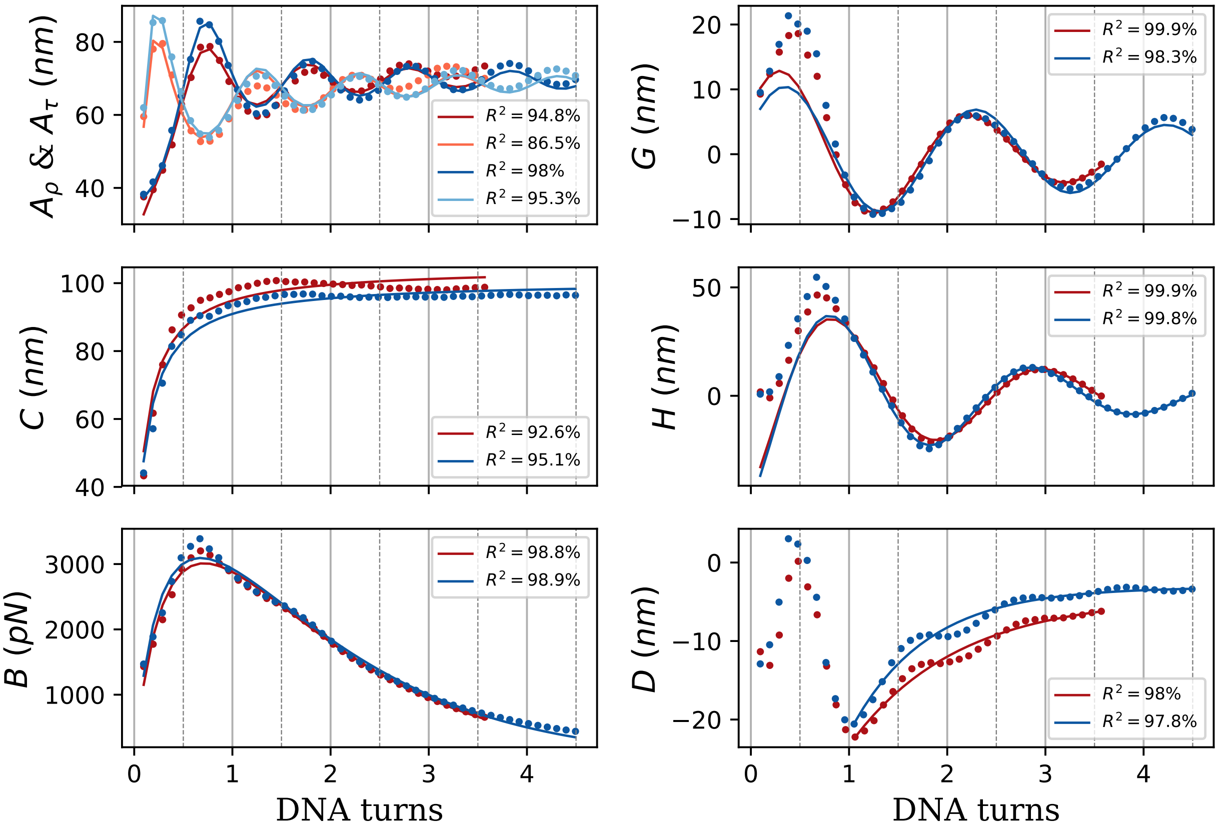


Figure 5.3: Fittings of the 7 non-zero elements of the elastic matrix as a function of length. Dots represent averaged data measured with SerraNA while shaded areas are their standard deviations. Lines represent the fitted curves where legends indicate their accuracy measured by the R^2 parameter. Red colours correspond to data from the 52mer and blue colours to the 62mer. Lighter colours in the A_ρ & A_τ panel represent tilt and darker colours roll. Fitted curves for lengths less than 1 DNA turn (≈10.5 bp) are not shown for D as they go out of the ranges of the y-axis.

findings contribute to a better understanding in the relationship between the DNA elastic couplings, how their behaviour changes as a function of length, and how proteins might benefit from these relationships.

5.1.2 Modelling the DNA flexibility as a function of length

Now we move on to mathematically describing the 7 non-zero elements of the elastic matrix F as a function of length (see equation 2.52). Here, we will empirically propose equations that describe how the elastic constants change as a function of length, then we will evaluate capacity of our model to describe the elastic behaviour by performing curve fittings with the corresponding elastic profiles. For the curve fittings, we use the curve_fit function from the scipy python library [161] that uses the Levenberg-Marquardt algorithm [99] for solving non-linear least squares problems. The performance of the curve fittings be measured through the coefficient of determination R^2 [32], which measures the proportion of variance that our model predicts and its percentage form is defined as:

Parameter	52mer	62mer	Average
a_ρ (turns)	0.7 ± 0.1	0.0 ± 0.1	0.4 ± 0.3
b_ρ	15.7 ± 0.1	16.1 ± 0.1	15.9 ± 0.2
c_ρ (turns)	1.8 ± 0.1	2.3 ± 0.1	2.1 ± 0.2
f_ρ (turns^{-1})	0.98 ± 0.02	0.97 ± 0.01	0.98 ± 0.01
ϕ_ρ (degrees)	0 ± 6	0 ± 2	0 ± 0
a_τ (turns)	1.5 ± 0.2	1.2 ± 0.1	1.4 ± 0.1
b_τ	15.9 ± 0.1	16.0 ± 0.1	15.95 ± 0.07
c_τ (turns)	2.0 ± 0.1	2.1 ± 0.1	2.05 ± 0.07
f_τ (turns^{-1})	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.00
ϕ_τ (degrees)	180 ± 4	174 ± 2	177 ± 3
a_C (turns)	1.1 ± 0.1	1.2 ± 0.1	1.15 ± 0.07
b_C	10.7 ± 0.1	11.1 ± 0.1	10.9 ± 0.2
a_B (nm turns)	-0.013 ± 0.001	-0.017 ± 0.001	-0.015 ± 0.002
b_B (nm turns)	0.023 ± 0.001	0.025 ± 0.001	0.024 ± 0.001
c_B (turns^{-1})	1.00 ± 0.02	0.96 ± 0.02	0.99 ± 0.02
a_G (nm)	-0.15 ± 0.04	-0.20 ± 0.13	-0.18 ± 0.02
b_G (nm)	14.7 ± 0.3	11.4 ± 0.7	13 ± 2
c_G (turns)	2.59 ± 0.07	4.80 ± 0.57	4 ± 2
f_G (turns^{-1})	0.516 ± 0.002	0.515 ± 0.004	0.516 ± 0.001
ϕ_G (degrees)	-331 ± 2	-338 ± 4	-334 ± 5
a_H (nm)	0.2 ± 0.1	-0.4 ± 0.1	-0.1 ± 0.3
b_H (nm)	54 ± 2	56 ± 1	55 ± 1
c_H (turns)	2.02 ± 0.06	2.02 ± 0.05	2.020 ± 0.002
f_H (turns^{-1})	0.467 ± 0.002	0.487 ± 0.002	0.48 ± 0.01
ϕ_H (degrees)	-57 ± 2	-60 ± 1	-58 ± 1
a_D (nm)	-4.4 ± 0.7	-3.1 ± 0.3	-3.8 ± 0.6
b_D (nm)	-47 ± 4	-67 ± 6	-57 ± 14
c_D (turns)	1.1 ± 0.1	0.8 ± 0.1	0.9 ± 0.2

Table 5.1: Parameters obtained from the curve fittings of figure 5.3, where the symbol ± indicates standard deviations. Units of the fitted parameters are shown in parenthesis, where b_ρ, b_τ and b_C are dimensionless. Unit "turns" correspond to DNA turns which corresponds to approximately 10.5 bp.

$$R^2 = \left(1 - \frac{\sum_i^n (y_i - f_i)^2}{\sum_i^n (y_i - \bar{y})^2}\right) 100\% \tag{5.1}$$

where y_i is the measured data, f the fitted curve and n the number of data points. Our aim will be to obtain expressions that describe at least 80% of the data.

Beginning with the bending elastic constants, we observe that both roll and tilt profiles correspond to plateaus that oscillate as a function of length (see figure 4.4b-c). Therefore, the following equations describe this observed behaviour:

$$A_\rho(l) = \frac{r_d^2 bl}{a_\rho + b_\rho l + c_\rho sin(2\pi f_\rho l + \phi_\rho)} \tag{5.2}$$

$$A_\tau(l) = \frac{r_d^2 bl}{a_\tau + b_\tau l + c_\tau sin(2\pi f_\tau l + \phi_\tau)} \tag{5.3}$$

where $r_d = 180°/\pi$ and b is the average bp rise of B-DNA. The fittings of these two equations have high accuracy for both 52mer and 62mer, where the R^2 parameters are higher than 90% except for the tilt elastic constant of the 52mer which has an accuracy of 86.5% (see figure 5.3). The reason why it presents lower precision is because the A_τ plateau is not precisely constant in the 52mer, and its magnitude slightly changes as the length increases, where equation 5.3 is not capable of describing it. However, it is still a good fit since it is higher than our 80% threshold. Table 5.1 shows the fitted parameters, which are similar between both structures for tilt and roll. Frequency components (f_ρ & f_τ) indicate that both elastic constants have a period of one DNA turn, but as expected have a phase difference of half turn as indicated by their phase components ((ϕ_ρ & ϕ_τ)). Lastly, if we calculate the limits of equations 5.2 & 5.3 they would correspond to sinusoidal waves that oscillate around the plateau $\frac{r_d^2 b}{b}$ (see appendix D.1 & D.2). Plugging the fitted parameters of table 5.1 we can obtain the elastic constants of tilt and roll. Finally, using equation 2.61 we can obtain the second estimation of the dynamic persistence length being 70.09 nm which is slightly stiffer than values calculated by SerraNA (see table 4.1) and is between experimental values reported in the literature [10, 162] (A_d between 50 and 80 nm).

Similar to tilt and roll, the twist elastic constant is characterised by a plateau around one DNA turn, however, it does not present periodic behaviour (see figure 4.4a). Therefore, its elastic profile is approximated by the following equation:

$$C(l) = \frac{r_d^2 bl}{a_\Omega + b_\Omega l} \tag{5.4}$$

For both structures, this equation describes the twist elastic profiles with an accuracy higher than 90% (see figure 5.3), where their two fitted parameters are in good agreement (see table 5.1). Similar to the bending elastic constants, we can estimate the twist elastic constant by calculating the limit (see appendix D.3) which yields $\frac{r_d^2 b}{b_C}$. These plateaus correspond to 102.42nm on average, which again are slightly stiffer than the ones calculated by SerraNA (see table 4.1) and are in good agreement with experimental torque measurements which yield a twist modulus between 90 and 120 nm [13, 85, 107].

Lastly, as previously stated the profile of the stretch modulus B follows a complex behaviour (see figure 4.4e). We observe that the reciprocal of an exponential function describes well the non-monotonic behaviour of the stretch modulus:

$$B(l) = \frac{k_B T b l}{a_L + b_L e^{c_L l}} \tag{5.5}$$

Curve fittings of this equation yield R^2 parameters of almost 99%, which are extremely good fits as the equation is capable of describing the complex behaviour of B with only three parameters. As in the previous cases, the fitted parameters are in agreement between the 52mer and 62mer (see table 5.1). But in contrast to the A_ρ, A_τ and C, the elastic constant B tends to 0 at long lengths, hence we cannot estimate its elastic constant by calculating the limit approach. However, the excellent fits indicate that the proposed equation does very well in describing the stretch modulus as a function of length as predicted by the LDEM [114] (see figure 2.8a).

Moving to the elastic couplings, we observe that the roll couplings G and H can be described as damped oscillations, while the twist-stretch coupling as an exponential decay for lengths longer than 1 turn:

$$G(l) = a_G + b_G e^{-l/c_G} sin(2\pi f_G l + \phi_G) \tag{5.6}$$

$$H(l) = a_H + b_H e^{-l/c_H} sin(2\pi f_H l + \phi_H) \tag{5.7}$$

$$D(l) = a_D + b_D e^{-l/c_D} \tag{5.8}$$

In general, from the curve fittings of figure 5.3 we observe that the three couplings greatly deviate at lengths less than one DNA turn. This could be due to the transition between local and bulk flexibility which as described in the LDEM occurs within one DNA turn [114]. This transition drastically changes the coupling's behaviour, and is even more evident in D as it completely transitions from a periodic behaviour for a more constant-like behaviour that resembles the profile of the twist elastic constant. Therefore, we filtered lengths less than one DNA turn when performing the curve fittings. For G and H, we show how the elastic curves would behave if they were to follow the same pattern at all lengths, where both couplings would be lower in magnitude at lengths shorter than one turn and G would get out of phase. We do not show this for D as it would greatly deviate from the ranges of the y-axis.

Further, the quality of the fittings is good as R^2 is greater than 95% in the three cases. Regarding D, its equation describes the bulk flexibility well, where we observe that the decaying rate c_D in both structures is similar. However, the parameter b_D is higher in magnitude in the 62mer. From equation 5.8, it can be easily seen that at long length-scales, the twist-stretch coupling tend to a constant value a_D, which is of -3.78nm on average (see table 5.1). This result qualitatively agrees with force-extension measurements that measured a negative twist-stretch coupling [48,49,86,134]. Regarding G and H, notice that from equations 5.6 and 5.7 they would tend to a_G and a_H at long length-scales; which are relatively low values (see table 5.1). However, the amplitude of G (b_G) is of 13 nm on average, which is differs from the twist-bend couplings calculated from force-torque measurements (20-30 nm [106, 111, 141]). Nonetheless, it is worth pointing out that both G and H fittings highly deviate from the stronger amplitudes observed at small lengths. And as expected, the frequencies of G and H indicate that both couplings have a period of approximately 2 DNA turns, and their respective phases indicate that they are out of phase by approximately 2.5bp. These parameters indicate that these two couplings would be changing signs every DNA turn and that at the critical lengths of $\frac{(2n+1)}{4}$ (with $n = 1, 2, ...$), one coupling would be at a local maximum/minimum and the other would be approximately zero. On the other hand, the critical lengths of the twist-stretch coupling are at half and one DNA turn, where in the first one the twisting of DNA would not affect its end-to-end distance and on the other one the two variables would be at its maximum correlation (stretching overwinds the DNA). Putting these observations altogether, the set of critical lengths correspond to: $\left[\frac{2}{4}, \frac{3}{4}, \frac{4}{4}, \frac{5}{4}, \frac{7}{4}, ..., \frac{(2n+1)}{4}\right]$ DNA turns with $n = 1, 2, 3$,...

It is important to note that non-linear least-squares methods can be limited when fitting curves on small datasets with multiple parameters, as they are sensitive to initial guesses and the risk of over-fitting is high. However, for all our cases, we are dealing with overestimation, where the number of observations is larger than the number fitting parameters [67]. For example, for the periodic elastic couplings of the 52 mer we have 42 observations and 5 fitting parameters. We overcome these limitations by carefully selecting initial guesses, verifying the solutions are realistic, and ensuring convergence. Large errors can be a sign of over-fitting which did not occur in any of our cases. Although testing on different DNA simulations is needed to further validate our proposed functions, the good agreement between both of our simulations demonstrate that the proposed equations provide a good mathematical description of the DNA elastic couplings.

The analysis presented in this section provides a mathematical description of the elastic constants. This analysis allowed us to identify critical lengths in which couplings are most correlated or decoupled. These critical lengths agree with the DNA-protein complexes described in the last section, where the GCN4 transcription factor deforms the DNA at approximately a distance of half a turn [37, 72], the IHF protein grabs and bends the DNA by two bp separated by almost one helical turn [1] and the 434 repressor binds a region around 1.25 turns [129] (see figure 1.3). These equations might aid in the modelling of more coarse-grained models where couplings and interactions beyond the nearest-neighbour approximations can be considered. However, we consider that these are early results and further analysis and evidence are required for deriving a complete theory and for exploring sequence-dependent features.

5.2 Calculating the DNA essential modes

In this section, we aim to identify the essential modes that describe the flexibility of double stranded DNA. To achieve this, we utilize principal component analysis (PCA), a powerful technique that has been used previously to capture complex movements, including curved trajectories [138]. While vectors in Cartesian space are linear, a vector of dimension $N \times 3$ can represent more complex movements such as rotations. PCA has been applied to analyse and capture a variety of movements, such as those of professional dancers [12], as well as DNA movements from MD simulations [39, 116].

Here, we combine PCA with SerraNA to associate a principal mode that primarily explains each elastic constant of interest at specific lengths. Our approach focuses on the covariance matrix formed by the four structural parameters used in the calculation of the elastic matrix: tilt, roll, twist and the end-to-end distance. To associate essential modes with elastic constants, we perform PCA on a trajectory and then rebuild it using one of the modes. Subsequently, SerraNA analyses each rebuilt trajectory and calculates the covariance matrices V. Finally, an essential mode is associated with one elastic variable if it yields the highest variance/covariance (element of V).

While a large combination of multiple principal modes would be necessary to fully capture the complex flexibility of DNA, our method focuses on associating individual modes with each elastic constant. This approach enables us to investigate the primary movements that contribute to the flexibility of DNA at various length scales, up to 4 DNA turns. It is akin to dissecting the system, aiming to identify key movements that contribute to the variations in our four structural variables: twist, roll, tilt and end-to-end distance (stretch). By uncovering the origin of the elastic properties of DNA, this analysis sheds light on the fundamental mechanisms governing its flexibility.

5.2.1 Associating essential modes to elastic constants (first classification)

The aim of our classification process is to identify the principal modes that originate the flexibility of DNA. We implement PCA to calculate the first 20 principal modes that are responsible for most of the dynamics in the MD simulations as they represent more than 90% of the system variance (see methods section 2.6). We classify them according to their contributions to DNA flexibility in terms of the four diagonal elastic constants of roll (A_ρ), tilt (A_τ), twist C and stretch B, and the three elastic couplings of twist-stretch (D), twist-roll (G) and roll-stretch (H). Only the mode that causes the major contribution to a particular elastic constant is associated with it. However, we want to obtain four independent modes that are related to the four diagonal elastic constants to see how the couplings emerge. In reality, the principal modes can cause relevant correlations in multiple elastic constants; hence we allow the principal modes associated with one diagonal elastic variable to be associated with the elastic couplings as they arise from correlations between the diagonal components. For example, if one mode is associated with roll it cannot be associated with twist, but it can be associated with the twist-roll coupling. It is worth mentioning that with this process, only a handful of modes out of the 20 will be associated with the elastic constants.

Similar to the LDEM (see figure 1.5), our method selects all possible sub-fragments for all possible lengths l within the given molecule. We form sub-fragments of lengths $l = 2, 4, ...38$ bp, where in total we have N_l number of sub-fragments for each length l. For example, for the 52mer $N_l^{52} = (N + 1) - (l + 4)$, where N is the total number of bp and the number 4 represents the two bp discarded at each end. We then perform PCA using PCAsuite on each sub-fragment trajectory to calculate the first 20 essential modes that explain more than 90% of the system variance (ν) (see methods section 2.6 and figure 5.4a). PCAsuite [138] sorts the modes according to the amount of variance that they contribute to the system, where the first mode m = 1 is the one that yields the highest system variance followed by m = 2, 3, ... , 20. In total, we obtain

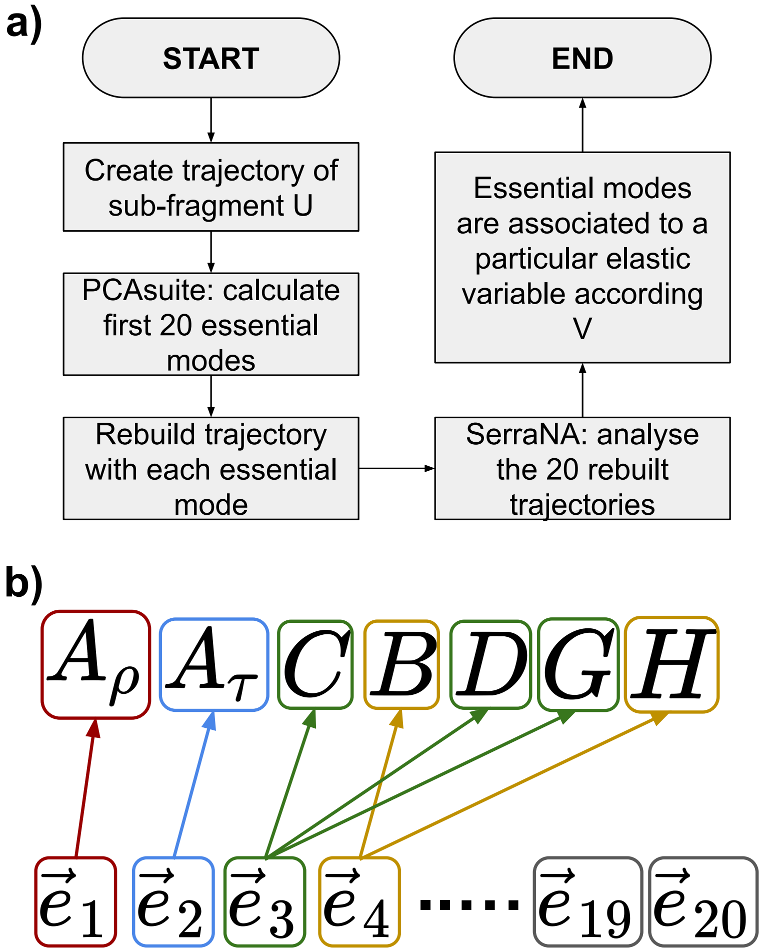

a)
START
Create trajectory of sub-fragment U
PCAsuite: calculate first 20 essential modes
Rebuild trajectory with each essential mode
SerraNA: analyse the 20 rebuilt trajectories
Essential modes are associated to a particular elastic variable according V
END
b)
A_ρ
A_τ
C
B
D
G
H
$\vec{e}_1$
$\vec{e}_2$
$\vec{e}_3$
$\vec{e}_4$
.....
$\vec{e}_{19}$
$\vec{e}_{20}$

Figure 5.4: a) General process of the first classification in which the first 20 principal modes of each sub-fragment U are associated with the elastic constants. First, each sub-fragment is stripped from its molecule (52mer or 62mer) to then be analysed with PCAsuite [138] for calculating its first 20 essential modes. Then, the trajectory of U is rebuilt using one mode at a time to be analysed with SerraNA to calculate the covariance matrix V. Lastly, one mode is assigned to each particular elastic constant if it yields the largest variance/covariance in that parameter. b) Representation in which the first 20 essential modes ($\vec{e}$) of a given sub-fragment are associated with the elastic constants. In this classification, only the first 4 modes are associated with the 7 elastic constants, while the remaining 16 modes are unassigned. Consequently, some modes were assigned to more than one elastic constant, where $\vec{e}_3$ was associated with C, D and G, while $\vec{e}_4$ was associated with B and H. Notice that in this case one mode is assigned to each diagonal elastic constant.

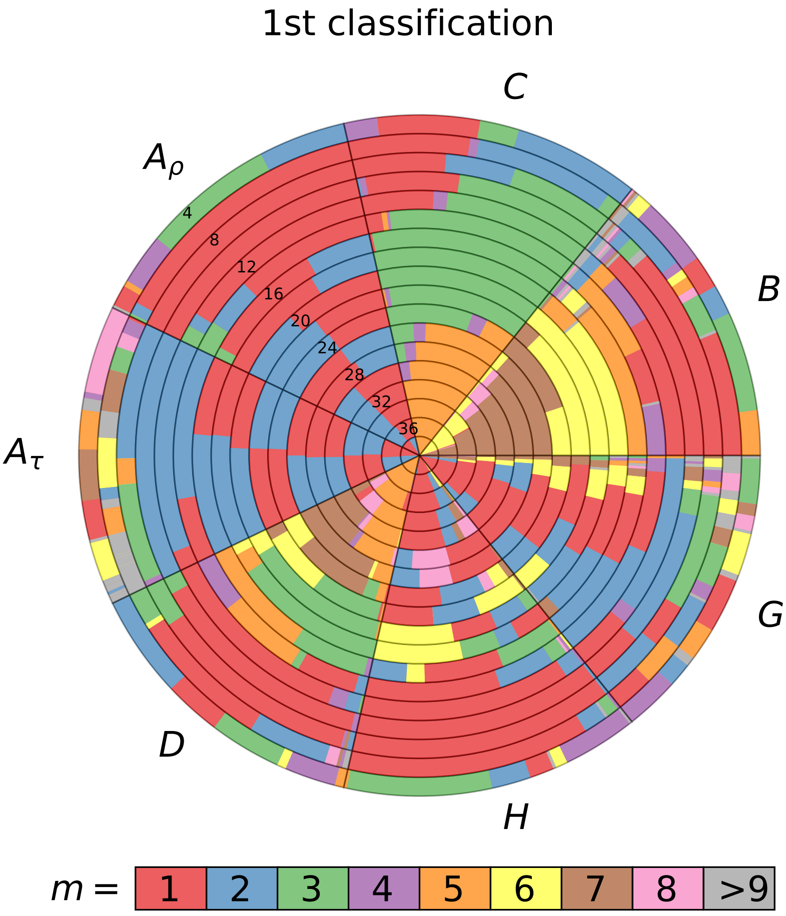


Figure 5.5: Donut chart summarising the mode association with the 7 elastic constants resulted from the first classification, and for the sub-fragments extracted from the 52mer and 62mer. Each radius represents a different length, being 4 bp the most external and 38 bp at the centre. Colour is associated with the different modes m, being grey for $m \geq 9$. Wedge width represents the percentage of modes with m that were associated to a particular elastic constant at a particular length.

a set of $20N_l$ (with $N_l = N_l^{52} + N_l^{62}$) essential modes per length. Each trajectory is then rebuilt according to one principal mode at a time using equation 2.114 (see figure 5.4a). Then, these trajectories are analysed by SerraNA to calculate the covariance matrices V. Finally, we associate one essential mode to one elastic constant if it yields the biggest magnitude of the corresponding variance or covariance (see figure 5.4b).

In figure 5.5, the donut chart shows a summary of the classification results for both 52mer and 62mer. However, the presence of significant noise (large variations in color) in the plot suggest that our current process may lack the desired accuracy. We noticed that some modes are associated with multiple diagonal elastic constants. For instance, at the length of 10 bp, some modes associated with roll are also associated with twist. This behaviour comes from the fact that although the essential modes are perpendicular to each other, they can still induce significant variations in multiple structural variables. In the case of twist at 10 bp, some trajectories exhibit larger variations in the twist angle (Ω) caused by the projections of the mode associated with roll, rather than the mode that should have been associated with twist. Consequently, this is an indicator of an incorrect classification. In reality, most principal modes induce variations in all structural parameters. However, before making further observations, it is essential to refine our method to align with the intended goal of associating single modes to each of the diagonal elastic constants (A_ρ, A_τ, C, B).

5.2.2 Modes comparison and second classification

For improving our classification process, we need a metric to find the mode that best represents that elastic constant. Since the modes are eigenvectors, we use the dot product as the metric to compare how similar two modes are. With this metric, we calculate one representative mode per each particular elastic constant at each length. Then, instead of associating according to the covariance matrix V, in our second classification we associate modes according to 'similarity' γ, where the set of representative modes are used as 'seeds' in which each represents a particular elastic constant. The 20 modes of each sub-fragment are then associated according to their similarity with the seeds.

Inspired by the work of Noy and collaborators [116], we compare modes using the dot product (see methods section 2.6.1) to evaluate the similarity (γ) between two modes. Before calculating the similarities, we apply the pre-process specified in the methods section 2.6.1, where for all sub-fragments, we strip the atoms that do not correspond to the backbone in order to be able to compare the principal modes of sub-fragments with different sequences. With this, we construct a 'dot product matrix' $M^{K,l}$, which contains information of how similar are the essential modes that were associated with a particular elastic constant ($K = A_\rho, A_\tau, B, C, D, G, H$) at length l. Each element of the elastic matrix $M^{K,l}$ compares the principal modes ($\vec{e}_m^{\,U}$ & $\vec{e}_{m'}^{\,U'}$) of two sub-fragments (U and U') of the same length (l) that were associated with the elastic constant (K):

$$M_{U,U'}^{K,l} = \gamma_{U,V} = |\vec{e}_m^{\,U} \cdot \vec{e}_{m'}^{\,U'}| \tag{5.9}$$

where both mode numbers m and m' can take the values $1, 2, ..., 20$, and are not necessarily different. Each row/column of matrix $M^{K,l}$ corresponds to a sub-fragment of either the 52mer or 62mer, where we order the sub-fragments as $U = 1, 2, ..., N_l^{52}, N_l^{52} + 1, ..., N_l^{52} + N_l^{62}$, with $U = 1$ corresponding to the sub-fragment that was stripped from the left end of the 52mer, $U = N_l^{52}$ from the right end of the 52mer, $U = N_l^{52} + 1$ the one stripped from the left end of the 62mer and $U = N_l^{52} + N_l^{62}$ the one from the right end of the 62mer. Once calculating every similarity index γ, the dot product matrices take the form:

$$M^{K,l} = \begin{bmatrix} 1 & \gamma_{1,2} & \cdots & \gamma_{1,N_l} \\ \gamma_{2,1} & 1 & \cdots & \gamma_{2,N_l} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{N_l,1} & \gamma_{N_l,2} & \cdots & 1 \end{bmatrix} \tag{5.10}$$

Notice that the dot product matrices are symmetrical with size $N_l \times N_l$ with $N_l = N_l^{52} + N_l^{62}$, and that every element $0 \leq \gamma_{U,U'} \leq 1$.

We can then use these dot product matrices to find the most representative mode for each elastic constant at each length. This representative mode is the vector whose sum of dot products yields the largest number or, in other words, it can be seen as the mode that is most similar to the rest of the modes within the same classification. In general, the most representative mode $\vec{v}_K$, of an elastic constant K, is found by calculating the row U whose sum yields the largest number:

$$\max_U \left(\sum_{U'=1}^{N_l} M_{U,U'}^{K,l} \right) = \max_U \left(\sum_{U'=1}^{N_l} \gamma_{U,U'} \right) \tag{5.11}$$

Once we have identified the most representative modes ($\vec{v}_K$) for every elastic constant and at every length, we use them in a second classification process as 'seeds' to associate one of the 20 essential modes of each sub-fragment U to each elastic constants according to their similarity with $\vec{v}_K$ rather than the covariance matrix V. This is done by finding the mode $\vec{e}_m^{\,U}$ (with $m = 1, 2, ..., 20$) that yields the largest product with $\vec{v}_K$:

$$\vec{e}_m^{\,U} = \max_m (|\vec{e}_m^{\,U} \cdot \vec{v}_K|) \tag{5.12}$$

where $\vec{e}_m^{\,U}$ is the essential mode of sub-fragment U associated with K. The mode $\vec{e}_m^{\,U}$ can also be seen as the vector that is most parallel to $\vec{v}_K$ from the set of 20 essential modes. Once we have associated the modes $\vec{e}_m^{\,U}$ with their respective elastic constants, we calculate a new dot product matrices $M^{K,l}$ and update the most representative modes $\vec{v}_K$ for each K at every length l.

Heatmaps of figure 5.6 show the dot product matrices ($M^{C,12}$) of the twist elastic constant C at the length of 12 bp, for the first and second classification. In the case of the first classification, it can be seen that many of the modes that were associated with C have low similarity ($\gamma < .25$), which is an indicator of an inaccurate classification. In the second classification, we see that most of the modes have high similarity ($\gamma > .75$), which demonstrates that our second classification increases the accuracy of associating modes with a particular elastic constant. The second classification also reduces the variability in mode number m that were related with an elastic constant K, where in case of $M^{C,12}$ in the second classification (see figure 5.6), all modes have $m = 3$.

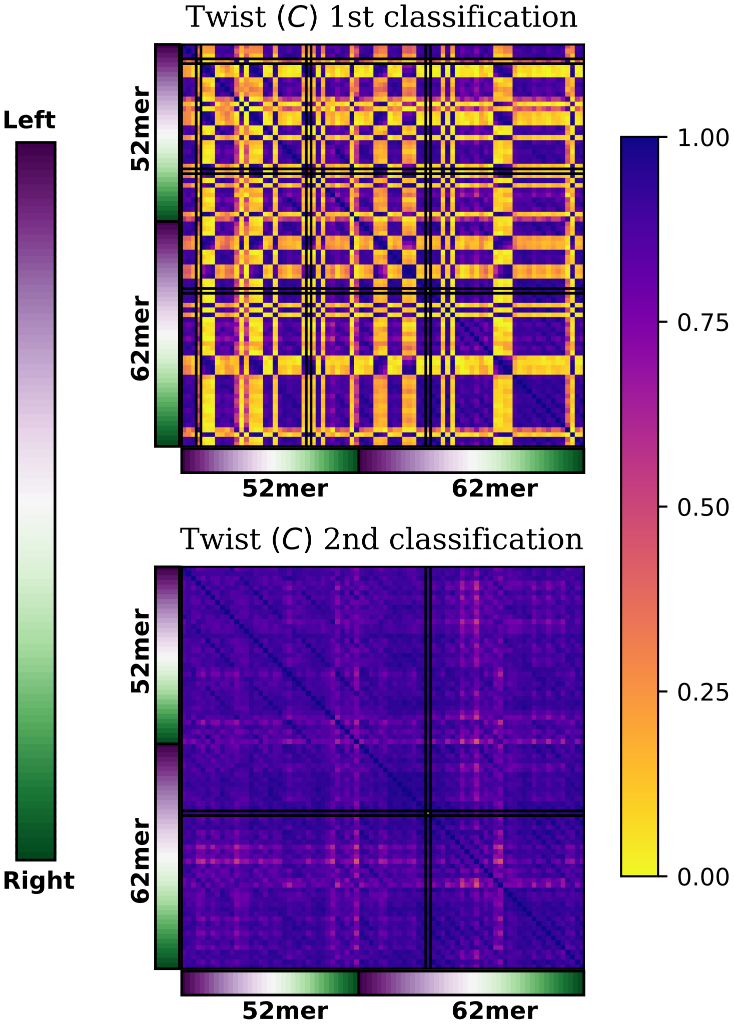


Figure 5.6: Heatmap of matrix $M^{C,12}$ which contains the dot products between eigenvectors associated with the twist elastic constant C at the length of 12 bp for the first and second classifications. Each row/column of $M^{C,12}$ corresponds to a sub-fragment U of either the 52mer or 62mer, which are sorted according to which end (left or right) they were extracted from. The black lines of the second classification indicate the representative mode $\vec{v}_C$ of the twist elastic constant at 12 bp, which corresponds to a sub-fragment extracted from the left end of the 62mer.

However, it is worth pointing out that the mode number m should not be used as a rule for relating modes with elastic constants, as the number m is just an indicator of how much system variance a mode contributes. For example, not always the mode $m = 3$ represents the twist elastic constant.

Donut plots shown in figure 5.7 summarize the second classification of elastic constants at lengths ranging from 4 to 38 base-pairs. The plots reveal that at short lengths (4 bp), the essential modes appear to be randomly associated with the elastic constants. This finding suggests that in order to capture the soft flexibility observed at short lengths, a combination of multiple modes, rather than single modes alone, is necessary. These results highlights the importance of adopting a multi-mode approach when investigating the flexibility of DNA at short lengths through the analysis of essential modes, thus providing valuable insights for future studies.

At longer length-scales (>4 bp), we observe the same mode numbers m tend to be associated with the same elastic constants (see figure 5.7). The first two modes ($m = 1, 2$) that contribute most of the system variance are generally associated with roll (A_ρ) and tilt (A_τ). In general, $m = 1$ tends to be associated with A_ρ while $m = 2$ with A_τ, but at lengths greater than 12 bp, these associations tend to oscillate more. This is related with the geometrical definition of tilt and roll in the LDEM, where their elastic profiles show that the stiffness of the two elastic constants oscillate, where at some lengths one is softer than the other (see figure 4.4b-c). Regarding the essential modes, the system variance is the quantity that oscillates when assigning either $m = 1$ or $m = 2$ to tilt or roll.

Focusing in the twist elastic constant (C), we see that at the lengths of 6 and 8 bp, some modes associated with A_ρ and A_τ are also associated with C (see figure 5.7), which indicates that most of the fluctuations in twist, are caused by the modes that are related to roll and tilt. At the length of 10 bp (approximately one DNA turn), the third mode ($m = 3$) tends to be associated with C, where we can say that this mode characterises the bulk flexibility in twist as it arises at the length in which the C elastic profile reaches bulk behaviour (see figure 4.4a). Then, at the length of 26 bp (2.5 turns) the mode associated with twist switches to $m = 5$. The modes related with the stretch modulus (B) follow a complex behaviour similar to the B elastic profile (see figure 4.4e). From 6 to 12 bp (.6 to 1.1 turns), the mode associated with A_ρ tends to also be related with B, which means that the fluctuations caused in B are mainly due to the mode that is related to A_ρ. This agrees with the H coupling profile since, at these lengths, H presents its largest maximum (see figure 5.1), which also corresponds to the length with the maximum correlation between roll and stretch. At longer length-scales, modes independent of A_ρ tend to be associated with B, which is $m = 5$ at 14 bp (1.3 turns), then $m = 6$ at 18 bp (1.5 turns) and $m = 7$ at 24 bp (2.3 turns). These lengths also coincide with the behaviour observed in the stretch profiles, where around one DNA turn B reaches a maximum, then presents relaxations approximately at 1.5 and 2.3 turns. Lastly, it is worth mentioning that even that the 62mer has 10 more sub-fragments than the 52mer at any length ($N_l^{62} > N_l^{52}$), the modes classification is almost identical in the two molecules (see figure 5.7).

When considering the modes associated with the twist-stretch (D), twist-roll (G),

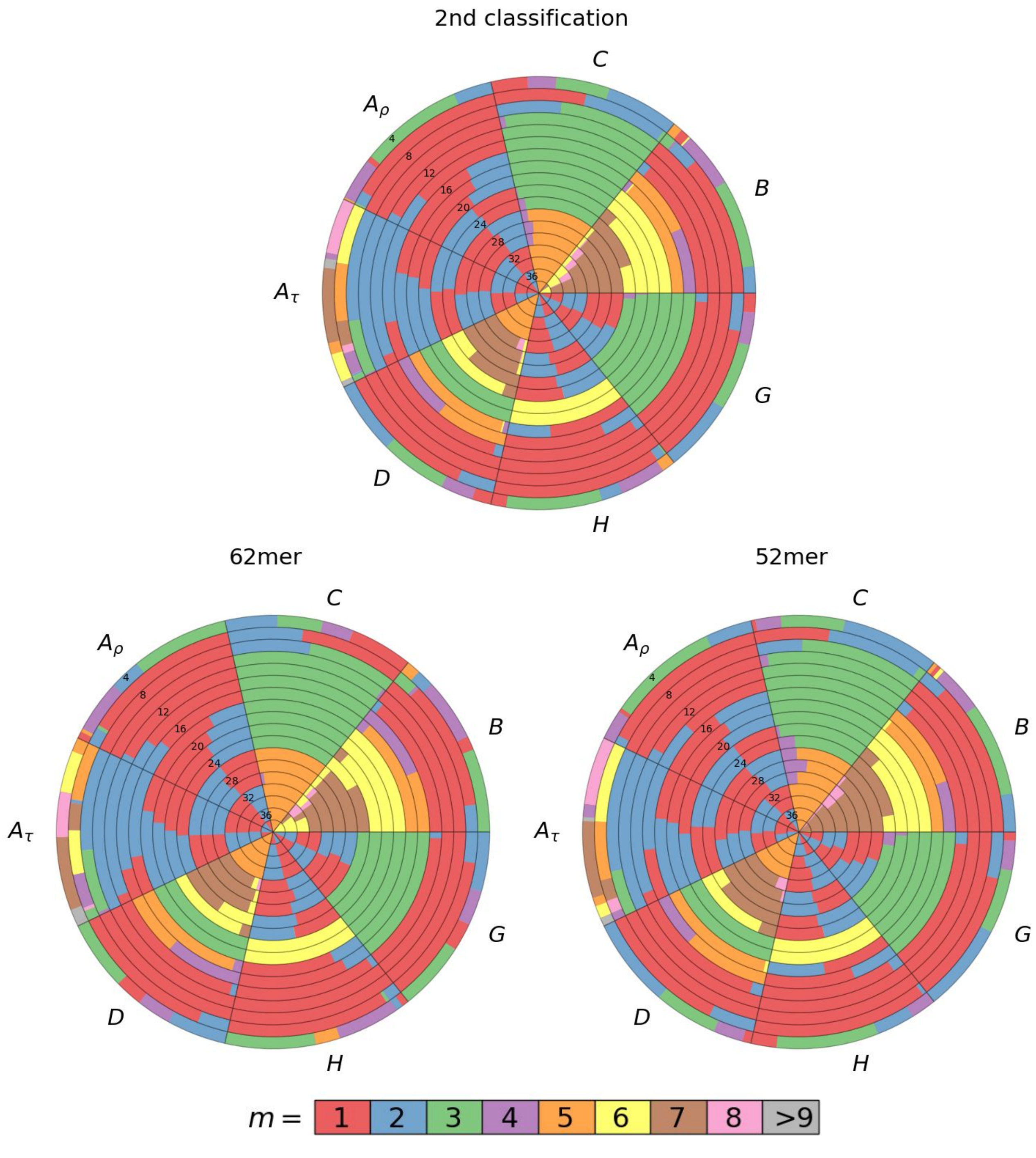


Figure 5.7: Donut charts summarising mode association with the 7 elastic constants resulted from the second classification for all sub-fragments. Top donut chart shows the classification considering the two DNA molecules, while the bottom charts correspond to the classifications of the 52mer and 62mer separately. Each radius represents a different length, being 4 bp the external up til 38 bp at the centre. Colour is associated with the mode number, being grey for numbers $m \geq 9$. Wedge width represents the percentage of modes with number m that were associated to the given elastic constant at length l.

and roll-stretch (H) couplings, we observe two interesting tendencies. Firstly, at shorter lengths ($l \leq$12 bp), the modes associated with roll (usually $m = 1$) are also associated with all three couplings. This result indicates that, at short length-scales, the mode related with A_ρ, which explains a significant portion of the system variance, is responsible for causing major fluctuations in the D, H and G couplings. Secondly, at longer lengths ($l \geq$14 bp), the modes associated with the couplings tend to be associated with one of the two modes assigned to the elastic constants they influence; for example, the modes associated with stretch (B) and twist (C) are also associated with twist-stretch (D). This transition is driven by the mode association of B, as its own mode emerges precisely at this length. These observations imply that the transition between local to bulk flexibility occurs between 12 to 14 bp (equivalent to 1.1 to 1.3 turns), primarily influenced by the essential mode associated with B, which is consistent with previous findings [114]. This transition may explain the distinct behaviors exhibited by the elastic profiles of B and the three couplings, where profiles present different behaviours at local and global levels. Additionally, it may explain the challenges faced by our mathematical expressions for accurately predicting the profiles of D, H, and G at short length scales (as shown in figure 5.3).

Lastly, the donut plots in Figure 5.7 provide insights into the relationship between modes and the elastic constants. Usually, modes $m = 1, 2$, and 3 correspond to roll, tilt, and twist, respectively. However, as the length increases, we observe higher mode numbers associated with twist (C) and stretch (B). For example, at the length of 26 bp, the mode $m = 5$ is associated with C, and mode $m = 7$ with B, while modes $m = 3, 4$, and 6 are not associated with any elastic constants. This result draws attention to two aspects. First, there are modes that significantly contribute to the system variance and influence DNA flexibility, although they do not cause major variations in our elastic constants. Second, our method is capable of capturing more system variance at certain lengths.

In future research, it would be valuable to investigate the interplay of multiple modes on each elastic variable as shown in figures 4.4 and 5.1. However, our current method is not suitable for studying these multiple modes. Nonetheless, these results align with our initial expectations that multiple modes affect DNA flexibility rather than a single mode for each elastic constant. Moving forward, our analysis will focus on the essential modes associated with the elastic constants identified in our second classification. These modes exhibit the greatest influence on their respective elastic constants, which will be the primary focus for the rest of this chapter.

The analysis presented in this subsection highlights the principal modes primarily influencing our set of four elastic constants and their respective non-zero couplings. We also observed that the modes directly associated with the diagonal elastic constants (A_ρ, A_τ, C, and B) also have direct associations with their respective couplings (D, G, and H). Additionally, these modes exhibit high similarity across different DNA sequences, and we have developed a methodology to identify and classify them accordingly. While our methodology can be further refined to include multiple modes, we anticipate its utilization in future research. Moreover, we believe that our method has the potential to capture a broader range of system variance at specific lengths, which will be further discussed in the next section 5.3.

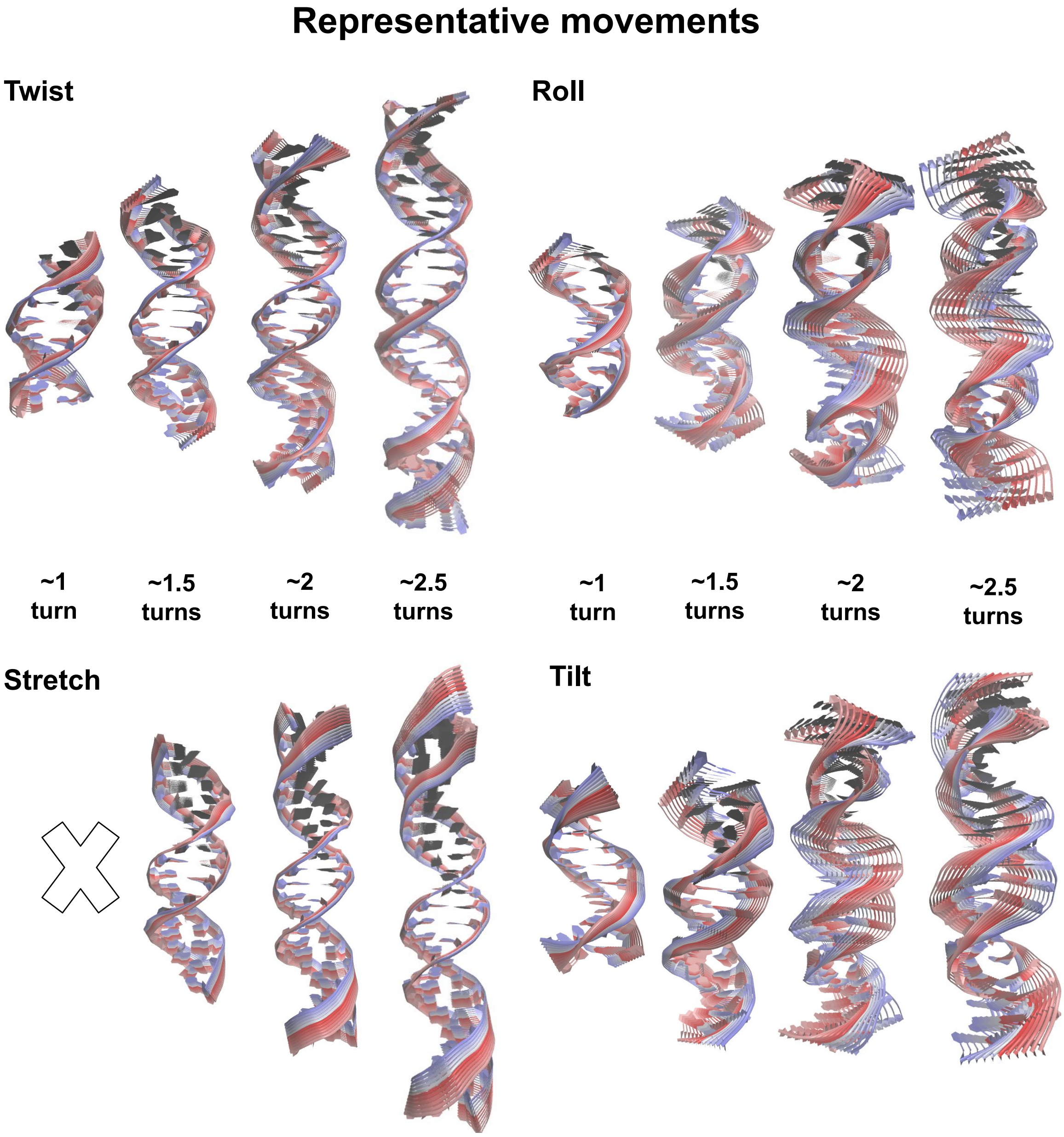


Figure 5.8: Superimposed animations of the representative modes of twist, roll, stretch and twist at key lengths, obtained with the PCAsuite [138] software. Red and blue colours indicate projections in opposite directions (negative and positive). Stretch does not have a proper mode at ∼1 DNA turn as it arises at longer lengths (see figure 5.7).

5.2.3 Representative modes

Figure 5.8 shows the animations of the four modes that best represent the diagonal elastic constants at the lengths of 1, 1.5, 2 and 2.5 DNA turns, and resulted from the second classification. From these animations, it seems that the modes associated with roll and tilt cause more variations in the molecule than the other two modes. Also, the modes related with roll and tilt are similar with the difference that they bend towards opposite directions. On the other hand, it seems that twist and stretch are highly correlated as they cause deformations that seem to have an impact on the twist and end-to-end distance parameters. Similarly, it seems that the modes associated with twist and stretch cause major increments in the molecule's distance. A more quantitative analysis of the structural deformations that the representative modes cause, will be performed in the next section, however, the animations presented here qualitatively indicate that the modes were associated correctly to their respective elastic constants.

For the rest of this thesis, instead of describing the essential modes with the number m, we will refer to them as roll (A_ρ), tilt (A_τ), twist (C) and stretch (B) modes. The reason for this is because the number m is not particularly related to the elastic constants as it is an indicator of how much a mode contributes to the system variance. Hence we prefer to refer to them by the elastic constant that they represent after the second classification has been performed.

5.3 Analysis of the essential modes of DNA

In the previous section, we presented the general process and results for calculating the essential modes of DNA and their mode association with the elastic constants. Here, we analyse in more detail these associated modes. Firstly, we examine their eigenvalues to understand the proportion of system variance they contribute. Subsequently, we perform a covariance analysis to explore how these modes directly influence the flexibility of DNA, particularly focusing on the variance of our structural variables across different lengths. Lastly, we gain insight into the structural deformations the associated modes induce by analysing the animations generated by the PCAsuite software [138]. These analyses provide a deeper understanding of the role played by the associated modes in shaping the behavior of the flexibility of DNA.

5.3.1 System variance explained by the essential modes

In this subsection, we analyse the system variance explained by the associated modes. Firstly, we are interested in investigating the amount of system variance the essential modes calculated from our method are able to explain. Additionally, we investigate the contributions of roll, tilt, twist and stretch modes to the system variance at different lengths. Furthermore, we examine the relationship between the elastic profiles (see figure 4.4) and the system variance explained by these modes. Lastly, we evaluate the effectiveness of our classification process by analysing the total amount of system variance our method is capable of capturing.

As mentioned earlier, usually a few modes are required to rebuild 90% of the system variance, and PCAsuite [138] sorts them according to the amount of variance that

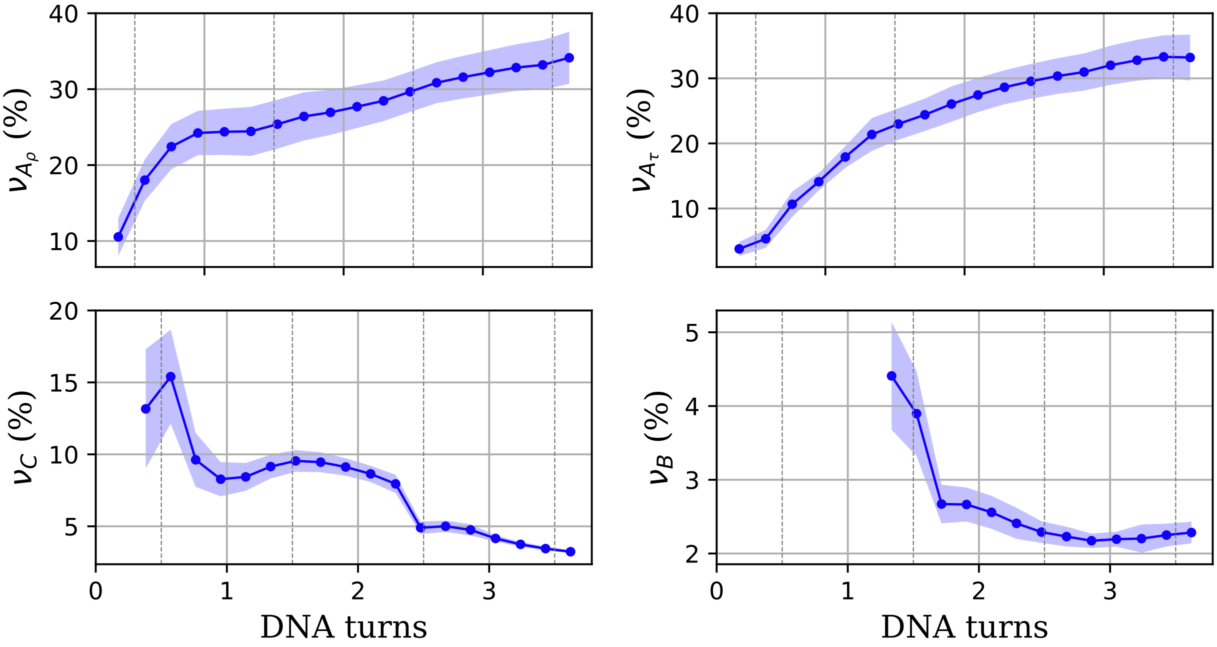


Figure 5.9: Proportion of system variance ν explained by the four essential modes associated with roll (A_ρ), tilt (A_τ), twist (C) and stretch (B). These proportions were obtained after the second classification was performed and results shown correspond to values from both 52mer and 62mer. Solid lines represent averages and shadowed areas standard deviations. The ν_B panel only shows proportions at lengths in which the stretch mode (B) is associated (higher than ~1.5 DNA turns).

they contribute. The eigenvalues λ, are related with the amount of system variance ν that each essential mode explains, where the sum of eigenvalues corresponds to the total system variance. Hence, using equation 2.116, we can calculate the proportion of system variance ν_K that the essential mode associated with the elastic constant K explains (see figure 5.9).

We observe that the amount of system variance that the roll and tilt modes (A_ρ and A_τ) explain increases as a function of length, arriving around 60% at 2 DNA turns. This reflects the qualitative results shown in the donut charts 5.7, where they indicate that these two modes usually take the numbers $m = 1$ and $m = 2$. This also agrees with the static animations shown in figure 5.8, where it can be observed that the bends caused by the roll and tilt modes induce greater variations in the bp positions in contrast to the twist and stretch modes. It is interesting to note that both roll and tilt modes have similar amounts of system variance at lengths longer than one DNA turn. This agrees with their elastic profiles (see figure 4.4b-c) which indicate that their stiffness have approximately the same magnitude (≈ 70 nm). Nonetheless, at lengths shorter than one DNA turn, A_ρ explains a greater amount of system variance than tilt. This observation aligns with the flexible nature of the roll elastic constant at short lengths, as illustrated in figure 4.4b-c. These findings further support the notion that the predictions for roll and tilt, obtained through the LDEM, are consistent with the principal modes. Finally, it is worth highlighting that the combined contribution of the modes of roll and tilt to the system variance is less than 25% at lengths shorter than one DNA turn. This observation suggests a more uniform distribution of eigenvalues at short lengths compared to longer lengths.

Regarding twist (C), we observe that it contributes less to the system variance ($\nu_C < 18\%$) than A_ρ and A_τ. This is expected since the mode $m = 3$ is usually associated with C (see figure 5.7). We also observe that the twist mode can be characterised by three different sections in which ν_C present sharp changes. These changes take place at lengths in which the twist mode transitions to modes that contribute less system variance. This is highlighted in the donut chart of figure 5.7, where the twist mode transitions from $m = 2$ to $m = 3$ at $l = 8$ bp, then from $m = 3$ to $m = 5$ at $l = 26$ bp. The stretch mode (B) presents the lowest contribution to the system variance at any length ($< 5\%$), and similar to the twist mode, its behaviour sharply changes at $l = 20$ bp as B transitions from $m = 5$ to $m = 7$. The relatively low contributions of C and B reflect the qualitative results of figure 5.7, since they are assigned to higher mode numbers m at longer length-scales. Lastly, the low contributions of twist and stretch are mainly due to the fact that those two modes tend to primarily affect the base-pairs at the ends of the molecule, in contrast of roll and tilt which affect the movement of all the base-pairs (see the superimposed animations of figure 5.8).

One important aspect to highlight is that our method calculated modes that can explain approximately 65% of the system variance. This is a clear indication that we are missing a few modes to at least explain 90% of the system variance, which is a standard percentage when implementing PCA. However, it is important to emphasize that our method specifically aims to identify the essential modes that exhibit the highest covariances in the structural variables. These essential modes may not necessarily coincide with the modes that contribute the most to the system variance. Consequently, the unaccounted modes might be responsible for combined deformations, which their capture is beyond the scope of our method. Nonetheless, their contributions remain significant and further investigation is needed to study them in depth.

Furthermore, the relatively low standard deviations of ν_K (see figure 5.9) is an indicator that our classification process effectively captures the most significant movements associated with the elastic constants.

The analysis presented in this subsection reveals that four key modes, primarily associated with roll, tilt, twist and stretch, collectively account for approximately 65% of the system variance. Specifically, roll and tilt modes demonstrate similar contributions of around 30% each, while the combined system variance explained by twist and stretch modes is less than 20%. To the best of our knowledge, no other studies have utilized PCA to directly establish a connection between structural deformations and essential modes. Some investigations have explored the similarity of movements between RNA and DNA molecules by analysing 500 essential modes [116], while others have focused on the essential modes capturing 90% of the system variance in naked DNA molecules and their ability to adopt protein-DNA binding conformations [39].

Nevertheless, our method has a limitation as it associates one mode with each elastic variable, limiting the capture of essential modes to a maximum of four per structure. Future investigations can enhance our method to explore the set of essential modes that capture at least 90% of the system variance and further investigate their relationship with the elastic variables.

5.3.2 Variances of the four structural parameters

Now that we have corroborated that our classification method is accurate in associating essential modes to elastic constants, and have analysed how much system variance these modes contribute, we proceed to analyse the elements of the covariance matrix (V). This analysis is performed by rebuilding trajectories using different combinations of modes:

- Rebuilt trajectories with one mode: Trajectories are rebuilt using equation 2.114 with individual modes (A_ρ, A_τ, C and B). Four trajectories are obtained for each sub-fragment, and a V matrix is calculated per trajectory.
- Rebuilt trajectories with combined modes: Trajectories are rebuilt by combining all four modes using equation 2.115. One trajectory is obtained for each sub-fragment, and a covariance matrix V is calculated. We denote this type of trajectory with the symbol ξ

Figure 5.10 shows the averages and standard deviations of the calculated variances for trajectories rebuilt according to one mode (A_ρ, A_τ, C, B), the four modes combined (ξ) and original trajectories (O). Averages are calculated for all sub-fragments with same length l.

Focusing on the variance of the roll angle (V_ρ), we observe that the A_ρ mode induces most of the fluctuations on this parameter while the A_τ mode can only induce small

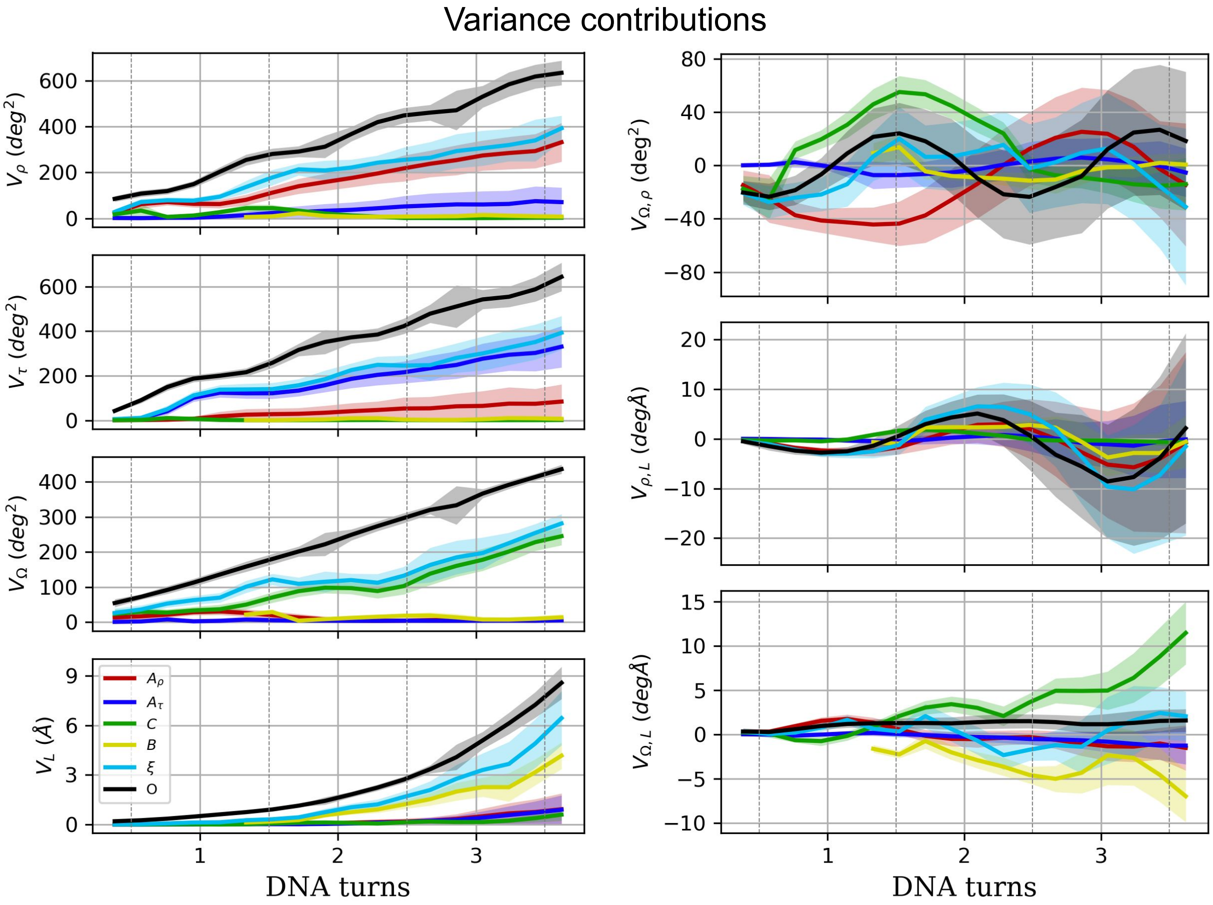


Figure 5.10: Average of variances (left) and covariances (right) as a function of length, measured from rebuilt trajectories according to one mode (A_ρ, A_τ, C, B), to the four modes combined (ξ) and to the original trajectories (O). Shadowed areas represent the standard deviations.

variations (see figure 5.10). Similarly, the A_τ mode provokes major fluctuations in the tilt angle (V_τ), where the A_ρ mode induces relatively low variations. This indicates that the modes have a strong influence in one parameter and a small influence in the other one. This agrees with previous studies, which have indicated that the coupling between tilt and roll is weak [74]. Lastly, when calculating the variances V_ρ and V_τ for the ξ trajectory, their resemblance with the original variances slightly improves, explaining 60.1% of the total variance in the case of roll and 57.4% in the case of tilt

Analysing the variances of the twist angle (V_Ω) and the end-to-end distance (V_L), we observe similar behaviour than tilt and roll, where the C mode mostly influences V_Ω and the B mode V_L. However, in contrast to tilt and roll, only the C and B modes are the major contributors to its variances. We observe that the ξ trajectories provide descriptions of the 54.9% of the variance in case of V_Ω and 56.4% in case of V_L when compared to the original trajectory (see figure 5.10). We also observe that the combined trajectories greatly increase the percentage in case of V_L, where the B mode only provides 32.27% of the description.

Regarding covariances (right panels of figure 5.10), we observe that $V_{\Omega,\rho}$ and $V_{\rho,L}$ present periodic components just as the related G and H couplings, respectively. In the case of $V_{\Omega,\rho}$, we observe that the main contributors are C and A_ρ modes, which surprisingly both induce periodic couplings with opposite signs. In contrast to previous estimations of the twist-roll coupling (G) [111,141], these results indicate that the coupling might be negative depending if the molecule is deformed following one of these modes. When combining modes (ξ), the overall description of the twist-roll covariances improves, although it is still very different from the original covariance (O). Vaguely similar, the A_ρ mode presents a periodic component in $V_{\rho,L}$, while the B mode also has considerable contributions in this covariance. When the four modes are combined, the shape of the covariance resembles the original covariance but slightly out of phase, which could be caused by the discarded modes.

A surprising behaviour is that C and B modes cause opposite covariances in twist-stretch ($V_{\Omega,L}$), which, in other words, means that the B mode provokes a positive D coupling while the C mode causes a negative D coupling as observed in B-DNA [48], [86], [49], [134]. It is barely visible but, at lengths less than one DNA turn, there is a sign switch in the C mode, which corresponds to a sign switch in D from positive to negative. This is the same behaviour previously observed in the coupling profiles (see figure 5.1). Lastly, similar to the covariance $V_{\Omega,\rho}$, when combining the modes (ξ) the covariance becomes more similar to the original trajectory; however, more modes are required to provide a more accurate description.

Putting all these observations together, we can conclude the following points:

- Each associated mode principally influences its related structural parameter, which again validates our classification method for associating modes with the elastic variable that they influence the most.

- Couplings between two structural parameters, are mainly affected by their two associated modes.

- For reproducing the original variances and covariances, the combination of the four modes provides more accurate descriptions as it would be expected. This demonstrates the effectiveness of our method in capturing and associating the dominant modes with the elastic constants.

- However, the current method has limitations in capturing the full range of essential movements that reproduce the flexibility of DNA, and, similar to the case of system variance, the maximum amount of variance in the structural parameters that the combination of four modes can recreate is around 60%. This limitation can be addressed in the future by enabling the capture of several additional modes, which would allow us to obtain descriptions of at least 90% of the original variance.

- The tilt mode (A_τ) has little influence in the covariances, and mainly affects the tilt variance. However, it can still have little influence in the roll variance.

- The two modes that influence their respective coupling, can provoke opposite correlations.

This last finding is particularly intriguing as it carries significant implications for DNA-protein interactions, where the behavior of the DNA molecule can vary depending on the specific mode of deformation. For example, the covariance $V_{\Omega,L}$ (see figure 5.10) reveals that twisting the DNA along the C mode leads to an increase in the end-to-end distance, while stretching the DNA along the B mode causes the molecule to untwist. These observations are backed up by recent studies on DNA-protein recognition, where it has been found that the DNA is mechanically deformed along its essential modes to adapt its structure to the interacting protein [39].

Our findings suggest the existence of an additional layer of complexity in the DNA couplings, where the sign of the twist-roll (G) and twist-stretch (D) couplings can change depending on the specific context of molecular deformation. This highlights the dynamic nature of DNA-protein interactions and the intricate interplay between structural parameters. Further exploration of these complex couplings is warranted to deepen our understanding of the mechanisms underlying DNA-protein recognition and its implications on the flexibility of DNA.

5.3.3 Structural deformations

Now, we move to analyse the structural deformations induced by the associated modes and investigate the conformational space accessible to each mode. To this end, we analyse the animations that the software PCAsuite [138] produces for each representative mode at each length (see figure 5.8). Basically, these animations correspond to projections in the form of equation 2.114. We use SerraNA to analyse how the structural parameters change in the form of:

$$\Delta x_k = x_0 - x_k \tag{5.13}$$

where x represents one of the four structural parameters (ρ, τ, Ω, L), x_0 represents the value of x in the average structure and Δx the change in x at each animation frame

k.

Figure 5.11 shows the conformational space that each representative mode can access to, where each panel shows the relationship between two structural parameters (Δy vs Δx) and their increments were calculated using equation 5.13. Before analysing these results, we want to indicate that deformations between 4 and 38 bp are plotted within the same panel, where each continuous line corresponds to one animation and the magnitude of deformations is proportional to the length of the sub-fragment. A general behaviour that we observe in all panels, is that each representative mode causes most of the deformations in their respective structural parameters, as in agreement with the covariance analysis.

Analysing $\Delta\Omega$ vs ΔL in more detail, we further confirm our observations from the previous subsection, where the B mode is mainly responsible for a positive D coupling, where elongations along this mode cause the untwisting of the molecule. In contrast, the C mode causes a negative D coupling in agreement with values reported in the literature [48], [86], [49], [134]. However, we can see that, at lengths smaller than one DNA turn (see figure 5.10), D is positive which indicates a sign switch as previously observed in the coupling profiles 5.1. The roll mode (A_ρ) causes deformations in both structural parameters (twist and stretch) and in both directions, which indicates that the signs of the G and H couplings oscillate in agreement with the coupling profiles. In contrast, A_τ only causes deformations in the L parameter.

The $\Delta\rho$ vs ΔL panel of figure 5.11 indicates that the A_ρ mode is the one that causes most of the roll deformations (as expected), while all the other modes can also induce deformations in L. Observing the dependence of roll and twist in the $\Delta\rho$ vs $\Delta\Omega$ panel, we notice that A_ρ and C are the modes that mostly induce structural deformations, which corroborates the previous covariance analysis. However, in this case it is very notorious that the A_ρ modes present a curved behaviour rather than linear. This indicates that the more the DNA is bent through the A_ρ mode, the more it will twist/untwist. This is a feature that as far as we know, it has not been reported in the literature before.

In contrast to the other panels of figure 5.11, the ones related with the tilt angle (τ) indicate that the A_τ mode mainly affects τ, and barely the other deformation variables. Similarly, the A_ρ, C and B modes almost do not have any impact on the τ angle. This agrees with previous observations both in this thesis and in the literature [74] that there is almost no correlation between the tilt angle and the rest of structural variables.

In summary, the structural analysis provides similar insights obtained from the variance analysis discussed in subsection **??**. Consistent with our previous findings, the modes associated with roll (A_ρ), twist (C), and stretch (B) primarily influence their corresponding structural parameters of roll (ρ), twist (Ω), and the end-to-end distance (L). On the other hand, the A_τ mode mainly impacts the tilt angle (τ). Furthermore, the structural analysis reveals that the cross-terms exhibit a linear relationship across the modes, except for the twist-roll coupling (G) in the A_ρ mode, which deviates from a linear response when the molecule is deformed along that mode (see the $\Delta\rho$ vs $\Delta\Omega$ panel of figure 5.11).

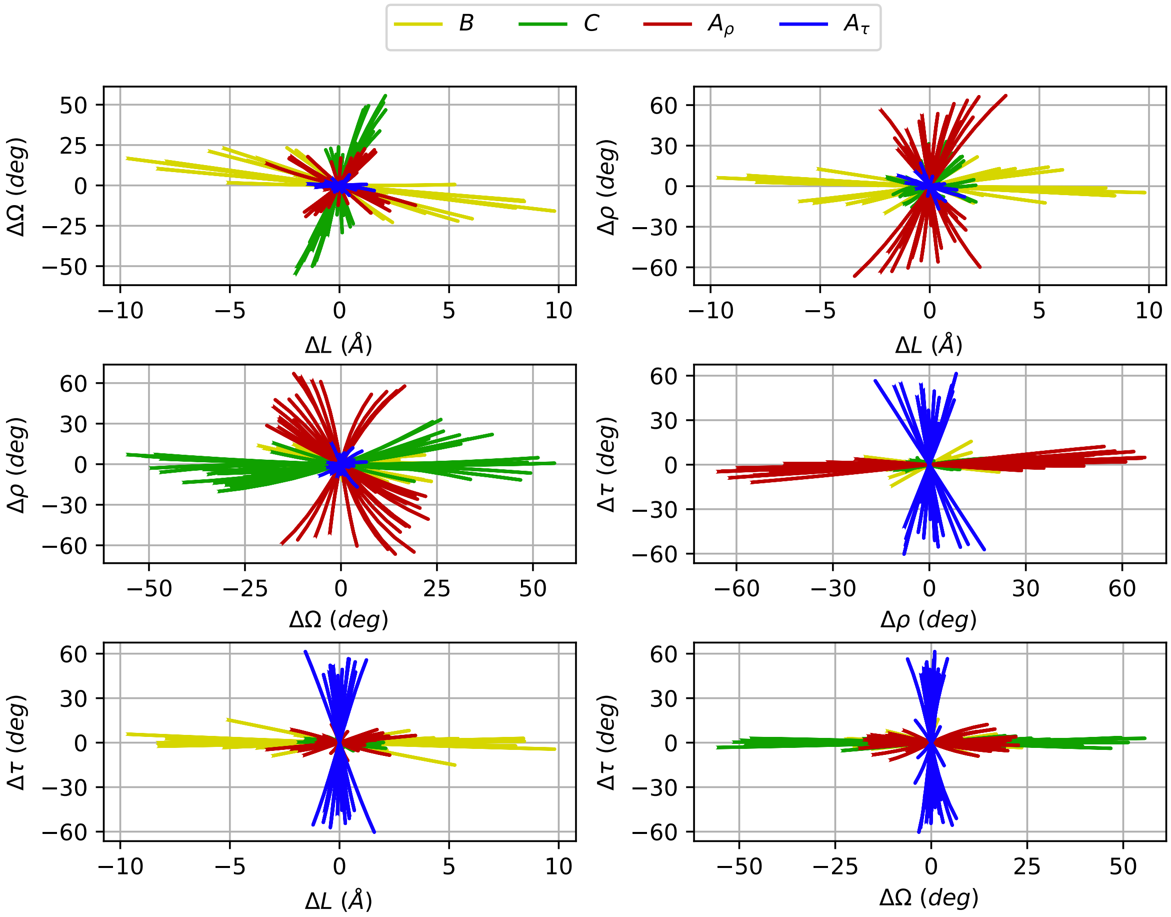


Figure 5.11: Structural deformations caused by the four modes associated to roll (A_ρ), tilt (A_τ), twist (C) and stretch (B), plotted against one another. The structural increments Δx were calculated using equation 5.13 from the animations of the representative modes (see figure 5.8). Results for all the analysed lengths (4 to 38 bp) are superimposed in each panel. mode colours have the same legend as in figure 5.10.

These findings provide further insights into the relationships between different structural parameters and their respective modes of deformation, revealing that the mechanical response of DNA can vary significantly depending on the direction/mode of deformation and length scale. By elucidating these relationships, we contribute to the broader knowledge on DNA conformation and its implications in DNA-protein interactions, as these deformations occur at length scales in which numerous proteins interact with the DNA molecule, including transcription factors [37, 72, 129] and nucleoid-associated proteins [1]. These assumptions are backed up by previous studies that have already suggested that the DNA undergoes deformation along its essential modes [39]. This knowledge can ultimately contribute to the development of more accurate models and predictive frameworks for understanding DNA-protein interactions. 1

5.4 Deforming DNA into complex 3D structures

In this section, our objective is to demonstrate the effectiveness of the modes associated with the elastic constants by testing their ability to adopt the structural shapes of more complex systems, such as supercoiled DNA minicircles or protein-DNA complexes such as nucleosomes.

To accomplish this, we have chosen two specific target structures. The first is a supercoiled DNA minicircle comprising 339 base pairs and with a linking difference of $\Delta Lk = -2$ ($\sigma \approx -0.06$) [125]. The second structure is the 1kx5 nucleosome, which consists of 147 base pairs and has been obtained from the BIGNASim database [59]. This nucleosome has previously been analyzed in the SerraNA chapter (see figure 4.8) [157].

5.4.1 Structure projection and RMSD minimization

It has been previously demonstrated by Orozco and co-workers [39] the ability of the DNA essential modes to adapt the DNA conformation to several DNA-protein complexes. They considered the set of essential modes that contribute 90% of the system variance to calculate the overlap between these essential modes and a vector $\vec{R}$ that measures the conformational transition between the bound (protein-DNA complex) and unbound (naked DNA) states. Inspired by this study, we employ a similar approach, where we project our four representative modes (see figure 5.8) onto the target trajectories. Our goal is test the capacity of these modes to adopt the structural conformations of deformed structures.

Our process consists in determining the projections p_K that minimise the RMSD with the target structure $\vec{X}$, which can be either a DNA minicircle or nucleosomal DNA. We initialize the process by selecting a structure of length l with an average structure $\vec{A}$ and its associated modes $\vec{e}_K = A\rho, A_\tau, C, B$. We choose $l = 16$bp because at this length, the associated modes correspond to the first five essential modes ($m = 1 - 5$), as indicated in the donut charts of figure 5.7. Moreover, at this length, the associated modes contribute the most to the system variance ($\nu \approx 60\%$), as shown in figure 5.9.

Next, sub-fragments of the same length l are extracted from the target structure (minicircle or nucleosome) contiguously until the whole molecule is covered. This results in n sub-fragments $\vec{X}_j$, where j represents the middle position of $\vec{X}_j$ within the structure $\vec{X}$. Here, n is given by $n = N_X - l + 1$, with N_X being the number of base-pairs in the target structure.

Similar to the procedure described in the second classification (see subsection 5.2.2), we apply a pre-process, where we remove atoms that are not part of the backbone for $\vec{A}$, $\vec{e}_K$ and $\vec{X}_j$ to ensure a fair comparison of sub-fragments with different sequences.

Finally, for each middle position j, we calculate n projections ($p_{K,j}$) by minimizing the following equation:

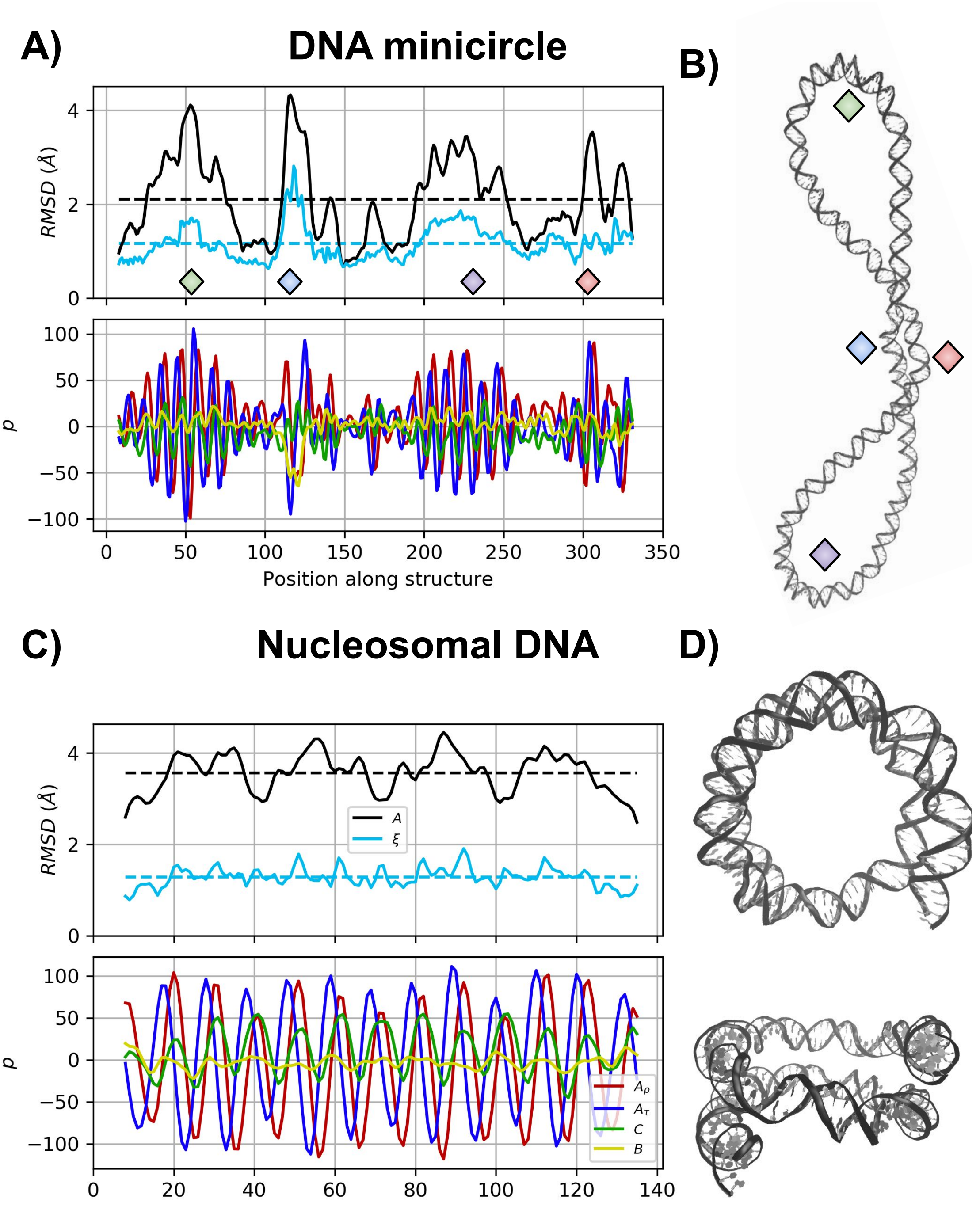


Figure 5.12: Projections of the essential modes representing roll tilt, twist and stretch deformations enables to get close to the DNA deformed structure due to supercoiling and nucleosome formation. Top, RMSDs between the projected and target structures of DNA minicircles (A,B) and nucleosomal DNA (C,D). Bottom, the extent of the projections (p), which is indicative of their relative importance. Black lines indicate RMSDs between the average structure A and the target structures. Cyan curves show the RMSD between the projected structure ξ and the target structure. Dashed lines correspond to the averaged RMSDs. Green and purple diamonds indicate high bends that are located at the U-turns of the supercoiled DNA minicircle, while blue and red diamonds indicate high bends at the crossing section. Here we consider sub-fragments of $l = 16$ bp.

$$\min_{p_{K,j}} f_j(p_{K,j}), \quad f_j(p_{K,j}) = RMSD\left(\vec{X}_j, \vec{\xi} = \vec{A} + \sum p_{K,j}\vec{e_K}\right) \tag{5.14}$$

where for each stripped sub-fragment, we obtain a function f_j, where the middle positions are given by $j = \frac{1+l}{2}, 1 + \frac{1+l}{2}, ..., n + \frac{1+l}{2}$. Notice that the only unknown variables in this equation are the projections $p_{K,j}$, which can be obtained by minimising the equation. To accomplish this, we use the Nelder-Mead algorithm [108] from the scipy python library [161] to minimise these functions f_j.

Our approach offers several advantages over Orozco's method [39]. Firstly, it allows for the comparison of structures with different sequences while maintaining the same length. This flexibility is beneficial when studying the effects of sequence variations on structural conformation and essential modes space. Secondly, the projected structures resulted from our approach provide atomistic detail, enabling further analysis using tools such as SerraNA or other software.

However, similar to previous subsections, there are limitations to our approach. One limitation is that the projected structures may struggle to adapt to complex conformations, such as those of DNA-protein complexes or supercoiled DNA. This is due to the fact that the four modes combined can only explain approximately 60% of the system variance of naked B-DNA. Increasing the number of modes considered may yield more accurate descriptions of complex conformations.

Another potential risk of our approach is that the flexibility of DNA is sequence-dependent. Therefore, it may be challenging to approximate complex conformations using projections obtained from different sequence contexts. However, even when considering this risk, our objective remains the same: to demonstrate that the essential modes associated with the elastic constants can be employed to approximate complex structures, even when derived from different sequences. We hypothesize that DNA possesses similar essential modes, regardless of the specific sequence context. This hypothesis is supported by the second classification process presented in Section 5.2.2, where modes were clustered based on similarity using the dot product between essential vectors. While this general assumption might be true for most of DNA sequences, it is worth noting that in very specific sequence contexts such as A-tracts, this assumption might not be valid as they could exhibit extreme mechanical responses, as highlighted in the SerraNA chapter 4.

To validate our method, we compared the minimised RMSD (ξ curve in figure 5.12) against the RMSD between the average structure A (it is the case in which $p_{K,j} = 0$) and the target structures. Figure 5.12 shows the RMSD calculated for the projected structure (ξ) and the average structure (A). In general, we observe that in comparison with the average structure (A), the projected trajectories (ξ) reduce the RMSD by 45% in case of the DNA minicircle and 74% in case of the nucleosomal DNA. Both reductions indicate that the associated modes can adopt structural conformations of deformed DNA as previously stated in [39]. However, the nucleosomal DNA can be better described by our associated modes than the DNA minicircle. The reason behind this comes from the fact that nucleosomal DNA is constantly bent around histones, where our associated modes can uniformly approximate any region along the molecule. This behavior is reflected by the constant RMSD shown in figure 5.12c. In the case of the DNA minicircle, the structure presents four strongly bent regions (marked with diamonds in figures 5.12A-B), which are induced by the imposed superhelical stress. These regions deviate from B-DNA as demonstrated by the RMSD between $\vec{X}_j$ and $\vec{A}$. The projected structures are able to reduce the RMSD at these locations, although it seems difficult to adopt certain conformations. As previously hypothesised, the RMSD could be improved by considering more modes. Another possible case is that other sequences might be able to provide better approximations, as the projected trajectory was taken from the 52mer, which does not have the same sequence as the target structures.

Regarding the projections p along the target structures, we observe some similarities between the DNA minicircle and the nucleosomal DNA (see figure 5.12). In general, the projections of all modes oscillate with 1 DNA turn of periodicity, being more evident for the roll (A_ρ), tilt (A_τ) and twist (C) modes. Roll and tilt have similar amplitudes but are out of phase by a quarter of helical turn, while twist has the same phase as roll. The major difference between the projection profiles of the minicircle and nucleosome, is in their oscillations, where the amplitude of A_ρ and A_τ increases in the highly bent regions of the supercoiled DNA (diamonds of figure 5.12A-B). In contrast, in case of the nucleosomal DNA, the amplitudes of roll and tilt tend to remain constant. An interesting behaviour that we observe, is that projections of the C mode present twist-waves previously described by Carlon and co-workers [110, 143], which are induced by bending deformations via the twist-roll coupling G. Our results indicate that this feature is also manifested in the C mode, which further confirms our previous observations from the covariance analysis, where both C and A_ρ modes influence the G coupling (see figure 5.10). These results suggest that the twist-bend coupling (G), is intrinsic to the DNA essential modes.

Overall, our results demonstrate the potential of the essential modes associated with elastic constants to approximate complex DNA structures. However, we observed limitations in their ability to adopt highly deformed conformations, such as plectonomic DNA (supercoiled). To address this, future research should focus on incorporating a greater number of modes and exploring different sequence contexts. This would expand the available conformational space, allowing the projected structures to adopt more complex conformations. Furthermore, our analysis reveals that individual projections exhibit twist-waves, previously observed in the literature [110, 143]. These twist-waves arise from the direct coupling between bending and twisting, which our results indicate that both twist-waves and the twist-bend coupling are intrinsic to the essential dynamics of DNA.

5.4.2 Analysing structural parameters of projected structures

In this subsection, we utilize SerraNA to analyse the structural parameters of the two target structures: a supercoiled DNA and a nucleosomal DNA. We then examine and compare the structural parameters of the corresponding projected structures, which approximate the shapes of the target structures. These projected structures were previously derived in subsection 5.4.1 by projecting the four modes associated with roll (A_ρ), tilt (A_τ), twist (C), and stretch (B) onto the target structures by minimizing the RMSD. Overall, the objective of this subsection is to analyse the structural parameters

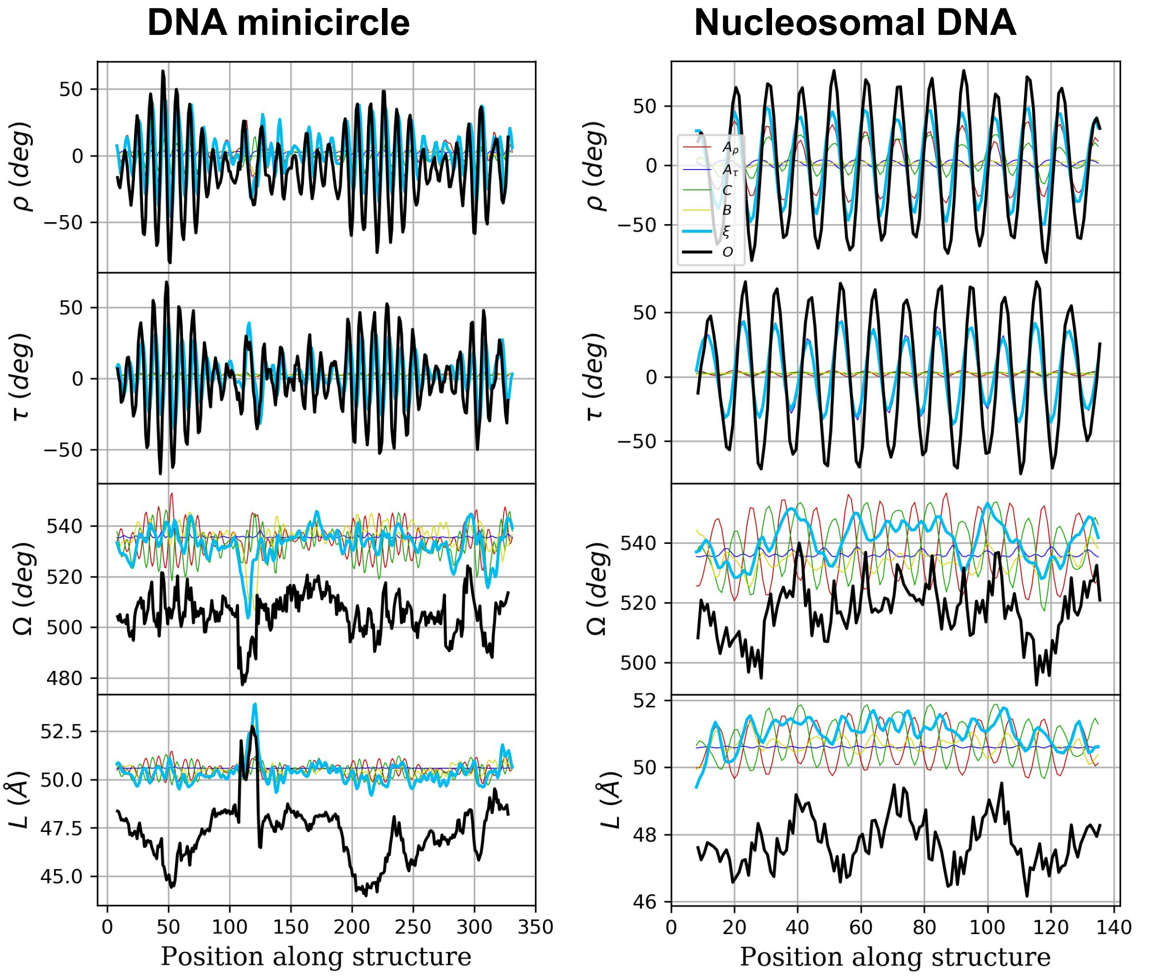


Figure 5.13: Structural parameters of the roll (ρ), tilt (τ) and twist (Ω) angles plus the end-to-end distance (L), for the original trajectory (O) (black), projected modes (ξ) (cyan) and individual projections of the roll A_ρ (red), tilt A_τ (blue), twist C (green) and stretch B (yellow) modes for the supercoiled DNA minicircle (left panel) and nucleosomal DNA (right panel). We consider sub-fragments of $l = 16$ bp.

of these target and projected structures, while simultaneously evaluating the accuracy and limitations of the projected structures in approximating the structural conformations of the target structures. Furthermore, we aim to explore the conformational space generated by these projected structures, as illustrated in figures 5.13 to 5.15.

Focusing in the roll and tilt angles, we observe that the projected structures (ξ) are able to qualitatively recreate the original measurements (O), where their roll angles are able to capture remarkably well the phase and periodicity of the oscillations, but fail to capture the amplitude of high bends (see figure 5.13). In case of the nucleosomal DNA it is more evident as the DNA is bent along the whole structure, but in case of the DNA minicircle this specially occurs in the U-turns located at 50 and 225 bp (see green and purple diamonds of figure 5.12). Regarding the oscillations in roll and tilt from the original nucleosomal trajectory, they have been previously observed in nucleosomal DNA [106] and relaxed minicircles [112] at the bp level. Here we calculated them at the level of l=16 bp so we are able to obtain clearer patterns since as investigated in the SerraNA section 4.2.5, nucleosomal DNA has a bendable structure that at lengths longer than the bp-step level, local bends can couple to give form to the global curvature. In case of the supercoiled DNA minicircle, the roll and tilt oscillations at the U-turns agree with the shapes predicted by theoretical models that consider the G coupling [110].

Regarding the twist angle Ω, the projected structures are able to qualitatively recreate the original curves for both supercoiled DNA minicircle and nucleosomal DNA, although, in general, they overestimate twist by 5.44% and 4.60% respectively (see figure 5.13). Certain periodicity in the twist angles can be observed in the original data for both cases. This periodicity becomes more evident in the projected trajectory ξ, which again highlights the existence of the twist-roll coupling (G) and the oscillations agree with previous studies [15, 110, 143], where twist waves (oscillations in the twist angle) have been observed and are originated by the coupling G between twist and roll.

In case of the end-to-end distance (L), an analogous behaviour is observed where the projected structures overestimate L by almost 7% in both structures and the periodicity in L is clearer, which again highlights the existence of the stretch-roll (H) coupling (see figure 5.13). One of the reasons that could explain why the oscillations are more clean in the projected structures ξ, is that several modes were filtered out and only four modes were considered. The filtered modes could be causing random fluctuations in the structures, which would be reflected in the noise that appears in the original structural parameters profiles.

Analysing the conformational space of individual projections, where the trajectories were rebuilt by using single modes, we can observe that these modes have contributions in multiple structural parameters for both supercoiled and nucleosomal DNA (see figures 5.14 and 5.15). For instance, in case of the roll angle, the main mode that causes deformations is precisely the roll mode (A_ρ), while twist (C) has a moderate contribution and stretch (B) a minor contribution (see ρ panels of figures 5.14 and 5.15). In case of the tilt angle, the main contributor is the tilt mode (A_τ), while the rest of the modes induce small variations in this structural parameter (see τ panels of figures 5.14 and 5.15). These two cases reflect the fact that roll and tilt modes are responsible

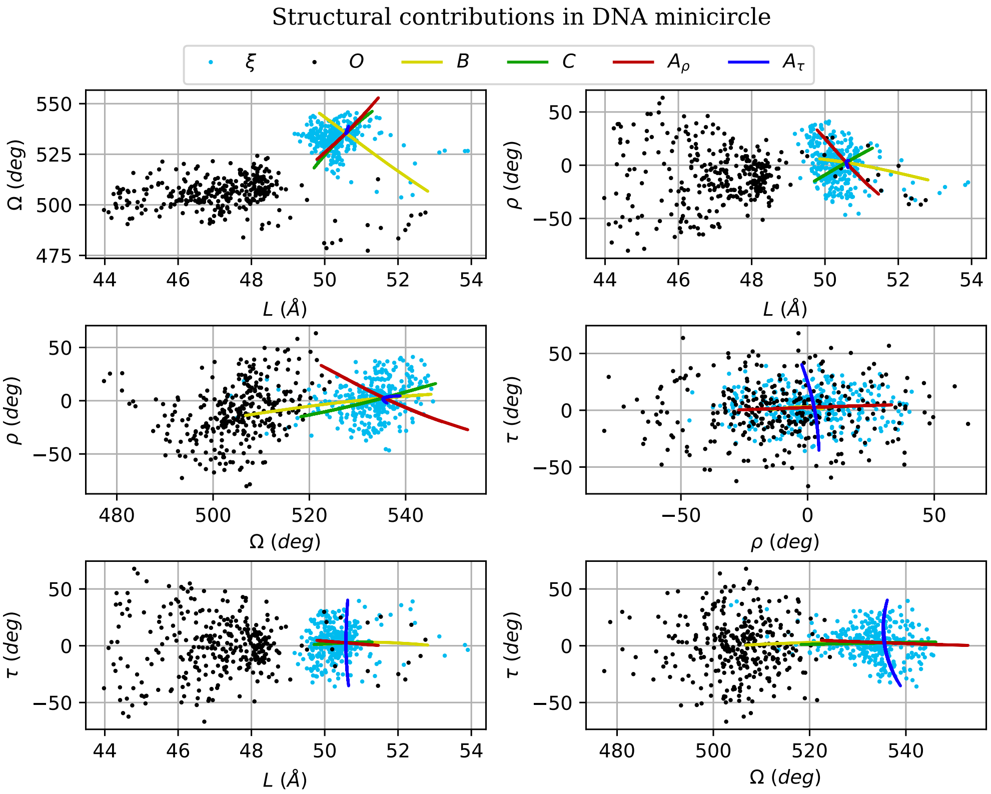


Figure 5.14: Conformational space sampled by the original trajectory O (black) and by the projected modes trajectory ξ (cyan) of the supercoiled DNA minicircle. Lines show the structural deformations along the roll A_ρ (red), tilt A_τ (blue), twist C (green) and stretch B (yellow) modes.

for most of the deformations regarding the roll and tilt angles (respectively), which corroborates the accuracy of our classification method.

Regarding the twist angle, the main contributors are the roll, stretch and twist modes (see Ω panels of figures 5.14 and 5.15). The roll and twist modes cause the same periodicity of 1 helical turn in Ω (see figure 5.13, 5.14 and 5.15). Although, they run in opposite directions, indicating a positive and negative twist-roll coupling as previously observed in the covariance analysis (see figure 5.10). These oscillations highly agree with the twist-waves observed in DNA loops and in nucleosomal DNA of figure 5.12D [110, 143]. We believe this oscillations are originated by couplings between the roll and twist angles. These couplings are intrinsic to the DNA essential modes, where each relevant mode (either A_ρ or C) have opposite signs and magnitudes.

The roll (A_ρ), twist (C) and stretch (B) modes cause major contributions in the end-to-end distance (see L panels of figures 5.14 & 5.15). A clear periodicity can be observed for these three modes and in different phases, as shown in figure 5.13. These oscillations in the end-to-end distance have not been reported in the literature, and because their similarity with the twist-waves [110, 143], we refer to them as stretch-waves. As stretch is correlated with roll, we deduce stretch-waves are induced by bending deformations and introduced via the stretch-roll coupling (H), highlighting the importance of this term.

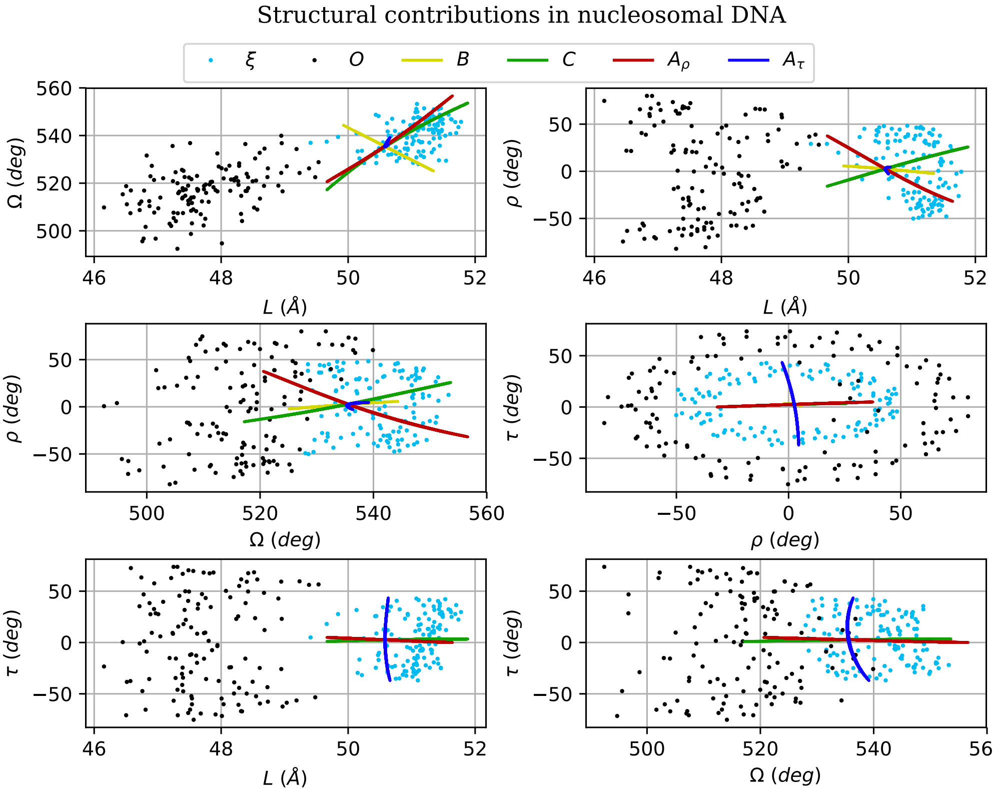


Figure 5.15: Conformational space sampled by the original trajectory O (black) and by the projected modes trajectory ξ (cyan) of the nucleosomal DNA [1kx5]. Lines show the structural deformations along the roll A_ρ (red), tilt A_τ (blue), twist C (green) and stretch B (yellow) modes.

As previously mentioned, one of the possible reasons as to why our projected structures are not able to obtain a higher accuracy when minimising the RMSD (see figure 5.12), is because the sub-fragment that we selected to calculate the projected structures does not have the same sequence as the target structures. In consequence, the conformations it can sample oscillate around its own average structure, which are not able to match the shape of the target structures (see figures 5.14 & 5.15). This is evident across several cases, where the projections fail to match the magnitudes of the original structural parameters (see figure 5.13). The accuracy may be improved by testing different sequences for calculating the projected structures, as well as considering more modes, which would be interesting to explore in future investigations.

Additionally, by examining the conformational space sampled by the projected structures illustrated in figures 5.14 and 5.15, we were able to identify that the modes associated with roll, twist and stretch are able to influence multiple structural parameters. This observation suggests that elastic couplings are intrinsic to the essential modes of DNA, and that these modes can exhibit opposite coupling behaviours in terms of sign and magnitude. These couplings give origin to oscillations presented in the twist angle and in the end-to-end distance when the DNA is bent, which have previously been denominated as "twist-waves" [110, 143], and similarly we term them as "stretch-waves" in case of stretching oscillations. On the other hand, we found that the tilt mode primarily influences the tilt angle, having minimal influence in the rest of structural parameters.

Finally, our analysis in this section aligns with Modesto's previous study [39], which suggests that the DNA benefits from its essential modes to adopt conformations that aid in DNA-protein recognition. Because modes can exhibit opposite behaviours, we hypothesize DNA binding enzymes might benefit from the couplings intrinsic to the essential modes, and may deform the DNA along specific modes to facilitate their binding. The analysis presented in this section further increases our understanding in the relationship between the DNA essential movements and DNA elasticity, including its elastic couplings.

5.5 CONCLUSION

In this chapter, we performed an in depth analysis of the terms that compose the elastic matrix F, and we have found that there are three non-zero off-diagonal components, which correspond to the twist-stretch (D), twist-roll (G) and stretch-roll (H) couplings, in agreement with the literature [74]. We found that these three couplings are length-dependent and oscillate as a function of length. In the case of the couplings related with the tilt angle, our results indicate that they can be neglected at the global level as they cancel out at the local level, in agreement with the MS model [98]. Analysing the six couplings along the sequence, we find that all couplings present sequence-dependent features.

We then mathematically described the elastic profiles of the 7 relevant elements of matrix F $(A_\rho, A_\tau, C, B, G, H, D)$. By performing curve fittings, we parameterised their functions, where we obtained an accuracy of around 90%. Our equations successfully describe G and H as damped waves, exhibiting a periodicity of two DNA turns, and amplitudes around 13 nm and 54.9 nm, respectively. Our predictions of the twist-roll coupling G, are about half the current accepted value for the twist-bend coupling [111, 141]. Interestingly, there are no reported calculations for H in the literature. Furthermore, the periodic nature of these two couplings has not been reported as well. Regarding twist-stretch coupling D, our model accurately captures its profile at lengths longer than 1 DNA turn as a negative exponential that tends to a plateau around -3.78nm, which agrees with previous experimental studies [48], [86], [49], [134]. Putting all these observations together, we identified critical lengths that maxmize and minimize the correlation between structural variables. More specifically, these ranges are defined as $\left[\frac{2}{4}, \frac{3}{4}, \frac{4}{4}, \frac{5}{4}, \frac{7}{4}, ..., \frac{(2n+1)}{4}\right]$ DNA turns, with n as an integer value. These key lengths could have important implications in biological processes where proteins may exploit the flexibility of DNA. For instance, the GCN4 transcription factor bends and slightly unwinds the DNA at approximately half a DNA turn [37, 72]. Similarly, the IHF protein bends the DNA within a distance of one helical turn [1], while the 434 repressor binds and overwinds the DNA around 1.25 turns [129] (see figure 1.3).

A flaw in our model is that it fails to describe the coupling profiles at the local level (lengths less than 1 DNA turn), which are characterised by sudden behavioural changes. The couplings G and H exhibit higher amplitudes, and D has a complex shape, where it transitions from an oscillatory behaviour to a negative exponential. We believe that the presented model could aid future studies in the predictions of cross-terms as well as the development of coarse grained models, and further analysis is required to address sequence-dependent effects.

To investigate the origin of the flexibility of DNA, we implemented PCA to associate essential modes with the elastic variables at different length scales. In general we found that there are four essential modes that cause most of the fluctuations in the structural parameters used for calculating the elastic matrix F. We denote these modes according to the flexibility they influence the most, being roll, tilt, twist and stretch essential modes. The essential mode classification reveals that there is a transition between local and bulk flexibility, where at the local level of less than 1 DNA turn, the roll mode is responsible for most of the stretching deformations, while at lengths of around 1.5 DNA turns, an essential mode arises and becomes the principal contributor to the stretching flexibility. In contrast to the stretching mode, the variables of roll, tilt and twist are assigned to independent modes at lengths higher than half helical turn. Our results indicate that the transition between local and bulk flexibility is mainly originated by this stretching mode, which is capable of radically changing the behaviour observed in the elastic profiles, specially in the case of the couplings. Interestingly, we found that the essential modes associated with the elastic variables are similar between sequences. Our classification method indicates that these associated modes are directly related to the elastic couplings. Lastly, it is worth to point out that, to the best of our knowledge, no other studies have tried to establish a connection between structural deformations and essential movements.

We then proceeded to analyse the system variance explained by the four associated modes. We found that the modes collectively account for approximately 65% of the system variance, with roll and tilt being the most significant contributors at around 30% each, while twist and stretch together account for less than 20%. While our classification method aims to relate each elastic variable with an essential mode, it is important to acknowledge a limitation of our approach: the relatively low amount of captured system variance. This limitation can be addressed in future research by expanding the method to capture a higher number of modes, aiming to collectively account for at least 90% of the system variance.

We then analysed the variance profiles calculated from the associated modes and found that they are able to reproduce the shapes of the original variances, although, they underestimate them in magnitude as multiple essential modes are filtered out. This is the reason our essential mode analysis is focused on the covariance matrix rather than the elastic matrix F. Furthermore, it is worth pointing out that our essential modes classification only associates modes with the elastic variable that yields the highest variance/covariance, however, there are multiple modes that can affect the flexibility and their study is beyond the scope of this project.

Analysing the covariance profiles, we found that the flexibility of DNA is better recovered when combining the associated modes, as expected, as they together allow the DNA to explore a wider variety of conformations. Interestingly, our analyses indicates that the twist and roll modes originate two oscillatory twist-roll couplings with opposite signs. Similarly, twist and stretch modes originate two opposite twist-stretch couplings. This adds an additional layer of complexity in the dynamics of DNA, where depending on the mode of deformation, the DNA could exhibit opposite mechanical responses. These findings may have important implications in understanding bio-molecular processes such as DNA-protein recognition, where it has been found that in this complexes, the DNA is deformed along its essential modes [39].

We then analysed the structural deformations the associated modes can induce. In general, we found that the roll, twist and stretch modes directly impact the roll and twist angles as well as the end-to-end distance. In contrast, the tilt mode is isolated and principally affects the tilt angle. Moreover, our analysis reveals a linear relationship between structural deformations and projections of the essential modes. These findings are again relevant for DNA-protein interactions, as these deformations are within the length scales in which a variety of proteins interact with the DNA, such as nucleoid associated proteins [1], transcription factors [37, 72] and repressors [129].

We then tested the capability of the essential modes to adopt conformations of complex deformed DNA. To this end, we projected the associated modes to a supercoiled DNA minicircle with -2 turns and a nucleosomal DNA. Similar to previous studies [39], our results indicated that the DNA uses the essential modes to adopt complex conformations, although the associated modes can hardly adopt conformations in which the DNA is highly bent. We believe that considering more modes may provide better descriptions as the DNA would be able to reach more regions in the conformational space. Not less importantly, sequence effects might also be important as different sequences would allow the DNA to also explore different regions of the conformational space. However, this study does not perform any emphasis on sequence effects and future work is required in this matter.

By analysing individual projections, we observed twist waves in the twist angle when projecting the roll and tilt modes, which agree with twist waves observed in looped DNA [110] and in circular DNA [143]. Similarly, we also observe an analogous behaviour regarding the end-to-end distance, where the structural parameter oscillates when projecting the roll, twist and stretch modes. Similar to the twist waves, we call this behaviour 'stretch waves'. We propose these stretch waves arise as a response of bending stress and are introduced via the stretch-roll and twist-roll couplings. The twist and stretch waves mechanisms might have biological relevance as they could facilitate protein binding.

Overall, our investigations have significantly contributed to our understanding of DNA elastic couplings and their intricate relationship with the DNA essential dynamics. The new mathematical framework we have developed allows for a comprehensive estimation and characterization of the DNA elastic couplings. Through our essential dynamics analysis, we have gained valuable insights into the complex interplay between DNA flexibility and its essential movements. A notable discovery from our research is the revelation that the transition from local to bulk flexibility is primarily driven by a stretching essential movement. Furthermore, our analysis has revealed that the DNA essential movements can exhibit opposite couplings, which holds significant implications for DNA-protein interactions, as the DNA could exhibit opposing mechanical responses depending on the mode of deformation induced by interacting proteins. Another important outcome of our investigations is the observation that the essential movements of DNA are relatively similar across most sequences. Lastly, our results indicate that the DNA is naturally deformed along its essential movements to approximate DNA-protein conformations, in agreement with previous studies.

Chapter 6

Conclusions

In this work, we presented the SerraLINE and SerraNA open softwares, which have been proved to be extremely useful in bridging the gap between experimental and computational studies, by yielding global parameters that allow the comparison between both approaches, while providing local parameters that allow the analysis of NA in detail. Both programs can provide structural and elastic parameters at different length-scales, from the dinucleotide level to the length of the whole molecule. Furthermore, we have validated both programs by analysing a wide range of atomistic MD simulations of DNA under different conditions and comparing our results with existing literature and experimental data.

We used SerraLINE in combination with AFM experiments to study how supercoiling affects the structure of double-stranded DNA minicircles. This multi approach allowed us to discover that, at the critical length of 16 bp ($\sim$ 5.3 nm), bending angles exceeding 75° can induce DNA defects. The onset of these defects is at the superhelical density of $\sigma \sim [-0.03, -0.06]$, which is within the ranges of superhelical levels found in DNA in vivo (approximately -0.06) [57]. Consequently, we found that these DNA defects serve as flexible hinges that allow the DNA to be highly bent, which in turn relaxes the imposed helical stress and causes a reduction of its aspect ratio. Our results indicate that DNA defects are frequently found in nature, and are relevant in DNA recognition processes, where the bending flexibility of DNA allows to reduce the distance between enhancer and promoter [88]. Additionally, our results indicate that DNA defects may be the underlying mechanism for reducing the DNA molecule size in biological processes such as DNA packaging.

SerraNA allowed us to observe how the bulk elastic properties of DNA arise from local bp fluctuations, for a variety of systems. In agreement with previous studies [114], we observed that the transition from local to bulk flexibility occurs at the length of one helical turn ($\sim$10.5 bp). From simulations of naked DNA, we estimated a stretch modulus around 1778 pN, a twist modulus of 97 nm and a persistence length of 57 nm, which agree with measurements from single-molecule experiments [9, 13, 50, 56, 85, 103, 107, 147, 165]. For DNA-protein complexes and sequence mismatches, even though some cases do not fall within the harmonic approximation, SerraNA was able to obtain valuable insights, indicating that SerraNA can still be used to study trajectories in which the DNA is heavily deformed due to the interaction with proteins. Our results indicate DNA-protein complexes tend to be more rigid as the DNA is restrained into particular conformations, which might be a mechanism for protein binding; on the other hand, sequence mismatches could be detected by cellular machinery as they exhibit higher flexibility.

We then analysed the set of unique 136 tetranucleotide sequences from the ABC simulation database [119] at the tetranucleotide level for the first time. Our findings indicate that flexibility is also strongly sequence-dependent at the tetranucleotide level, as some sequences are twice as rigid as others, where in general RRYY and RRRY sequences are the most rigid, while YRYR sequences are the most flexible. Remarkably, we observed that sequences containing central AT and AA base-steps are considerably more rigid than sequences with central TA (see figure 4.9). This finding suggests that AT-tracts can exhibit extreme mechanical properties when the sub-sequences are properly phased. These observations show that there is a complex correlation between dinucleotide steps [4,5], and that the interplay between sequence-dependent and length-dependent features is highly important and complex.

To further explore the mechanical behaviour of DNA and particularly the previously unexplored area of DNA elastic couplings, we utilized SerraNA to obtain for the first time length-dependent profiles of the elastic couplings. In agreement with the MS model [98], our results indicate that the twisting and bending deformations are coupled via roll, introduced by the anisotropy of DNA. The twist-roll (G) and stretch-roll (H) couplings exhibited oscillatory behavior as a function of length, resulted from the calculation of mid-step triads at different length-scales. Nonetheless, we mathematically described these couplings as damped oscillations with a period of two helical turns and an amplitude of 13 nm for twist-roll and 55 nm for stretch-roll. The amplitude of twist-roll is about half of the values reported in the literature [111, 141], while there are no estimations of stretch-roll. Furthermore, we found that the twist-stretch coupling exhibit oscillatory behaviour for lengths less than 1 DNA turn, where it then tends to a plateau around -3.78 nm, which qualitatively agrees with previous experimental evidence [48], [86], [49], [134]. Lastly, from our detailed analysis we found that couplings related with the tilt bending component were approximately zero at every length-scale, indicating that the DNA bending flexibility is coupled to twist and stretch via the roll angle. The mathematical functions we proposed for describing the couplings allowed us to estimate parameters that can be used to characterise the DNA elastic couplings. The proceedings for extracting these parameters could be employed as a new methodology for estimating the DNA global elastic couplings. Additionally, these functions could also aid in the development of future coarse-grained models of DNA that take into account its elastic couplings and interactions beyond the nearest-neighbour approximation.

Our comprehensive analysis of the elastic coupling profiles, allowed us to identify critical lengths in which the elastic variables are most correlated or decoupled. The interplay between these elastic variables might be relevant for biological processes, as interacting proteins could deform the DNA at key lengths in order to exploit the length-dependent correlations between the elastic variables. Complexes where the DNA is bound to proteins such as IHF [1], the GCN4 transcription factor [37, 72] and the 434 repressor [129] corroborate our observations, where deformations of the double helix present different behaviour depending on the length and deformation mechanism.

Lastly, similar to the diagonal elastic constants, we observed that the crossover from local to bulk flexibility occurs around one DNA turn, where the behaviour of the elastic couplings in these two regimes are completely different.

To investigate the origin of DNA flexibility, we combined PCA with SerraNA to associate one particular essential mode to each type of deformation at different length-scales. We observed that the modes associated with bending deformations (roll and tilt), are responsible for most of the variance in the simulations, while modes associated with twist and stretch primarily affect the base-pairs at the ends of the DNA fragments. Interestingly, we found that DNA fragments with different sequences exhibited similar essential modes, indicating a common mechanism governing their flexibility. Our results reflected previous observations [114], where we found that DNA transitions from local to bulk flexibility at approximately 14 bp (around 1.5 helical turns). This transition is driven by an emerging mode which our classification process associates to stretching deformations and is capable of radically changing the interplay between elastic constants. While the modes associated with roll, twist and stretch primarily influence the structural deformations they were associated with, our analysis of the accessible conformational space revealed their significant impact on the elastic couplings, including the ability to exhibit couplings with opposite signs. In contrast, the tilt mode remained isolated, capturing only deformations of the tilt angle. These findings suggest that the DNA could exhibit distinct mechanistic responses when deformed along specific essential movements, which may be exploited by DNA-binding proteins in various biological processes. Furthermore, we examined this possibility by testing the capacity of the associated modes to adopt complex structures in which the DNA molecule is highly deformed, such as superoiled DNA and nucleosomal DNA. Remarkably, the associated modes were able to approximately adopt the shapes of these deformed DNA structures, with the exception of regions where the DNA was severely bent. In contrast to previous studies [163], it is important to note that our method utilised four essential modes from test DNA sequences that differed from the target trajectories (e.g., deformed DNA), which provides evidence that the DNA essential modes are generally similar across fragments with different sequences. Nevertheless, future improvements could involve considering a greater number of modes and emphasizing the influence of sequence effects to enhance the accuracy of our predictions. Lastly, analogous to the twist waves [110, 143], we discovered the existence of stretch waves induced by the modes associated with roll, twist, and stretch deformations. Similar to the twist waves, the stretch waves are introduced through the couplings via the roll component, and might have biological relevance in DNA-protein interactions.

Overall, our investigation has contributed to the understanding of structural and elastic properties of DNA at various length scales. The development of SerraLINE and SerraNA facilitates the integration of experimental and computational studies, offering valuable insights into the mechanical behavior of DNA. By investigating the effects of supercoiling, analysing unique tetranucleotide sequences, analysing flexibility of DNA-protein complexes, exploring the elastic couplings, and associating deformations with essential dynamics to explore the origin of the DNA flexibility, we have expanded our knowledge of DNA's flexibility and its relevance in biological processes. Our findings pave the way for further research in the field of DNA structure and flexibility, contributing to the scientific community and promoting advancements in this important and particular biophysical area of study.

6.1 Future work

SerraNA and SerraLINE have proved to be particularly useful in investigating these properties, specially when combined with experimental approaches, and have been expanded to provide parameters that further increase their compatibility with experiments. On one hand, there are a variety of new features that could be implemented in both programs in order to expand their suitability with new experimental setups. On the other hand, the programs are currently written in Fortran, but they could be translated to Python in order to further expand its functionality by combining it with multiple libraries, including pytraj [131] (Python implementation of cpptraj), which is useful for processing and analysing MD trajectories. This would be beneficial for analysing a wider range of more complex systems such as damaged DNA or hybrid structures; however, this expansion would have an impact in the program's performance. Lastly, both programs have started to draw some attention from the scientific community, and we hope they will further aid in increasing our knowledge about DNA and its mechanical properties.

The flexibility of a handful of systems were analysed in this work, with no particular emphasis in sequence-dependent features, however, SerraNA could be implemented to analyse how the elastic profiles behave on systems with characteristic sequences such as A-tracts. These profiles might exhibit drastic changes in their behaviour, and might provide further insight about the role of sequence and length-scale in building bulk flexibility.

Regarding the elastic couplings, there is a lack of methods that estimate their values. Our proposed models of chapter 5, could be of further aid in the design of both computational and experimental strategies to estimate their quantities. Furthermore, in our essential modes analysis, we filtered multiple modes that considerably contribute to the system variance, however, it might be worth analysing these essential modes to determine which aspects of DNA flexibility they contribute. It would also be interesting to analyse simulations in which DNA interacts with proteins such as IHF, the GCN4 transcription factor or the 434 repressor, to further corroborate our observations in which the DNA couplings exhibit different mechanical properties depending on the length-scale of deformations. Additionally, PCA could be implemented to simulations of naked DNA with these same sequences, in order to associate the essential modes to the four elastic variables. Then, same as the analysis performed in chapter 5, the calculated modes could be projected to the simulations with bound proteins, to finally demonstrate that the associated essential modes originate the flexibility of DNA and proteins deform DNA along them.

Appendix A

DNA sequences and molecular dynamics conditions

The simulations that are shown in this thesis project, were not produced by me. However, here we provide some key information regarding their simulation procedures and conditions.

A.1 MD simulations of DNA minicircles

This set consists of 10 circular DNA simulations conformed by a 260 bp DNA extracted from [115] and a 339 bp DNA with same sequence as the experimental minicircles used in [125]. The set of simulations is conformed by these two circular molecules with distinct levels of negative supercoilings. For the 260 bp molecules, systems were built with 0 (relaxed), -1, and -2 turns undertwisted, while for the 339 bp systems, 0 (relaxed), -1, -2, -3 and -6 turns were removed. Additionally, two replicas were created for the 339 bp minicircle with -2 and -3 turns. The simulation protocols involve a combination of implicit and explicit solvent techniques. The AMBER99 forcefield [61] with parmBSC0 [122], parmOL4 [69] and parmOL1 [175] corrections were used to describe the DNAs. The starting 20 ns of simulation time were produced following the protocol [17] with the SANDER module within AMBER12 in implicit solvent using the Generalised Born/Solvent Accessible area method [156] at 200mM salt concentration and 300 K. Then, representative structures were selected and solvated in TIP3P rectangular boxes. Ca^{2+} counterions were added to neutralize the charges and additional $Ca^{2+}/2Cl^{-}$ ion pairs were added to achieve a 100mM concentration. Simulations were then ran for 100 ns using the CUDA version of AMBER16 [16]. The WrLINE program [150] was then used to calculate the molecular contour of the circular simulations, where only the last 30 ns were used for analysis. These simulations were produced by Dr Agnes Noy, and more detailed information regarding the simulation and experimental protocols can be found in [125].

A.1.1 260 bp DNA minicircle

TCTCTCTCTC TCTCTCTTAA AGGTATACAA GAAAGTTTGT TGGTCTTTTT
ACCTTCCCGT TTCGCTCCAA GTTAGTATAA AAAAGCTGAA CGAGGAAACG
TAAAATGATA TAAATATCAA TATATTAAAT TAGGATTTTG CATAAAAAAC
AGACTACATA ATACCTGTAA AACACAACAT ATGGCAGTCA CTATGAATCA

ACTAACTTAG ATGGTATTAG TGACCTGTAA CAGAGCCGAG GGCGATATCG CAGGAGTCCG

A.1.2 339 bp DNA minicircle

TTTATACTAA CTTGAGCGAA ACGGGAAGGG TTTTCACCGA TATCACCGAA ACGCGCGAGG CAGCTGTATG GCGAAATGAA AGAACAAACT TTCTTGTACG CGGTGGTGAG AGAGAGAGAG AGATACGACT ACTATCAGCC GGAAGCCTAT GTACCGAGTT CCGACACTTT CATTGAGAAA GATGCCTCAG CTCTGTTACA GGTCACTAAT ACCATCTAAG TAGTTGATTC ATAGTGACTG CATATGTTGT GTTTTACAGT ATTATGTAGT CTGTTTTTTA TGCAAAATCT AATTTAATAT ATTGATATTT ATATCATTTT ACGTTTCTCG TTCAGCTTT

A.2 MD simulations of linear DNA

This set of MD simulations were produced by Dr Agnes Noy. The 32mer, 42mer, 52mer and 62mer DNA duplexes were extracted from sequences that consist of 170-200 bp, and are respectively named $\gamma 3$, $\gamma 1$, $\gamma 4$ and $\gamma 2$ in [104], and NoSeq, CA, TATA and CAG in [160]. These simulations were produced with AMBER16 suite [16] using the AMBER parm99 [21] forcefield with parmbsc0 and parmbsc1 corrections [63,122]. The 32mer sequence was also ran using parmOL15 forcefield [173,174] and named as 32ol15. All structures were ran in explicit solvent with 200 mM Na^+ and Cl^- counter-ions [146] and in TIP3P octahedral boxes [65], where productive MD simulation was ran for 1 μs.

A.2.1 32mer and 32ol15 sequence

CGACTATCGC ATCCCGCTTAGCTATACCTA CG

A.2.2 42mer sequence

CGCATGCATA CACACATACA TACACATACT AACACATACA CG

A.2.3 52mer and 52s sequence

CGTATGAACG TCTATAAACGTCTATAAACG CCTATAAACG CCTATAAACG CG

A.2.4 62mer sequence

GCAGCAGCAC TAACGACAGC AGCAGCAGTA GCAGTAATAG AAGCAGCAGC AGCAGCAGTA GC

A.3 MD simulation of DNA pulling

The simulation named 52s is composed by a 52 bp-long DNA with same sequence as the 52mer, and was produced by Dr Jack Shepherd. It was ran in explicit solvent and stretched by a series of umbrella sampling simulations following the protocols stated in [135]. The length of the molecule was increased by a series of stretching steps each consisting of 1 Åof extension and 1ns of simulation time, resulting in a 8ns trajectory with a total extension of 8 Åfrom the relaxed structure. The maximum relative extension of the pulled DNA is of approximately 5%, where in all the stretching steps the DNA conserved all the hydrogen bonds and stacking interactions.

A.4 MD simulations from the BIGNASim database

The following simulations were obtained from the BIGNASim database[*] [59]. Note that for all the following simulations, monovalent ions were used to neutralize the system and the bsc1 forcefield was used [63]

A.4.1 32rand

32 bp-long DNA with random sequence ran for 1 μs with the TIP3P water model.

ATGGATCCAT AGACCAGAAC ATGATGTTCT CA

A.4.2 Nucleosome 1kx5

147 bp-long nucleosome with PDB ID 1kx5, consisting of 500 ns of simulation time using the TIP3P water model and bsc1 forcefield.

ATCAATATCCACCTGCAGATACTACCAAAAGTGTATTTGGAAACTGCT
CCATCAAAAGGCATGTTCAGCTGGAATCCAGCTGAACATGCCTTTTGATG
GAGCAGTTTCCAAATACACTTTTGGTAGTATCTGCAGGTGGATATTGAT

A.4.3 Transcription factor 2dgc

A protein-DNA complex with PDB ID 2dgc and composed of a 18 bp-long DNA bound to the transcription factor GCN4, consisting of a 500 ns of simulation using the SPCE water model.

GGAGATGACGTCATCTCC

A.4.4 A:A mismatch

500 ns simulation using TIP3P waters of a 13 bp-long DNA with an A:A mismatch at the middle of the sequence denoted as A.

CCATACAATACGG

A.4.5 G:G mismatch

A DNA with same sequence than A:A but with a G:G mismatch instead, which similarly was run for 500 ns and using the TIP3P water model.

[*] `https://mmb.irbbarcelona.org/BIGNASim/`

CCATACAATACGG

A.5 MD simulations with the tetranucleotide sequences from the ABC database

The ABC consortium created a simulation database[†] [119] composed by 39 oligomers that together contain all the distinct 136 tetranucleotide sequences. Each oligomer is made of 18 bp and is simulated for 1 μs using the parmbsc0 forcefield [122] and SPC/E water model [11] with ion concentration 150 mM K^+Cl^- [24].

Each oligomer is built with a repeating sequence 'GC-CD-ABCD-ABCD-ABCD-GC' where 'ABCD' denotes the tetranucleotide sequence. More information can be found on their web-page located in the footnote of this page.

[†] `https://bisi.ibcp.fr/ABC/Sequences.html`

Appendix B

SerraLINE supportive information

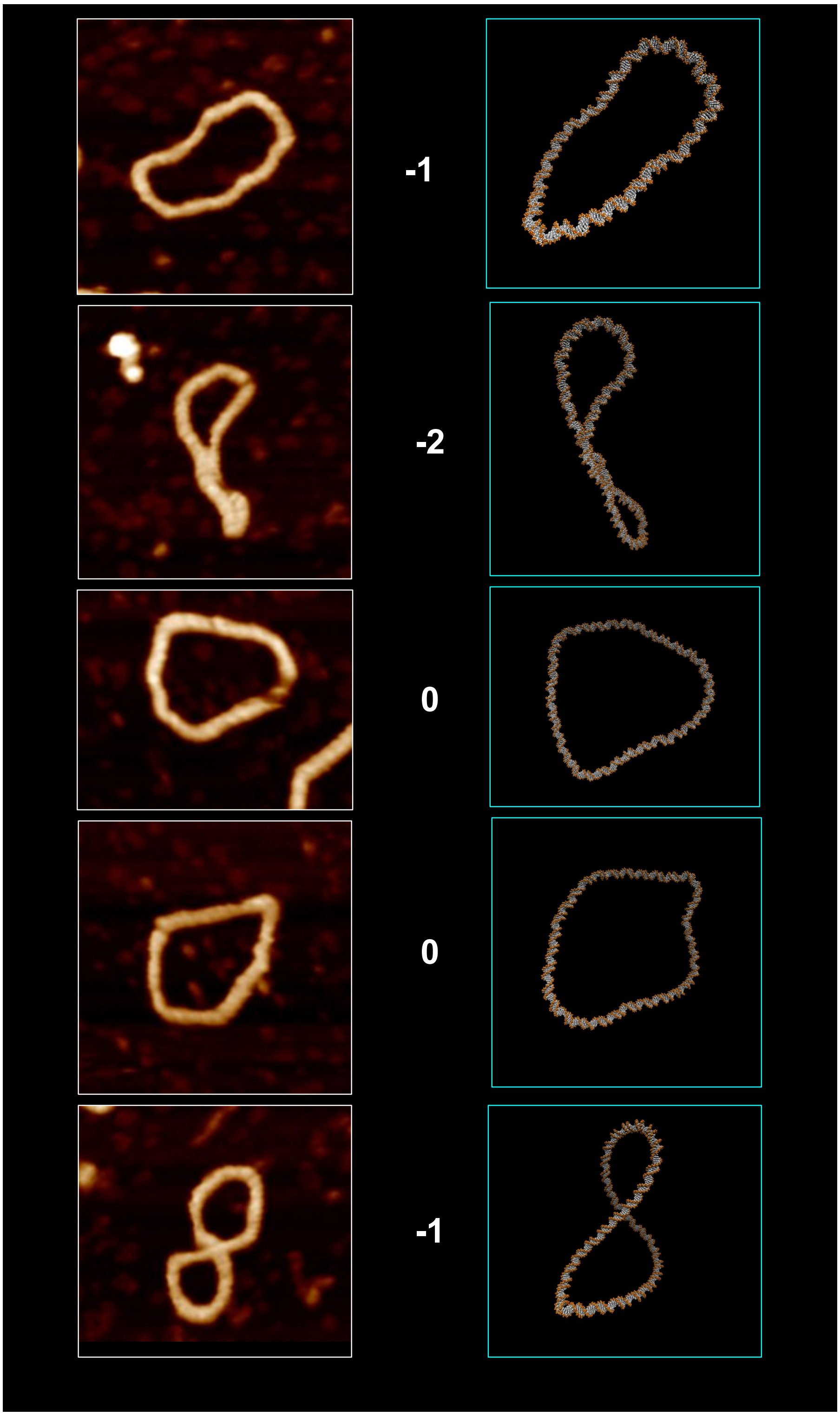


Figure B.1: (Left column) High-resolution AFM images of supercoiled DNA minicircles. (Right column) Snapshots of atomistic MD simulations of supercoiled DNA minicircles. Number between columns indicate the number of DNA turns removed on both experimental and simulation approaches. AFM images are taken from [125]. See the appendix section A.1 for more information regarding simulation conditions.

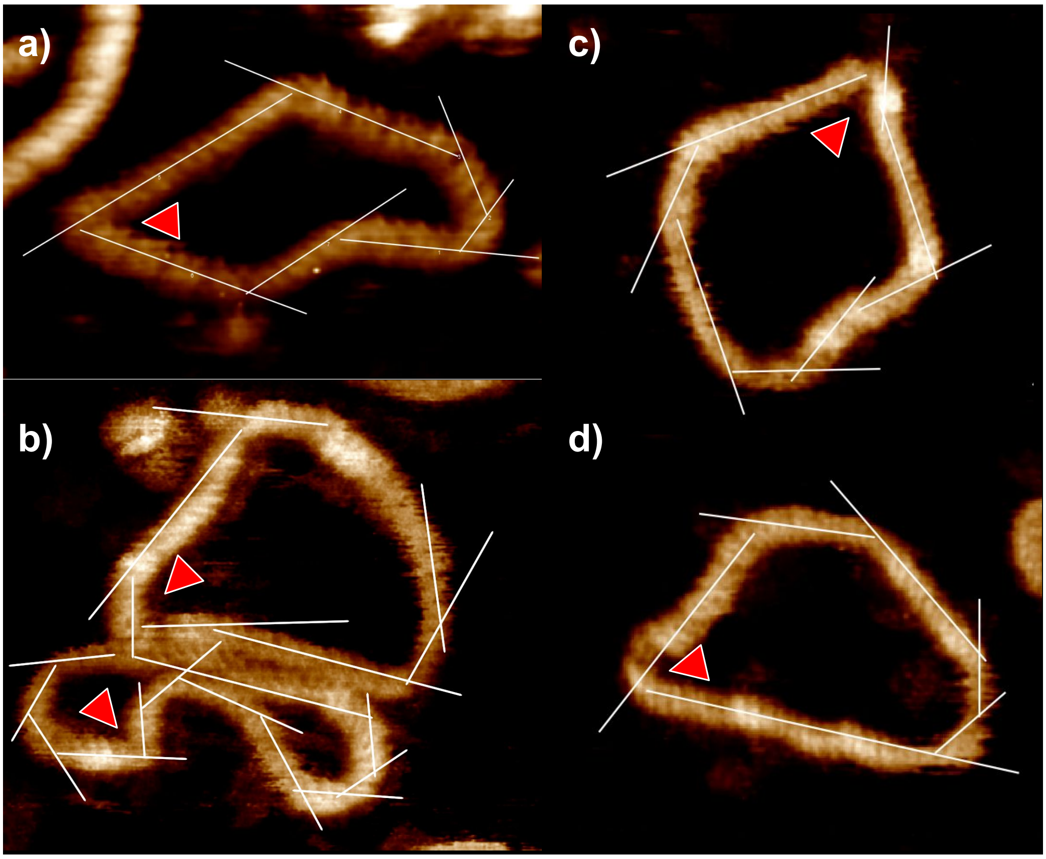


Figure B.2: Analysis of bend angles by high-resolution AFM images of 251 (a) and 339 (b,c,d) bp minicircles. Red triangles indicate kinked regions. Bending angles were measured of 51 ± 14° and 106 ± 16° on average for bent and kinked regions respectively. Data and images from high-resolution AFM were taken from [125].

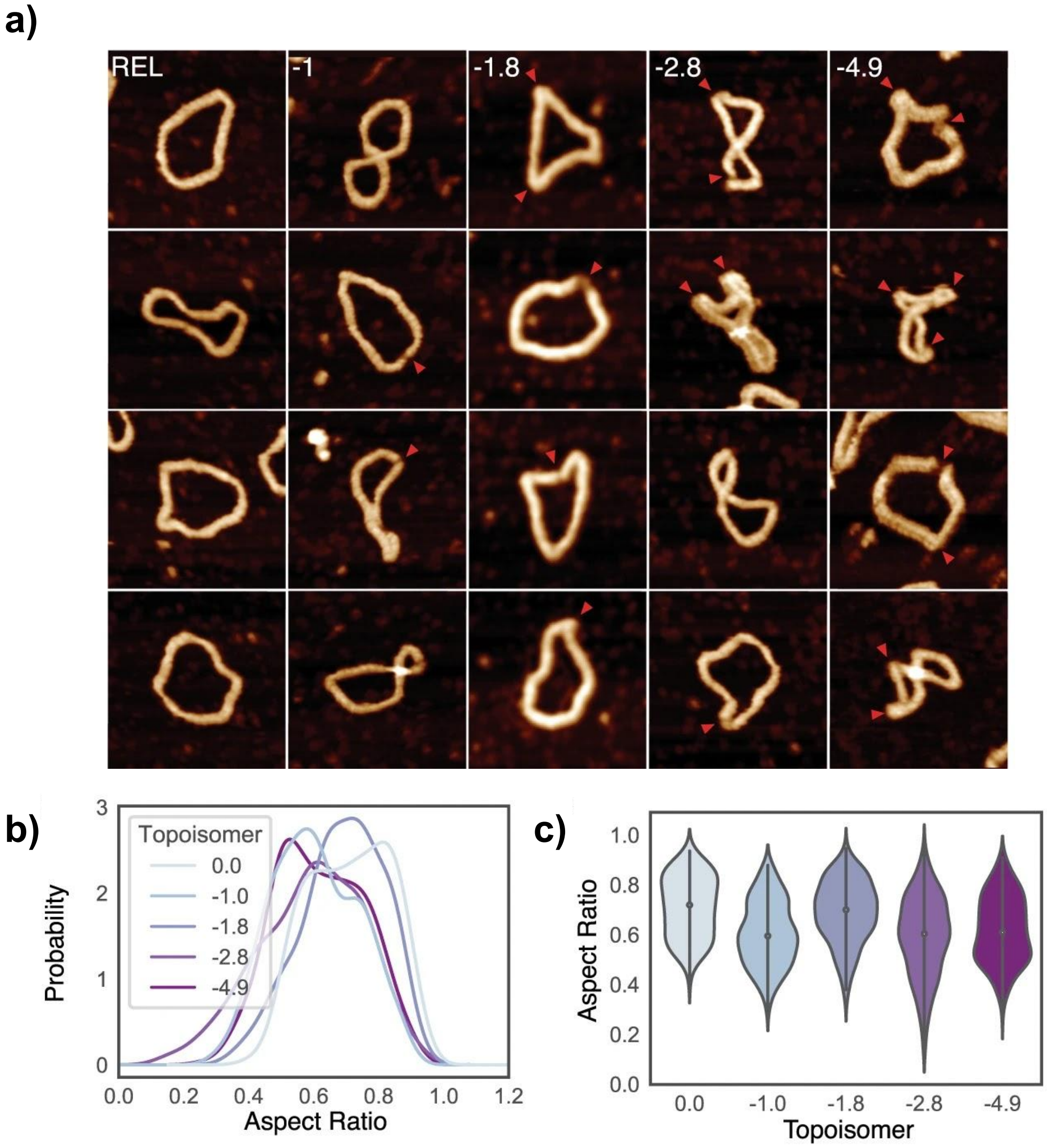


Figure B.3: a) Visualization of representative images of 339 bp minicircles where global compaction is increased as a response to negative supercoiling. White letters indicate the average number of turns removed in a particular column where the REL column indicates relaxed minicircle and red triangles indicate defects formation. b) Kernel Density Estimate (KDE) plots of the probability distribution of aspect ratios for each toposoimer (N = 1375). c) Relationship between aspect ratios and level of negative supercoiling represented as violin plots. Data, images and results were taken from [125].

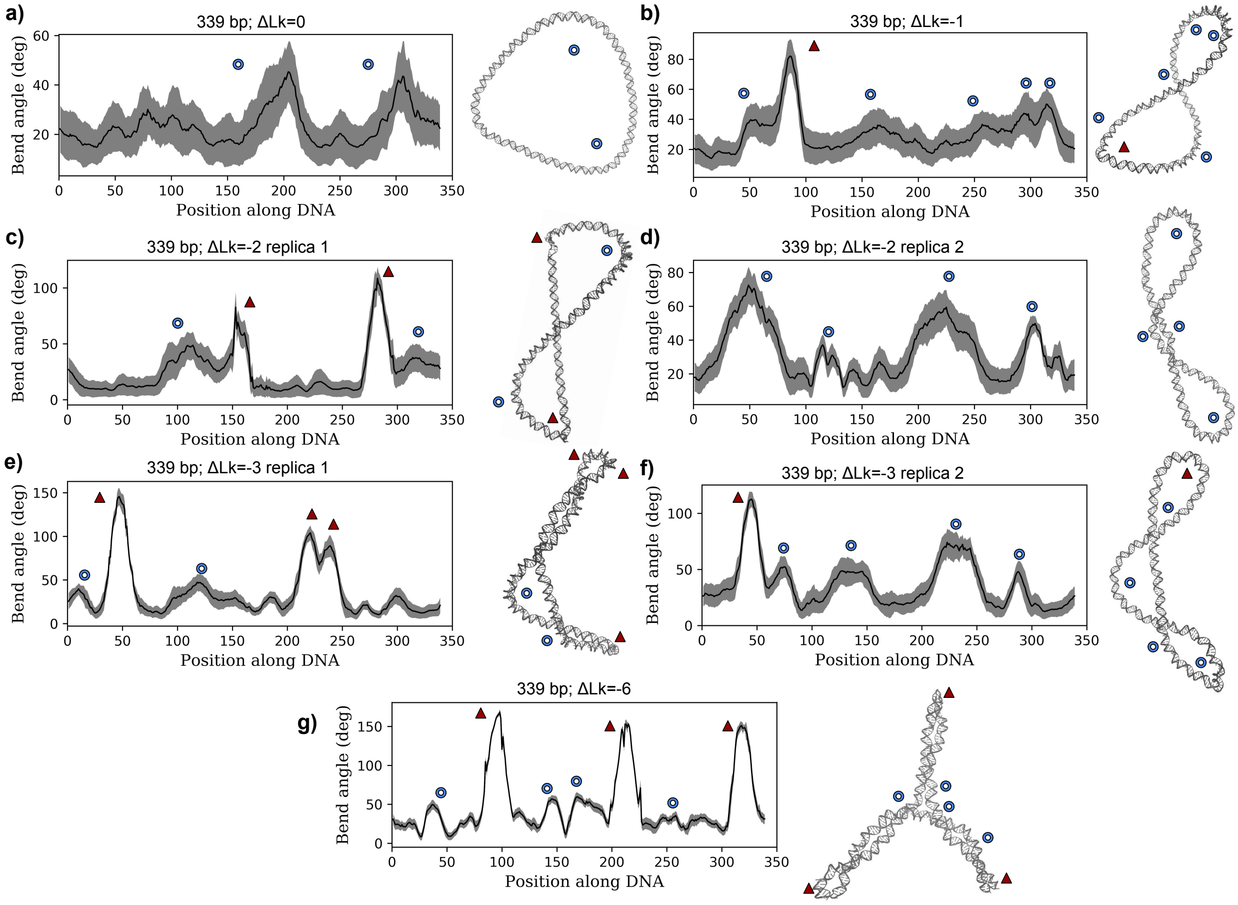


Figure B.4: Bending angle profiles at the length of 16 bp with their corresponding supercoiled structures for the 339 minicircles with diverse liking differences (ΔLk). Blue circles indicate bent regions classified as B-DNA and red triangles as bent regions classified as defects. These classifications correspond to the ones shown in figure 3.8. Shaded areas represent standard deviations.

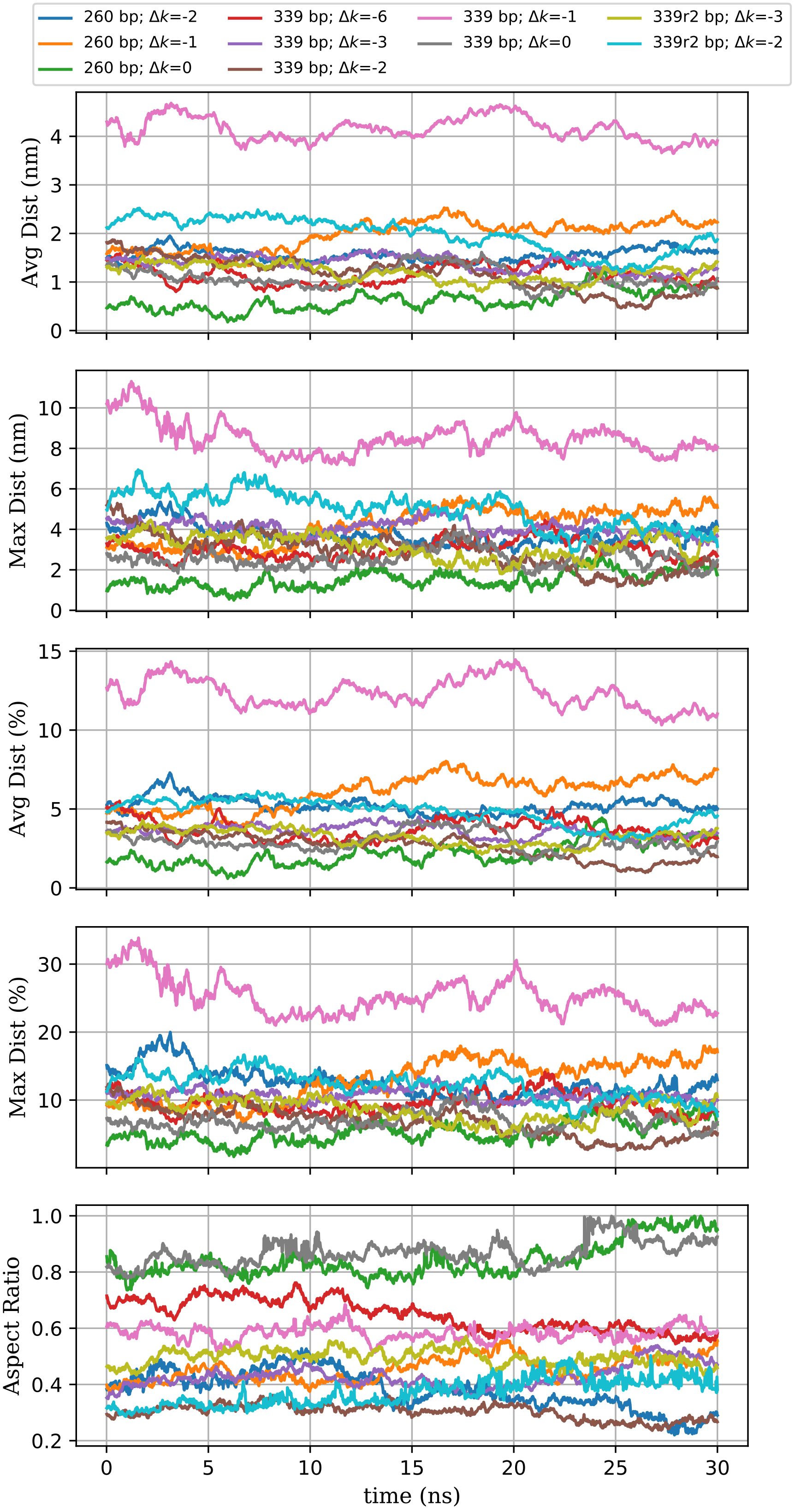


Figure B.5: Time-series of the last 30 ns of simulation time for the deviation from planarity parameters and the aspect ratio of all minicircles. These time-series were used to calculate the distributions shown in figure 3.9.

Appendix C

SerraNA supportive information

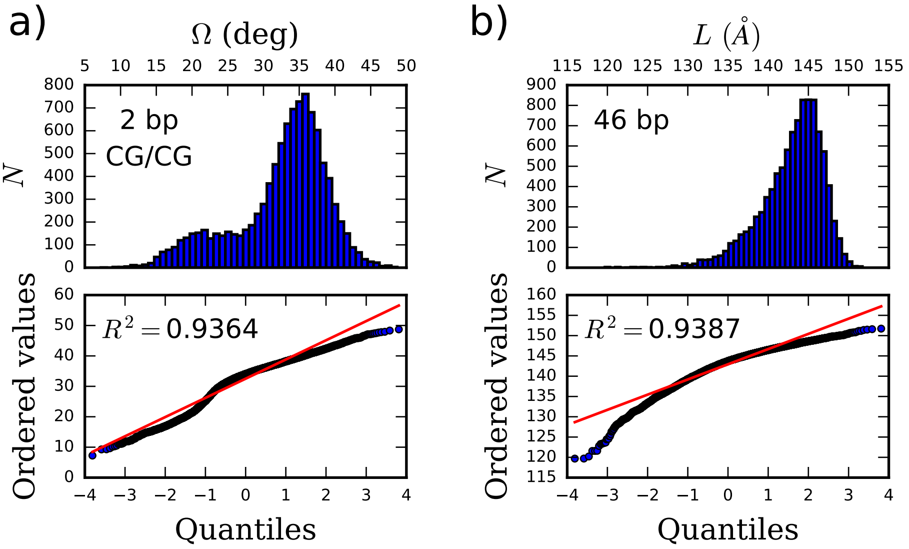


Figure C.1: Histograms and Q-Q plots for the two less Gaussian cases in the set of linear DNA. (a) Bimodal twist at the CG bp step from the 32mer fragment and (b) asymmetrical distribution of the end-to-end distance distribution at the longest possible oligomer length (46 bp) within the 62mer fragment.

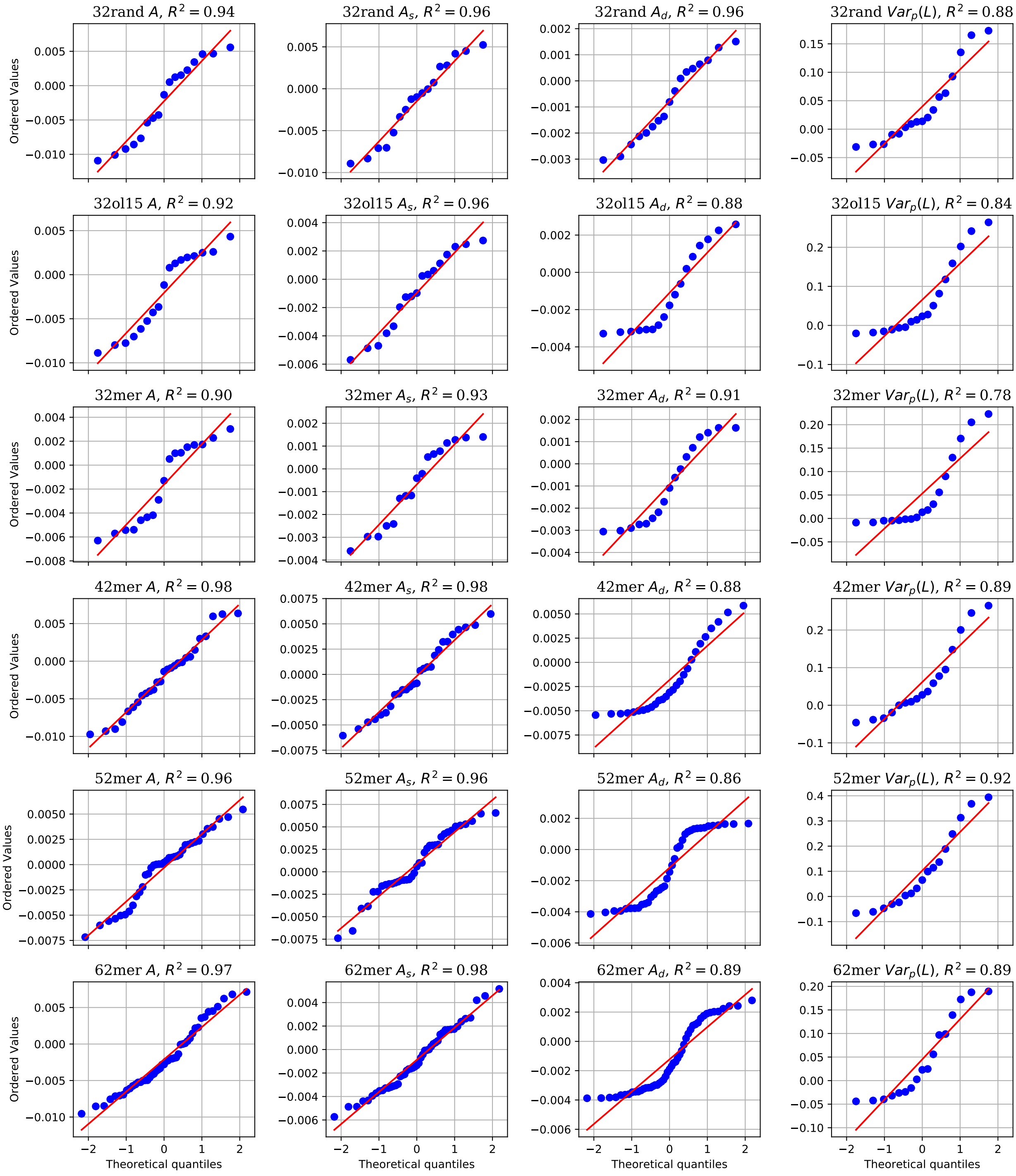


Figure C.2: Q-Q plots of the residuals obtained from the linear fittings presented in figure 4.5 are shown for the set of linear free DNA simulations. The fittings were applied to the elastic constants of persistence length A, static persistence length A_s, dynamic persistence length A_d, and the partial variance of the end-to-end distance $V6ar_p(L)$, which was used to calculate the stretch modulus B. The residuals were obtained by applying the SerraNA default Analysis method as described in subsection 4.1.2.

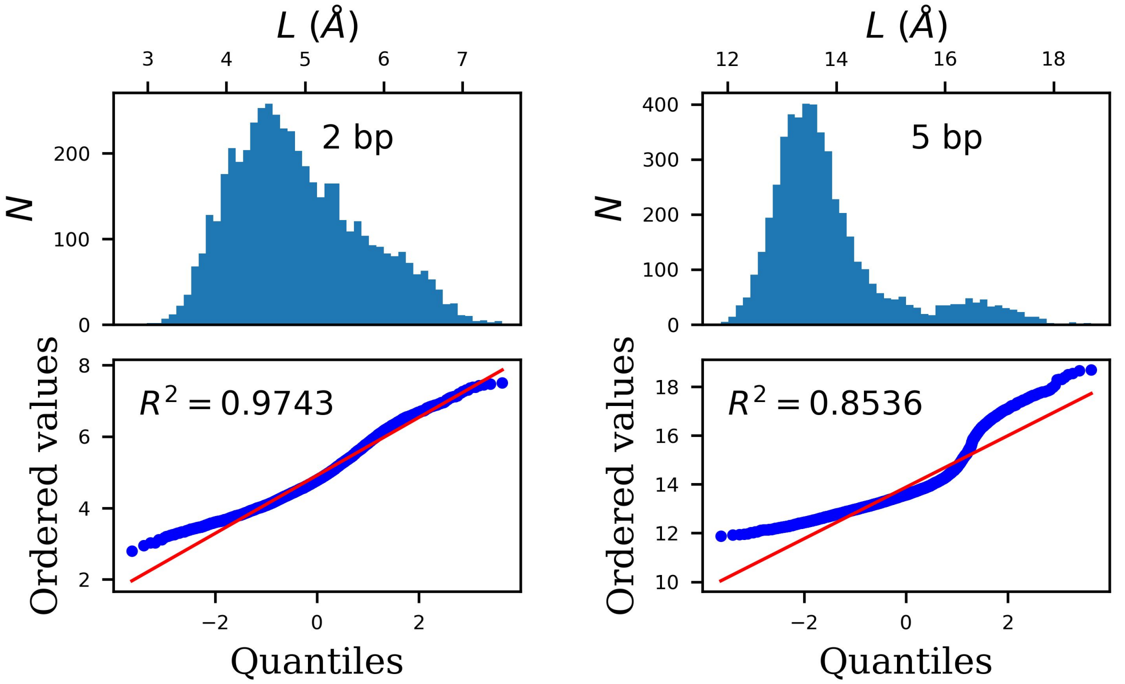


Figure C.3: Histograms and Q-Q plots of the end-to-end distance L for the GG mismatch trajectory. (a) Slightly asymmetric distribution for the bp-step located at the mismatch. In this case, the Gaussianity test is passed as the R^2 value is higher than 0.90. (b) Skewed bimodal distribution at t6he length of $l = 4$, with the GG mismatch located at the middle of the corresponding sub-fragment. This case has the slowest R^2 for the simulations analysed in the SerraNA chapter (see figure 4.2).

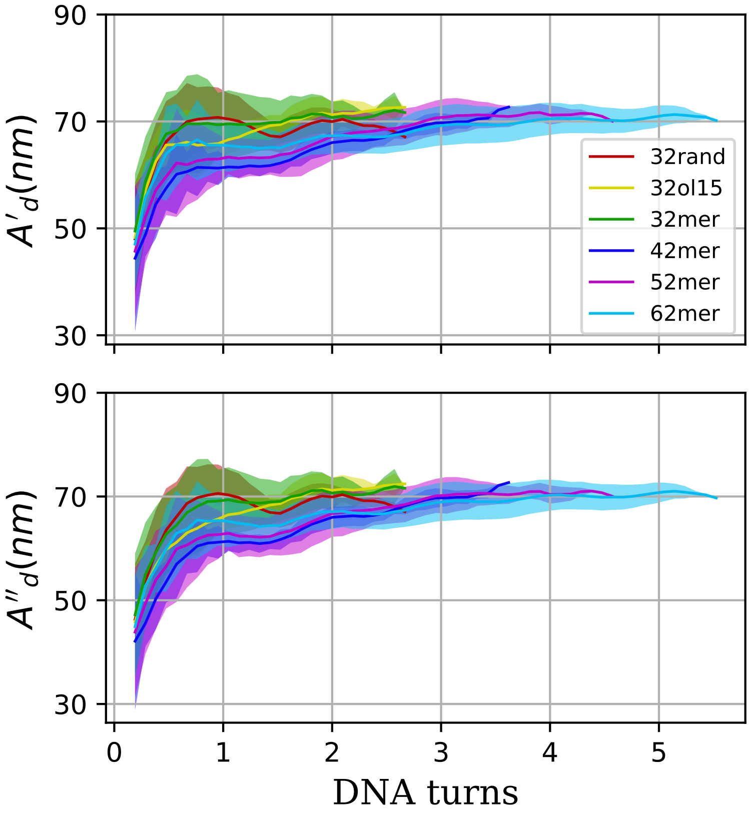


Figure C.4: Second (A'_d) and third (A''_d) estimations of the dynamic persistence length as a function of time for the set of 6 linear DNA MD simulations. Lines represent average values, while shaded areas standard deviations.

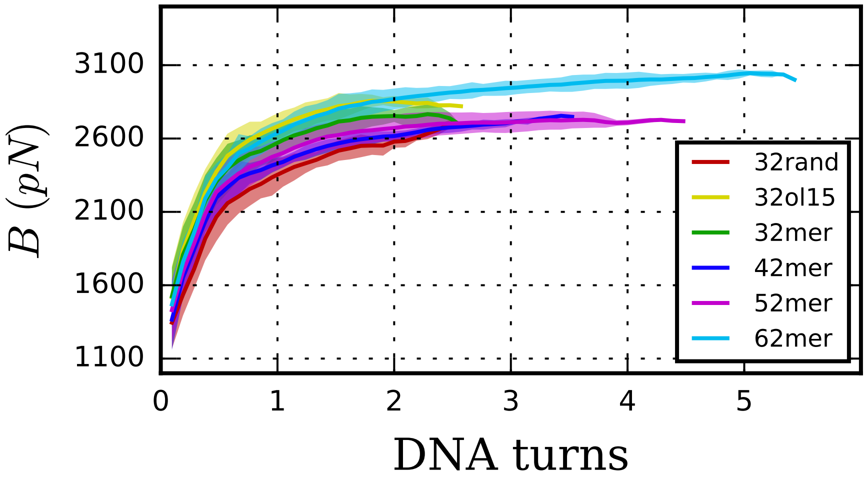


Figure C.5: Elastic profiles of the stretch modulus B obtained through the contour length (L_{CL}) instead of the end-to-end distance (L), for the set of linear DNA simulations.

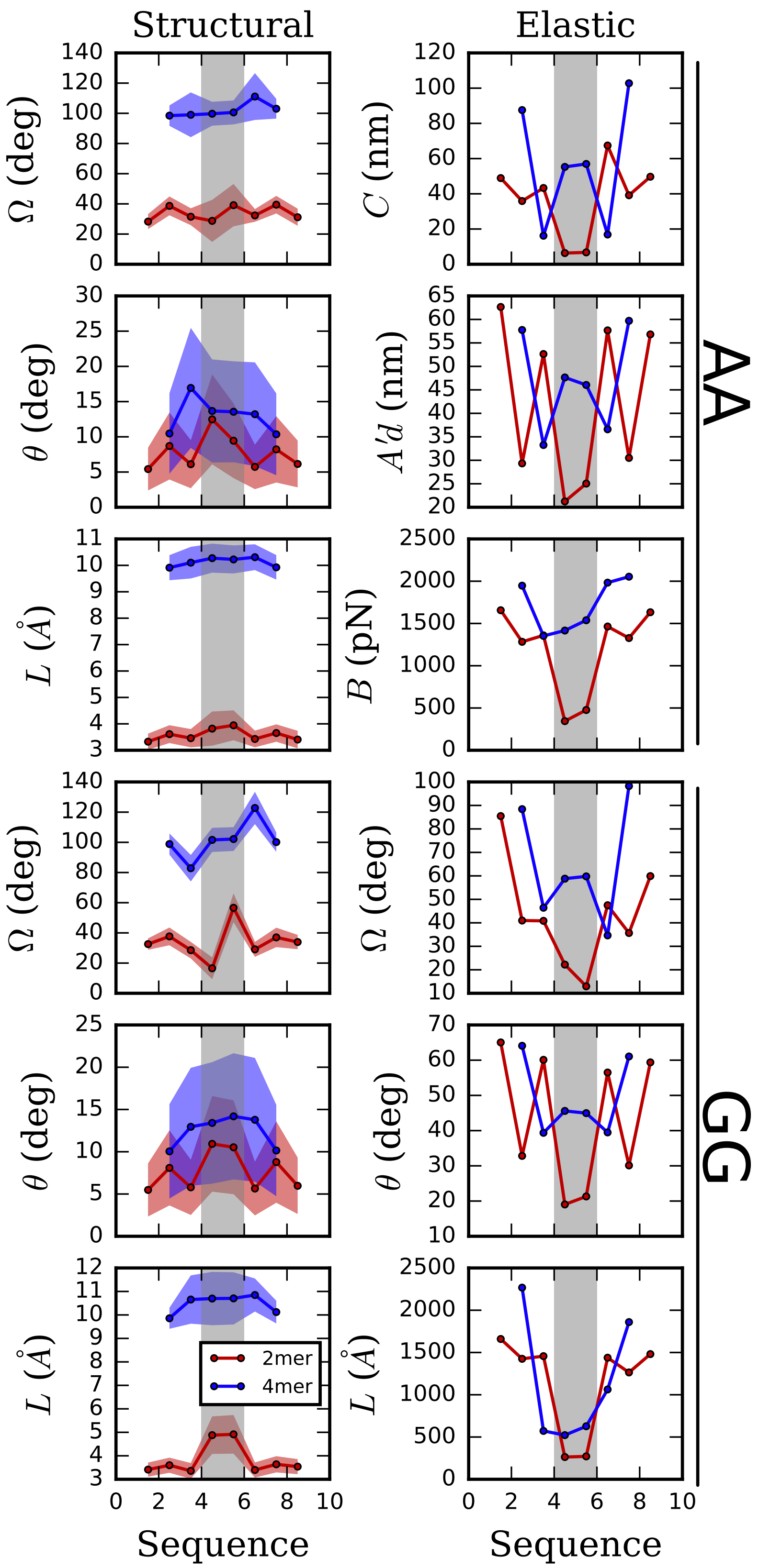


Figure C.6: Structural and elastic profiles along the sequence at the dinucleotide and tetramer level, for the AA (top) and GG (bottom) mismatches, where the shaded gray area represents the mismatch location.

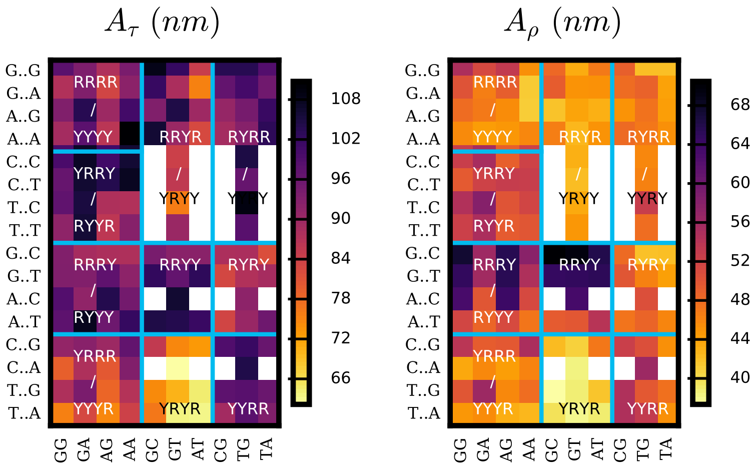


Figure C.7: Sequence-dependant elastic constants of tilt and roll (A_τ and A_ρ respectively) at the tetramer level for the 136 tetra-nucleotide sequences from the ABC simulation database. Horizontal axis represents the central base-steps while vertical axis represents the flanking base-pairs. Cyan lines sort the sequences according their purine (R) or/and pyrimidine (Y) type. Duplicated sequences are colored as white squares.

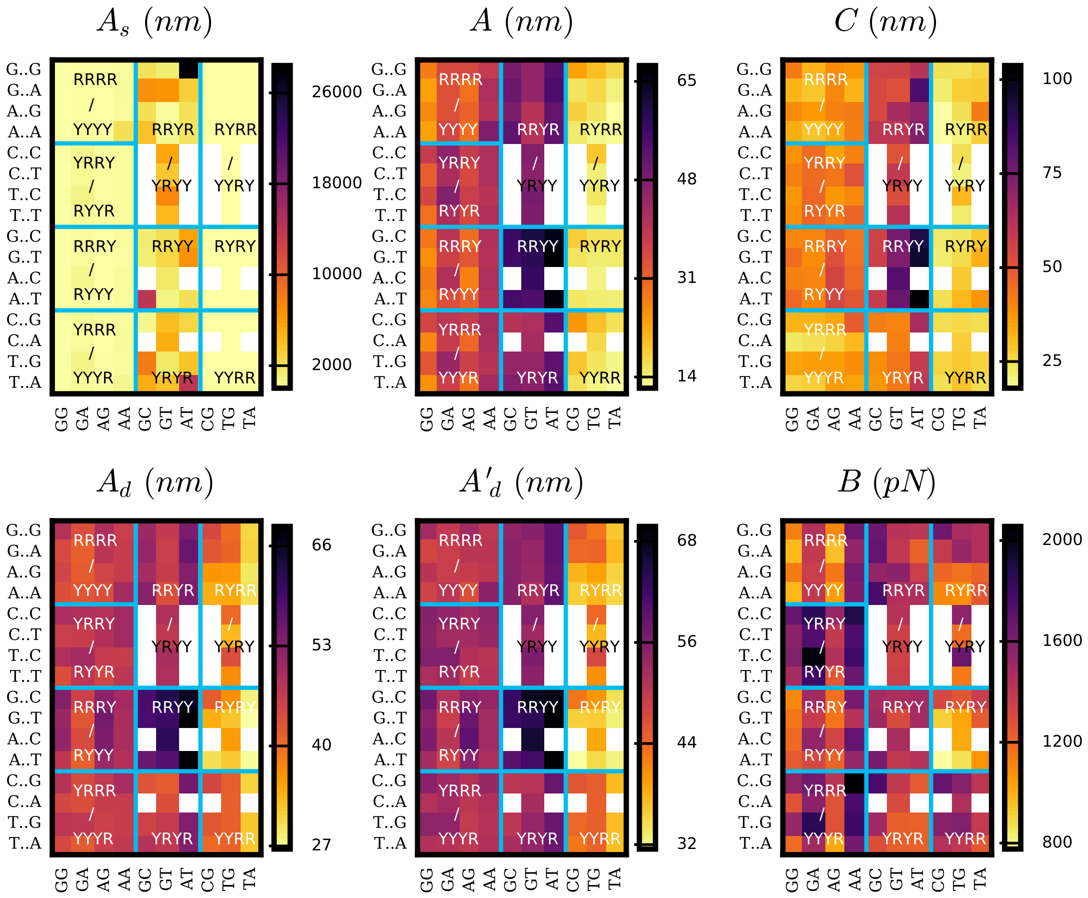


Figure C.8: Sequence-dependant elastic constants at the dinucleotide level for the set of 136 tetra-nucleotide sequences from the ABC simulation database. Flexibility is measured at the central bp-step. The persistence length (A) as well as its static (A_s) and dynamic (A_d) components are estimated through the directional decays at the bp-step level ($l = 1$bp) using equations 2.57-2.59. Twist (C), the stretch modulus (B) and the second estimation of the dynamic persistence length (A'_d) were calculated from the inverse-covariance matrix at the bp-step length. Vertical axis indicate the flanking bases while the horizontal axis the central bp-step. Cyan lines sort sequences according their purine (R) and/or pyrimidine (Y) types. Sequence duplication is avoided through white squares.

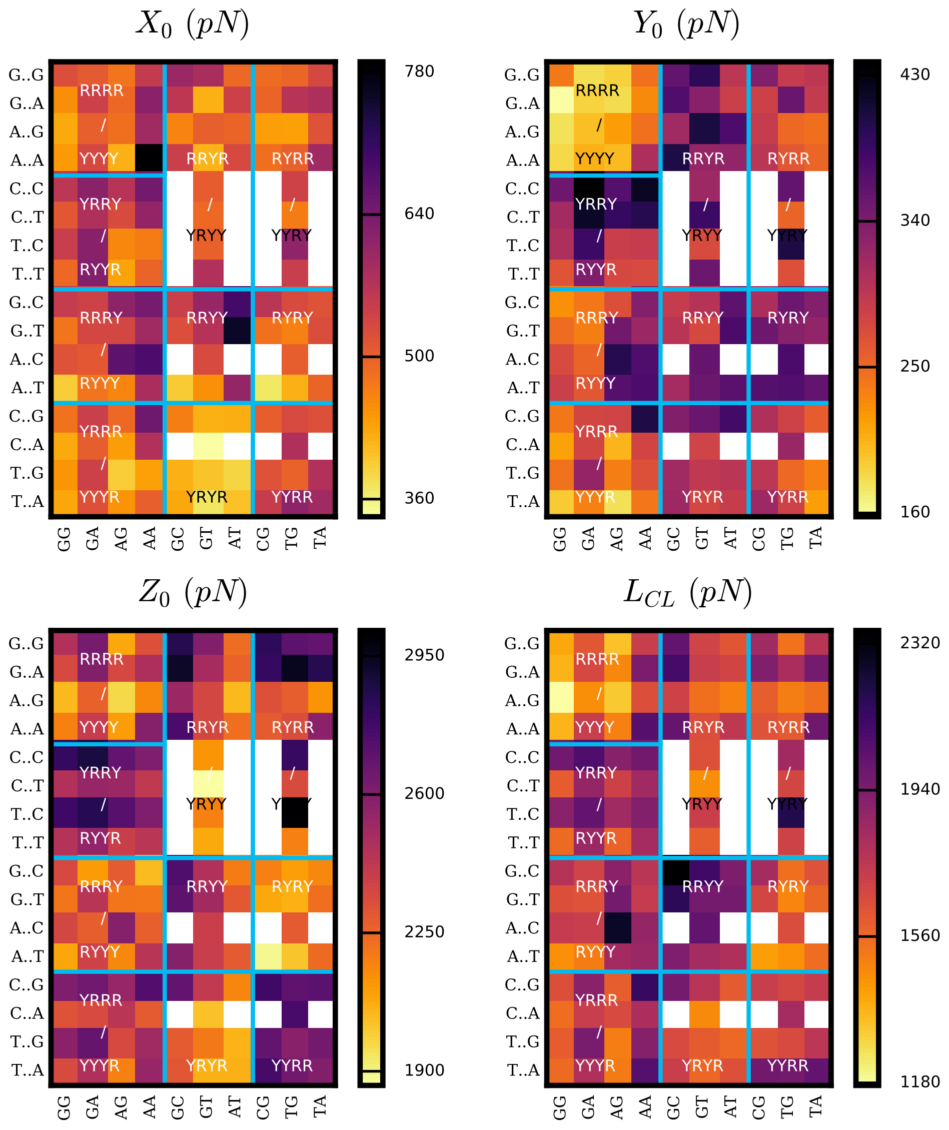


Figure C.9: Elastic constants associated with the structural parameters of added shift (X_0), slide (Y_0), rise (Z_0) and the contour length (L_{CL}) at the tetramer length for the 136 tetra-nucleotide sequences from the ABC simulation database. Horizontal axes indicate the central bp-steps while vertical axes the flanking bases. Cyan lines organize the sequences according their purine (R) and pyrimidine (Y) types. White squares avoid sequence duplication.

DNA	A'_d	A''_d
32rand	68.3 ± 1.0	68.0 ± 1.0
32mer	69.9 ± 0.5	69.2 ± 0.6
32ol15	69.0 ± 1.1	68.4 ± 1.1
42mer	64.7 ± 2.3	64.4 ± 2.5
52mer	67.8 ± 2.9	67.1 ± 3.0
62mer	68.2 ± 1.8	67.9 ± 1.9
Average	67.8 ± 1.7	67.5 ± 1.5

Table C.1: Second (A'_d) and third (A''_d) estimations of the dynamic persistence length of the set of 6 linear DNA MD simulations, calculated from plots of figure C.4 and using the SerraNA default options (see subsection 4.1.2).

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
GGGG	19.0 ± 0.6	65.4 ± 2.6	26.9 ± 1.1	79.1 ± 6.4	1541 ± 47
GGGA	24.0 ± 0.7	57.9 ± 1.1	41.0 ± 2.2	65.8 ± 2.0	1411 ± 34
AGGG	18.0 ± 0.5	59.3 ± 1.7	25.8 ± 0.7	63.7 ± 3.1	1188 ± 26
AGGA	23.5 ± 2.2	57.2 ± 0.8	40.3 ± 6.3	61.0 ± 0.8	1358 ± 39
GGAG	31.9 ± 0.6	60.7 ± 1.1	67.6 ± 3.9	69.3 ± 2.1	1863 ± 6
GGAA	36.7 ± 0.6	55.5 ± 0.7	108.1 ± 2.7	74.5 ± 4.9	1844 ± 44
AGAG	25.5 ± 2.1	60.5 ± 1.0	44.5 ± 6.4	66.9 ± 4.8	1637 ± 78
AGAA	35.4 ± 5.0	55.0 ± 4.3	101.8 ± 26.9	66.7 ± 1.7	1863 ± 136
GAGG	23.3 ± 1.6	60.6 ± 0.9	37.9 ± 3.8	58.6 ± 1.6	1652 ± 122
GAGA	27.7 ± 0.8	57.6 ± 0.5	53.4 ± 3.0	60.1 ± 1.7	1603 ± 38
AAGG	21.2 ± 1.0	56.8 ± 0.5	34.0 ± 2.7	59.2 ± 4.5	1311 ± 23
AAGA	27.4 ± 4.7	56.6 ± 2.6	54.7 ± 14.7	51.8 ± 8.4	1581 ± 84
GAAG	34.9 ± 1.3	55.4 ± 0.5	95.6 ± 10.6	61.3 ± 1.4	2028 ± 109
GAAA	46.9 ± 0.1	52.3 ± 0.1	454.1 ± 20.1	61.5 ± 2.0	2127 ± 24
AAAG	33.7 ± 0.4	54.6 ± 1.1	87.8 ± 0.1	56.8 ± 3.0	1754 ± 47
AAAA	55.8 ± 7.0	61.6 ± 3.4	970.0 ± 506.1	74.6 ± 6.6	2241 ± 88
Average	30.3 ± 9.9	57.9 ± 3.2	140.2 ± 236.0	64.4 ± 7.0	1688 ± 288

Table C.2: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of RRRR tetranucleotide sequences.

Sequence	$A\ (nm)$	$A_d\ (nm)$	$A_s\ (nm)$	$C\ (nm)$	$B\ (pN)$
TGGT	19.6 ± 0.8	61.1 ± 3.8	28.8 ± 1.0	72.9 ± 5.4	1700± 109
TGGC	28.4 ± 2.2	65.5 ± 2.4	50.4 ± 5.4	69.9 ± 2.1	2139± 82
CGGT	23.0 ± 0.5	62.7 ± 2.5	36.4 ± 0.6	69.0 ± 4.5	2072± 88
CGGC	32.7 ± 0.6	66.2 ± 2.3	64.9 ± 2.6	65.8 ± 2.7	2468± 82
TGAT	28.5 ± 1.6	68.7 ± 3.4	48.8 ± 3.2	82.3 ± 2.5	2153± 118
TGAC	42.3 ± 1.7	70.1 ± 1.9	106.7 ± 6.6	81.0 ± 0.4	2470± 39
CGAT	33.6 ± 0.7	68.1 ± 0.7	66.5 ± 2.9	68.3 ± 2.8	2421± 44
CGAC	41.5 ± 1.2	69.0 ± 2.7	104.2 ± 1.7	74.1 ± 1.4	2754± 65
TAGT	24.0 ± 1.9	57.2 ± 2.8	41.4 ± 4.3	68.0 ± 4.9	1771± 62
TAGC	33.1 ± 0.8	57.4 ± 0.9	78.1 ± 3.7	65.4 ± 1.9	2119± 49
CAGT	30.9 ± 1.1	66.9 ± 1.8	57.6 ± 3.2	78.2 ± 1.8	2097± 26
CAGC	35.5 ± 1.7	62.8 ± 2.4	81.8 ± 5.4	70.2 ± 2.0	2329± 70
TAAT	31.6 ± 2.5	57.4 ± 2.3	70.7 ± 9.0	69.5 ± 2.0	1923± 68
TAAC	38.8 ± 3.6	56.8 ± 2.4	124.0 ± 23.3	72.0 ± 2.6	2171± 16
CAAT	36.6 ± 2.5	67.4 ± 2.9	80.2 ± 7.9	80.5 ± 2.4	2147± 81
CAAC	44.6 ± 1.6	65.5 ± 2.4	139.7 ± 5.9	81.5 ± 0.9	2347± 125
Average	32.8 ± 6.9	63.9 ± 4.5	73.8 ± 30.7	73.0 ± 5.7	2193± 261

Table C.3: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of YRRY tetranucleotide sequences.

Sequence	$A\ (nm)$	$A_d\ (nm)$	$A_s\ (nm)$	$C\ (nm)$	$B\ (pN)$
GGGT	29.7 ± 1.3	60.8 ± 0.2	58.5 ± 5.3	58.4 ± 1.2	1682 ± 113
GGGC	35.9 ± 2.2	64.9 ± 2.2	80.5 ± 7.8	61.5 ± 2.0	1861 ± 105
AGGT	33.5 ± 1.7	56.7 ± 2.0	82.3 ± 7.8	60.5 ± 0.1	1498 ± 82
AGGC	36.5 ± 0.7	63.3 ± 1.4	86.5 ± 4.7	61.0 ± 3.0	1833 ± 71
GGAT	36.4 ± 0.6	57.6 ± 0.4	99.1 ± 5.8	63.9 ± 0.7	1670 ± 69
GGAC	41.9 ± 2.1	56.0 ± 0.9	169.0 ± 25.4	55.8 ± 0.0	1938 ± 46
AGAT	33.7 ± 0.4	57.6 ± 0.4	81.5 ± 3.3	74.0 ± 1.8	1589 ± 28
AGAC	41.6 ± 1.3	56.1 ± 0.3	163.2 ± 22.0	57.1 ± 0.5	1923 ± 52
GAGT	50.0 ± 2.2	64.5 ± 1.2	233.9 ± 55.2	63.2 ± 2.6	2219 ± 34
GAGC	37.4 ± 0.8	62.8 ± 2.3	93.5 ± 9.3	59.2 ± 2.4	2072 ± 24
AAGT	43.3 ± 1.6	57.6 ± 2.0	174.5 ± 6.7	62.8 ± 1.6	1993 ± 15
AAGC	40.8 ± 0.6	69.5 ± 1.4	99.1 ± 0.8	79.8 ± 3.1	2551 ± 52
GAAT	51.0 ± 2.4	57.0 ± 2.0	484.7 ± 69.6	54.8 ± 4.2	1981 ± 25
GAAC	52.5 ± 2.7	58.7 ± 3.1	496.5 ± 18.6	56.7 ± 2.2	2155 ± 97
AAAT	53.9 ± 0.2	59.2 ± 0.9	613.0 ± 78.7	62.5 ± 1.0	2185 ± 8
AAAC	50.4 ± 0.7	59.7 ± 1.6	327.7 ± 17.1	57.1 ± 2.4	2203 ± 47
Average	41.8 ± 7.4	60.1 ± 3.7	209.0 ± 170.6	61.8 ± 6.4	1960 ± 264

Table C.4: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of RRRY tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
TGGG	17.1 ± 0.3	59.8 ± 1.0	24.0 ± 0.5	60.6 ± 1.7	1673 ± 67
TGGA	18.0 ± 1.0	52.5 ± 0.8	27.6 ± 2.4	62.0 ± 0.8	1695 ± 82
CGGG	21.5 ± 1.4	65.4 ± 0.8	32.2 ± 2.9	56.6 ± 2.5	1974 ± 40
CGGA	21.9 ± 0.7	54.5 ± 0.6	36.6 ± 2.0	51.6 ± 1.9	2000 ± 44
TGAG	33.2 ± 1.0	65.5 ± 2.2	67.3 ± 2.2	72.7 ± 2.7	2335 ± 100
TGAA	28.0 ± 0.5	54.9 ± 1.8	57.5 ± 3.1	69.0 ± 1.1	2044 ± 28
CGAG	27.6 ± 0.7	61.9 ± 0.8	50.0 ± 2.3	69.3 ± 1.5	2377 ± 65
CGAA	31.6 ± 0.6	57.6 ± 2.2	70.2 ± 1.0	54.3 ± 1.3	2357 ± 100
TAGG	20.1 ± 0.7	52.6 ± 1.9	32.7 ± 1.2	59.8 ± 3.9	1663 ± 43
TAGA	16.8 ± 0.3	49.5 ± 0.6	25.5 ± 0.8	74.1 ± 3.3	1593 ± 40
CAGG	23.0 ± 1.3	60.9 ± 1.6	36.9 ± 2.9	63.5 ± 2.1	1996 ± 23
CAGA	22.4 ± 0.3	54.3 ± 0.4	38.1 ± 0.8	60.5 ± 2.4	1958 ± 83
TAAG	32.2 ± 1.5	55.8 ± 1.5	76.5 ± 6.9	67.5 ± 1.8	2186 ± 49
TAAA	35.0 ± 0.9	51.4 ± 2.2	110.4 ± 9.5	75.7 ± 1.8	2161 ± 63
CAAG	35.1 ± 1.6	66.9 ± 2.1	73.7 ± 4.9	81.2 ± 5.1	2566 ± 110
CAAA	34.6 ± 0.1	58.0 ± 0.7	85.5 ± 1.3	65.1 ± 3.2	2060 ± 87
Average	26.1 ± 6.5	57.6 ± 5.2	52.8 ± 24.6	65.2 ± 7.9	2040 ± 276

Table C.5: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of YRRR tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
GGTG	30.4 ± 0.1	57.6 ± 0.8	64.1 ± 0.7	66.0 ± 2.0	2076 ± 68
GGTA	29.2 ± 0.1	52.6 ± 0.6	65.8 ± 0.4	61.1 ± 0.9	1783 ± 53
AGTG	29.2 ± 0.7	59.5 ± 1.3	57.5 ± 2.4	66.0 ± 3.2	1996 ± 74
AGTA	22.7 ± 2.1	51.7 ± 1.5	40.6 ± 5.7	56.5 ± 3.2	1854 ± 13
TGTT	22.8 ± 0.1	58.9 ± 0.1	37.2 ± 0.2	52.5 ± 1.9	1832 ± 28
TGTC	26.1 ± 1.5	53.4 ± 0.7	51.3 ± 6.0	48.8 ± 1.3	2035 ± 114
CGTT	24.6 ± 0.1	57.2 ± 0.1	43.1 ± 0.3	41.1 ± 0.3	1931 ± 10
CGTC	26.6 ± 2.5	57.0 ± 0.6	50.7 ± 8.4	43.8 ± 1.8	2309 ± 108
GGCG	31.5 ± 0.3	65.2 ± 0.3	61.1 ± 0.9	56.3 ± 0.6	2462 ± 27
GGCA	28.1 ± 0.0	63.4 ± 1.3	50.4 ± 0.9	57.8 ± 0.3	2293 ± 41
AGCG	22.2 ± 3.1	56.6 ± 4.6	38.8 ± 12.4	60.5 ± 3.1	2171 ± 149
AGCA	19.8 ± 1.1	63.1 ± 3.6	29.0 ± 1.6	66.1 ± 6.9	2266 ± 46
GATG	25.8 ± 3.5	56.2 ± 0.4	49.0 ± 11.4	52.4 ± 2.9	2142 ± 191
GATA	20.9 ± 0.5	48.5 ± 0.2	36.7 ± 1.4	65.4 ± 1.3	1678 ± 74
AATG	24.5 ± 0.7	56.0 ± 0.2	43.6 ± 2.3	48.7 ± 1.1	1714 ± 43
AATA	25.8 ± 0.5	52.2 ± 1.7	50.9 ± 0.3	56.2 ± 0.8	1799 ± 6
Average	25.6 ± 3.3	56.8 ± 4.4	48.1 ± 10.2	56.2 ± 7.7	2021 ± 229

Table C.6: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of RRYR /YRYY tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
GGTT	38.4 ± 0.4	65.8 ± 0.8	92.4 ± 0.9	54.5 ± 0.1	2255 ± 24
GGTC	44.5 ± 0.1	65.6 ± 0.5	138.4 ± 3.3	65.1 ± 1.2	2493 ± 1
AGTT	48.2 ± 0.7	60.8 ± 0.9	232.1 ± 3.2	56.4 ± 0.4	2179 ± 12
AGTC	50.2 ± 5.8	68.3 ± 1.5	205.5 ± 65.0	62.3 ± 1.9	2533 ± 59
GGCT	42.1 ± 0.3	65.2 ± 0.6	118.9 ± 3.9	55.7 ± 2.3	2527 ± 28
GGCC	44.1 ± 0.4	65.3 ± 0.0	136.3 ± 4.0	60.7 ± 1.6	2677 ± 18
AGCT	41.5 ± 1.8	58.8 ± 1.4	141.7 ± 12.7	55.4 ± 0.5	2314 ± 67
GATT	59.4 ± 0.5	70.7 ± 0.3	370.7 ± 11.9	65.2 ± 2.6	2413 ± 49
GATC	58.1 ± 0.4	65.9 ± 0.3	495.3 ± 17.6	66.7 ± 0.9	2339 ± 15
AATT	61.7 ± 1.0	64.9 ± 0.7	1267.1 ± 144.2	57.9 ± 0.9	2209 ± 16
Average	48.8 ± 7.8	65.1 ± 3.2	319.8 ± 337.9	60.0 ± 4.4	2394 ± 154

Table C.7: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of RRYY tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
TGTG	17.5 ± 0.9	47.4 ± 2.1	28.0 ± 3.1	54.7 ± 4.5	1847 ± 136
TGTA	17.5 ± 0.4	43.3 ± 0.9	29.3 ± 0.9	64.6 ± 4.3	1682 ± 40
CGTG	19.9 ± 0.1	51.3 ± 0.8	32.7 ± 0.5	54.5 ± 1.5	2173 ± 51
CGTA	19.0 ± 0.5	44.0 ± 1.0	33.6 ± 2.0	50.9 ± 3.5	1749 ± 52
TGCG	18.9 ± 0.3	48.8 ± 0.9	30.7 ± 0.9	50.8 ± 1.8	2095 ± 53
TGCA	15.1 ± 0.1	49.7 ± 0.3	21.7 ± 0.2	69.7 ± 1.1	1850 ± 30
CGCG	25.6 ± 0.7	56.4 ± 1.9	47.1 ± 2.6	52.1 ± 1.0	2660 ± 63
TATG	17.7 ± 0.8	43.9 ± 0.3	29.8 ± 2.3	60.6 ± 0.2	1676 ± 12
TATA	17.6 ± 0.8	41.1 ± 0.8	31.0 ± 2.8	70.1 ± 4.1	1672 ± 30
CATG	16.3 ± 0.5	53.2 ± 0.9	23.6 ± 0.9	60.4 ± 0.2	1887 ± 38
Average	18.5 ± 2.7	47.9 ± 4.6	30.8 ± 6.5	58.9 ± 7.0	1929 ± 293

Table C.8: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of YRYR tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
GTGG	30.0 ± 0.6	56.0 ± 1.7	64.9 ± 0.6	85.1 ± 1.6	2271 ± 131
GTGA	46.1 ± 0.6	61.5 ± 0.2	184.6 ± 7.5	92.2 ± 1.8	2645 ± 14
ATGG	20.7 ± 1.4	59.9 ± 1.2	31.7 ± 3.7	79.3 ± 0.8	1746 ± 203
ATGA	26.6 ± 1.2	59.1 ± 1.1	48.7 ± 4.1	83.3 ± 0.6	1875 ± 54
TTGT	29.6 ± 1.4	61.3 ± 2.1	57.3 ± 3.5	74.4 ± 5.4	1898 ± 31
TTGC	43.6 ± 2.1	67.8 ± 1.3	122.6 ± 13.1	93.4 ± 1.7	2829 ± 164
CTGT	24.0 ± 0.7	59.9 ± 0.4	40.1 ± 1.9	81.7 ± 2.5	1889 ± 52
CTGC	37.9 ± 0.4	62.1 ± 2.4	97.7 ± 4.1	82.2 ± 6.3	2497 ± 87
GTAG	34.9 ± 2.2	54.9 ± 3.0	96.1 ± 7.4	81.1 ± 4.5	2147 ± 148
GTAA	40.3 ± 2.6	54.0 ± 1.3	161.7 ± 28.9	84.9 ± 1.4	2378 ± 111
ATAG	19.7 ± 0.5	55.9 ± 0.5	30.4 ± 1.2	97.3 ± 2.7	1504 ± 23
ATAA	32.0 ± 3.2	54.4 ± 1.0	79.7 ± 16.8	83.8 ± 3.5	1919 ± 58
GCGG	32.6 ± 0.1	65.2 ± 2.1	65.5 ± 2.4	69.3 ± 6.4	2659 ± 116
GCGA	30.2 ± 0.1	62.5 ± 0.3	58.5 ± 0.8	87.1 ± 1.2	2503 ± 66
ACGG	21.7 ± 0.8	60.0 ± 0.9	34.0 ± 2.0	74.4 ± 1.3	2050 ± 116
ACGA	26.9 ± 0.9	62.0 ± 1.5	47.4 ± 2.0	62.0 ± 1.4	2111 ± 48
Average	31.1 ± 7.7	59.8 ± 3.8	76.3 ± 44.6	82.0 ± 8.7	2183 ± 363

Table C.9: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of RYRR/YYRY tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
GTGT	29.6 ± 1.3	55.7 ± 0.7	63.2 ± 5.4	72.6 ± 1.9	1599 ± 47
GTGC	35.1 ± 0.2	53.9 ± 0.8	100.8 ± 4.3	70.7 ± 1.7	2014 ± 38
ATGT	31.0 ± 0.4	56.4 ± 0.6	69.0 ± 0.8	83.8 ± 3.7	1455 ± 54
ATGC	29.4 ± 1.0	59.8 ± 1.7	58.5 ± 6.0	79.7 ± 3.4	1785 ± 106
GTAT	30.6 ± 0.7	50.3 ± 0.0	78.4 ± 4.6	91.0 ± 0.1	1550 ± 22
GTAC	31.5 ± 0.2	49.0 ± 0.1	88.0 ± 2.2	82.2 ± 3.1	1792 ± 78
ATAT	31.6 ± 0.9	51.9 ± 1.1	81.0 ± 4.9	97.5 ± 3.8	1448 ± 22
GCGT	34.0 ± 0.5	56.7 ± 0.7	85.0 ± 1.8	58.6 ± 1.2	2047 ± 37
GCGC	46.4 ± 1.7	57.7 ± 0.8	239.7 ± 31.4	59.9 ± 1.5	2642 ± 90
ACGT	30.9 ± 1.7	54.9 ± 0.7	71.2 ± 8.5	65.8 ± 1.6	1569 ± 41
Average	33.0 ± 4.8	54.6 ± 3.2	93.5 ± 50.2	76.2 ± 12.3	1790 ± 349

Table C.10: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of RYRY tetranucleotide sequences.

Sequence	A (nm)	A_d (nm)	A_s (nm)	C (nm)	B (pN)
TTGG	26.7 ± 0.8	59.6 ± 2.7	48.5 ± 1.1	88.0 ± 5.0	2313 ± 81
TTGA	32.1 ± 0.7	64.6 ± 2.1	64.0 ± 1.5	93.7 ± 1.2	2466 ± 86
CTGG	30.8 ± 1.5	61.6 ± 1.6	61.7 ± 4.6	87.3 ± 1.3	2529 ± 66
CTGA	36.0 ± 2.0	67.1 ± 2.4	78.5 ± 10.5	94.4 ± 2.4	2597 ± 107
TTAG	32.9 ± 1.3	57.2 ± 1.6	77.3 ± 4.2	84.4 ± 2.0	2245 ± 70
TTAA	33.7 ± 0.9	55.7 ± 0.6	85.5 ± 4.3	80.4 ± 1.9	2276 ± 55
CTAG	34.5 ± 1.3	56.8 ± 1.9	87.7 ± 4.1	78.7 ± 8.1	2365 ± 97
TCGG	26.3 ± 0.3	66.0 ± 1.5	43.7 ± 0.9	83.3 ± 3.0	2546 ± 132
TCGA	29.6 ± 1.4	65.4 ± 0.3	54.4 ± 4.5	81.1 ± 2.4	2698 ± 45
CCGG	27.4 ± 0.6	65.0 ± 0.8	47.3 ± 1.5	83.0 ± 2.3	2755 ± 61
Average	31.0 ± 3.2	61.9 ± 4.1	64.9 ± 15.6	85.4 ± 5.1	2479 ± 168

Table C.11: Persistence length A, its static A_s and dynamic A_d components and twist C and stretch modulus B of YYRR tetranucleotide sequences.

Parameter	Average
A (nm)	31.8 ± 1.0
A_d (nm)	58.8 ± 5.9
A'_d (nm)	64.3 ± 7.1
A_s (nm)	108.1 ± 158.9
C (nm)	68.0 ± 12.0
B (pN)	2054 ± 354

Table C.12: Averages and standard deviations of tetranucleotide elastic constants.

Appendix D

Chapter 3 appendix

Elastic constant	52mer	62mer	Average
A_ρ (nm)	71.09	69.33	70.21 (1.24)
A_τ (nm)	70.20	69.76	69.98 (0.31)
A'_d (nm)	70.64	69.54	70.09 (0.78)
C (nm)	104.3	100.55	102.42 (2.65)

Table D.1: Elastic constants calculated with parameters from the curve fittings of figure 5.3. Parameters in parenthesis represent standard deviations.

D.1 Roll limit

The limit of the roll elastic constant is:

$$\lim_{l\to\infty} A_\rho(l) = \lim_{l\to\infty} \frac{r_d^2 b_0 l}{a_\rho + b_\rho l + c_\rho sin(\omega_\rho l + \phi_\rho l)} = \frac{\infty}{\infty} \tag{D.1}$$

And applying l'Hospital's rule once more and ignoring the periodic component:

$$\lim_{l\to\infty} A_\rho(l) = \lim_{l\to\infty} \frac{r_d^2 b_0}{b_\rho} = \frac{r_d^2 b_0}{b_\rho} \tag{D.2}$$

This last expression corresponds to a plateau, but if we do not ignore the periodic function in the denominator, and supposing that we are at very long lengths, the roll elastic constant would adopt the following shape:

$$A_\rho(l) = \frac{r_d^2 b_0}{b_\rho + c_\rho \omega_\rho sin(\omega_\rho l + \phi_\rho)} \tag{D.3}$$

D.2 Tilt limit

Similarly than the roll elastic constant and applying l'Hospital's rule, as l approaches infinity, the tilt elastic constant tends to:

$$\lim_{l\to\infty} A_\tau(l) = \lim_{l\to\infty} \frac{r_d^2 b_0 l}{a_\tau + b_\tau l + c_\tau sin(\omega_\tau l + \phi_\tau l)} = \frac{r_d^2 b_0}{b_\tau} \tag{D.4}$$

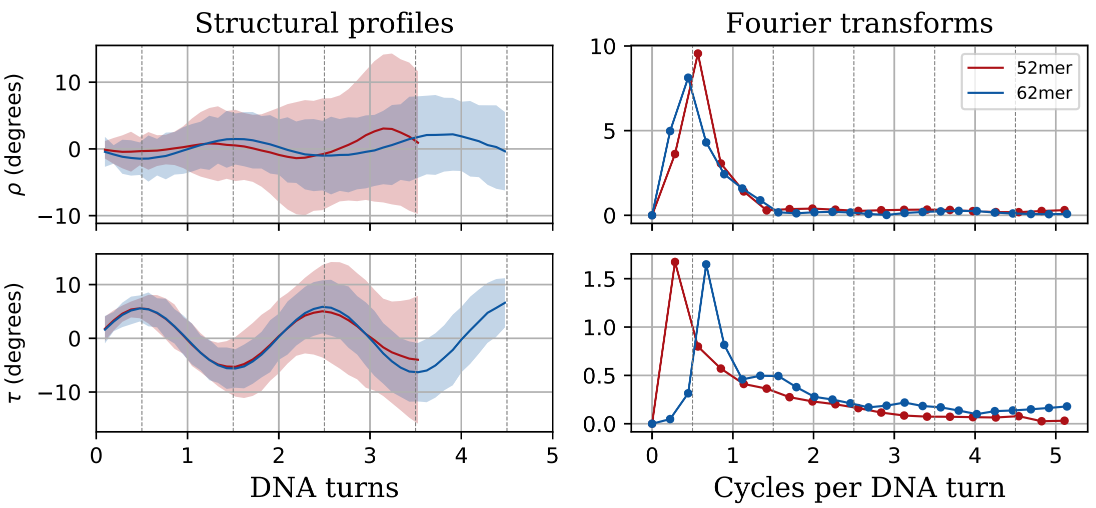


Figure D.1: Structural profiles and their corresponding Fourier transforms of the roll (ρ) and tilt (τ) angles for the 52mer and 62mer. The shaded areas indicate the standard deviation.

D.3 Twist limit

The limit of the twist elastic constant is calculated as:

$$\lim_{l\to\infty} C(l) = \lim_{l\to\infty} \frac{r_d^2 b_0 l}{a_\Omega + b_\Omega l} = \frac{\infty}{\infty} \tag{D.5}$$

And applying l'Hospital's rule we get:

$$\lim_{l\to\infty} C(l) = \lim_{l\to\infty} \frac{r_d^2 b_0}{b_\Omega} = \frac{r_d^2 b_0}{b_\omega} \tag{D.6}$$

Appendix E

Software access and installation instructions

E.1 SerraLINE access and installation

SerraLINE is a software tool developed for calculating bending angles, width, height, aspect ratio, and deviation from planarity distributions from simulations of the global molecular contour of DNA molecules. This molecular contour can be generated using the WrLINE program [150], where the molecular contour is defined by a set of coordinates, each representing a base-pair. The SerraLINE software is designed to process both closed DNA structures (e.g., minicircles) and open structures (e.g., linear DNA), making it a versatile tool for structural analysis. SerraLINE also offers the capability to project structures onto a best-fit plane, simulating single-molecule experiments such as atomic force microscopy (AFM) where the molecule is visualised in 2D. The outputs generated by SerraLINE are suitable for comparison with experimental structural data.

E.1.1 Overview

SerraLINE is an autocontained software written purely in Fortran, requiring no external libraries. It is available under version 3.0 of the GNU Lesser General Public License[*], and can be accessed from the agnesnoy/SerraLINE GitHub repository[†].

The software comprises two main programs:

- **SerraLINE**: This program processes input structures, performs mathematical procedures to calculate structural parameters, and outputs the results in a human-readable format.

- **Extract**: A supportive program that filters the structural parameters for visualisation and analysis purposes.

The repository also includes an example demonstrating how to run SerraLINE and process its outputs for analysis and plotting.

[*] `https://www.gnu.org/licenses/gpl-3.0.en.html`

[†] `https://github.com/agnesnoy/SerraLINE`

E.1.2 Requirements

To compile SerraLINE, you only need a Fortran compiler (e.g. gfortran).

E.1.3 Software Download

You can download SerraLINE from the GitHub repository by two ways:

Manual Download

1. Navigate to the repository: Visit the GitHub repository at:

   https://github.com/agnesnoy/SerraLINE

2. Download the repository: On the repository's main page, click the green "Code" button (top right) and select "Download ZIP". This action will download a ZIP archive (SerraLINE-master.zip) containing the entire repository.

3. Extract the ZIP archive: After the download, locate the ZIP file, right-click it, and select "Extract All" to reveal the repository contents.

Download via Command Line

1. Open terminal/command prompt: Launch your terminal or command prompt.

2. Navigate to the installation directory: Use the `cd` command to navigate to the directory where you want to install SerraLINE, for example:

   `cd /path/to/SerraLINE_directory`.

3. Download the repository: Clone the repository using Git:

   ```
   git clone https://github.com/agnesnoy/SerraLINE.git
   ```

4. Navigate to the repository directory: Move into the downloaded repository directory using `cd SerraLINE`.

E.1.4 Compilation and Installation

Once inside the SerraLINE directory, execute the command `make all`. This command compiles both the main SerraLINE program and the supportive Extract program, resulting in two executable files ready for use.

For detailed instructions on operating the program and running the example provided in the repository, please refer to SerraLINE's manual F.

E.2 SerraNA access and installation

SerraNA is a software designed for calculating structural and flexibility parameters from atomistic molecular dynamics simulations of nucleic acid structures [157] based on the Length-Dependent Elastic Model [114]. This program can process both single-stranded and double-stranded nucleic acid molecules, as well as closed (e.g., minicircles) and open (linear) structures. SerraNA generates outputs that enable the analysis of structural and elastic parameters at various length scales, while also offering insights into sequence-dependent properties. Additionally, a key feature of SerraNA is its capability to infer global elastic constants, including bending persistence length, twist modulus, and stretch modulus. These constants provide a characterisation of the overall flexibility of the molecule and can be compared with experimental measurements.

E.2.1 Overview

SerraNA is a software written in Fortran that operates independently, without the need for external libraries. It is distributed under version 3.0 of the GNU Lesser General Public License* and is available from the agnesnoy/SerraNA GitHub repository†.

The software package consists of three primary programs:

- **SerraNA**: The core program responsible for processing input molecular structures and calculating structural and flexibility parameters. It produces results into various files in human-readable format.
- **Analysis**: This program processes the output parameters generated by SerraNA and employs mathematical procedures to estimate the global elastic constants.
- **Extract**: A supporting program that processes SerraNA outputs, creating simplified files that are ready for plotting.

The repository includes detailed compilation and execution instructions as well as an example with a small trajectory to guide users through preparing input files, executing the main programs, and analysing and visualising the generated data.

E.2.2 Requirements

The only requirement for compiling and using SerraNA is a Fortran compiler (e.g. gfortran).

E.2.3 Software Download

SerraNA can be acquired from the GitHub repository using either of the following methods:

Manual Download

1. Navigate to the repository: Visit the GitHub repository at:

   https://github.com/agnesnoy/SerraNA

2. Download the repository: On the repository's main page, click the green "Code" button (top right) and select "Download ZIP". This action will download a ZIP archive named `SerraNA-master.zip` containing the entire repository.
3. Extract the ZIP archive: After the download, locate the ZIP file, right-click it, and select "Extract All" to reveal the repository contents.

*https://www.gnu.org/licenses/gpl-3.0.en.html

†https://github.com/agnesnoy/SerraNA

Download via Command Line

1. Open terminal/command prompt: Launch your terminal or command prompt.
2. Navigate to the installation directory: Use the `cd` command to navigate to the directory where you want to install SerraNA, for example:

   `cd /path/to/SerraNA_directory`.
3. Download the repository: Clone the repository using Git:

   ```
   git clone https://github.com/agnesnoy/SerraNA.git
   ```
4. Navigate to the repository directory: Move into the downloaded repository directory using `cd SerraNA`.

E.2.4 Compilation and Installation

Once you have downloaded SerraNA and are inside the directory, execute the command `make all`. This command will compile the main programs `SerraNA` and `Analysis`, as well as the supportive `Extract` program. This will produce three executable files that are ready for use.

For comprehensive instructions on operating the software and running the provided example from the repository, please refer to SerraNA's manual F.

Appendix F

Software manuals

SerraNA

VERSION 1.0

by Victor Velasco and Agnes Noy

York, UK, 2019

Introduction

SerraNA constitutes a software for analysing elastic and structural properties of nucleic acids using ensembles obtained by molecular dynamics (MD) or Monte Carlo (MC) simulations. By analysing all sub-fragment lengths, the program allows to infer global elastic constants describing fragment's overall flexibility (Figure 1). It is composed by three executables, ***SerraNA, Analysis*** and ***Extract***. The workflow is summarized in Figure 2.

Installation

The only requirement for running *SerraNA* is a FORTRAN compiler. The program can be compiled on a terminal by typing:

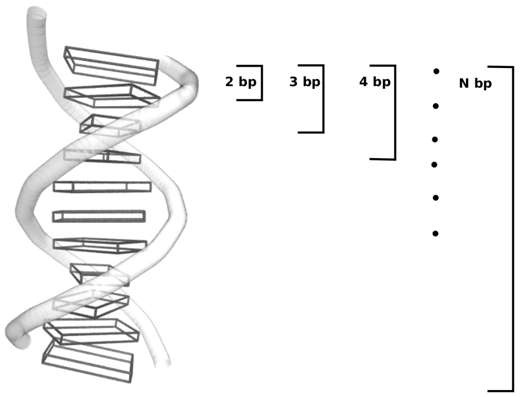


Figure 1: Parameters are calculated between every two bp comprising an oligomer whose length ranges from 2 bp to N, being N the total number of bp of the oligomer. If molecule is linear, two bp for each end are discarded

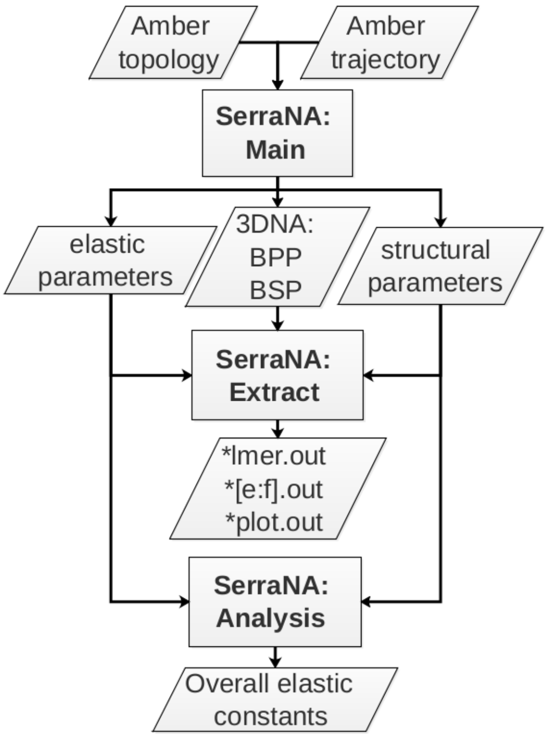


Figure 2: General workflow of **SerraNA**

```
$ make all
```

This will produce the three executables. They can also be compiled separately:

```
$ make SerraNA
$ make Analysis
$ make Extract
```

SerraNA

It is the main program that processes the DNA trajectory and calculates structural and elastic parameters at all sub-fragment lengths. It will run by typing:

```
$./SerraNA < s_NA.in
```

A trajectory file and a topology file in AMBER style format is needed (10F8.3 for the trajectory). The files can contain ions or other residues and *SerraNA* will ignore them.

s_NA.in is the input file that indicates:

- The path for topology and trajectory.
- If the structure is double-stranded (typing ”2”) or if it is single-stranded (typing ”1”).
- If the structure is linear (”1”) or closed (”2”). For linear nucleic acids, *SerraNA* ignores the two base-pairs at each end for avoiding end effects.

The program generates four outputs:

1. **BPP.out** contains the six base-pair (bp) parameters: shear, stretch, stagger, buckle, propeller and opening as they are calculated in 3DNA. It presents averages

and standard deviations over the whole MD or MC ensemble. The output is only written for double-stranded DNA.

2. **BSP.out** contains the six bp-step parameter (shift, slide, rise, tilt, roll and twist) plus bending angle. It presents averages and standard deviations over the whole ensemble. Values should be directly comparable to the ones obtained by 3DNA.

3. **structural_parameters.out** which have variables describing the geometry of the DNA molecule for all possible sub-fragments using an extension of CEHS algorithm as it is explained on Figure 3 and on [1]. It presents averages and standard deviations over the whole ensemble:

   - Twist and bending angles, roll and tilt, which denote bending towards the major groove and backbone, respectively, at the mid-point of the specified fragment
   - Added-shift, added-slide, added-rise, which are the counterparts of the translational bp-step parameters for longer lengths and are defined simply by the addition of values at 2 bp level.
   - End-to-end distance and contour length
   - $< \theta >$ ("Bending" as it's labelled on the output), $< \theta^2 >$ ("Bending**2"), and $< \cos\theta >$ ("D correlation")
   - From averaged structure: $< \theta_s >$ ("AVSTR B"), $< \theta_s^2 >$ ("AVSTR B**2"), and $< \cos\theta_s >$ ("AVSTR D C"), where θ_s is the static curvature. Average structure is built with mean values of base-step parameters at 2 bp level

   Translation are in Å and rotations are in degrees. Note that some of the variables for 2-mers will be directly compatible with those printed in the **BSP.out** output and with 3DNA.

4. **elastic_parameters.out**, which containts the following parameters for all sub-fragments:

   - elastic constants for stretch (pN), twist (nm), roll (nm), tilt (nm), as well as their couplings (nm). These are the terms of elastic matrix $F = k_B T b N V^{-1}$, where V is the corresponding covariance matrix, b is average rise and N is the number of bp-steps.
   - Dynamic persistence lengths defined via $A_d^{-1} = 1/2(A_{tilt}^{-1} + A_{roll}^{-1})$.
   - Variance and partial variance for end-to-end distance (in Å^2) as they are relevant for the calculation of the global stretch modulus.

   This file is only written if the trajectory has more than 1 snapshot.

structural_parameters.out and **elastic_parameters.out** have (N-1)! values for each variable, being N the total number of bp, if it is circular DNA, and the total number of bp minus two for each end, if it is linear DNA.

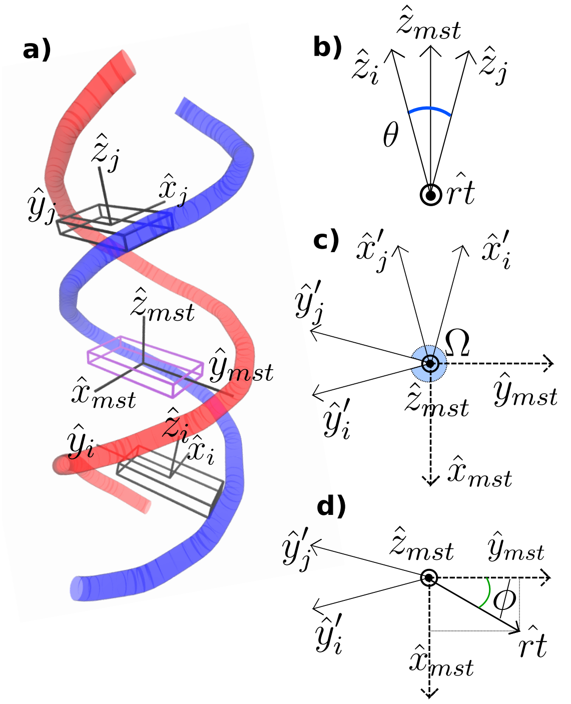


Figure 3: Schematic diagrams of SerraNA's method. (a) Bp-triads and mid-base triad are defined as in 3DNA using bending angle (θ) and roll-tilt axis ($\hat{rt}$) (b). Co-planar vectors $\hat{y}'_i$, $\hat{y}'_j$, $\hat{x}_{mst}$ and $\hat{y}_{mst}$ define twist angle Ω (c) and roll and tilt bending angles (d) with the help of auxiliary angle ϕ.

Analysis

This is the program that calculates the elastic constants at a more global level describing the whole DNA fragment. For execution, simply type:

```
$./Analysis < ov_NA.in
```

ov_NA.in is the input file that indicates:

1. The path to **elastic_parameters.out** and **structural_parameters.out**
2. The part of the molecule that will be used to calculate each of the global elastic constants. Two ranges should be provided:
   - The first one defines the region of the molecule used (from bp "a" to bp "b").
   - DEFAULT OPTION a=b=0 considers the whole fragment except for the stretch modulus where only the central 18-mer is taken to avoid long end-effects.
   - The second indicates the range of sub-lengths analyzed, being from "c" to "d" bp-steps. Note that c > 0 and d <= b-a.
   - DEFAULT OPTION c=d=0 applies the recommended methodology described in [1]:
     - For twist and dynamic persistence length, c=11 for avoiding local irregularities and d=N-10 to have at least ten different values for each sub-length.

  - For stretch modulus, c=8 for avoiding short-ranged stacking effects and d=17 due to only central 18-mer is used

Analysis outputs information on screen (see Figure 4) regarding the global elastic constants (for more information see [1]):

1. Total, static and dynamic persistence lengths obtained through the linear fitting of the corresponding directional correlation decays (labelled as A^a, A^a_s and A^a_d, respectively). Note that the fitting always uses sub-lengths ranging from 1 bp-step to N-10 bp-steps.
2. Total persistence length recalculated through $1/A = 1/A_s + 1/A_d$ using A^a_s and A^a_d and being labelled as A^b. A^a and A^b should be almost equal.
3. Torsion modulus, together with global values of tilt, roll and the associated A^c_d obtained through averages of the mean values by length. Standard errors are printed for these variables.
   Note that for sufficiently long molecules containing a few DNA turns, roll and tilt converges with A^c_d as the imbalance between directions facing towards grooves and backbone dissipates.
   Also note that $A^c_d > A^a_d$ as A^c_d is based on partial variances. These are the reciprocal of V^-1 diagonal terms and are the residual variance left after removing the influence from other variables
4. Total persistence length recalculated through $1/A^d = 1/A^a_s + 1/A^c_d$.
   Note that A^d should be bigger than all other estimations of A due to the use of partial variances on A^c_d
5. Stretch modulus calculated through the linear fitting of end-to-end partial variances using the central 18-mer for avoiding long end-effects

The interval of confidence are calculated as [1] for the variables obtained through linear fits.

Extract

Extract program process *SerraNA* ouputs (**BPP.out**, **BSP.out**,**elastic_parameters.out** and **structural_parameters.out**) creating simple files ready to plot. For each parameter, you can filter a particular sub-length l to produce structurals/elastic profiles along the molecule (Figure 4a) or you can extract averages and standard deviations as a function of length from a particular region (Figure 4b) or from the whole molecule (Figure 4c)[1]. For running it type:

```
$./Extract < ex_NA.in
```

ex_NA.in is the input file that indicates:

1. Path to either BPP, BSP, structural or elastic parameters output file. If you selected to extract BPP.out or BSP.out, then all other inputs will be ignored.
2. Type "0" for extracting a sub-length or "1" for getting avg+-sd as a function of length
3. The following entry is used to indicate:
   - The length (l) you want to process, which should be 0 < l < N bp-steps, if you typed "0" before

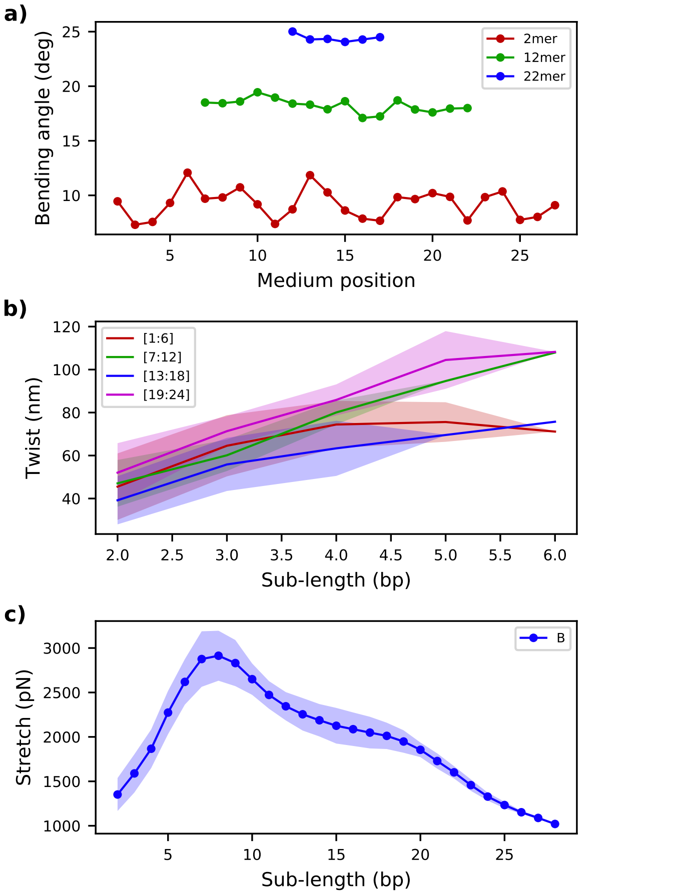


Figure 4: *Extract* tool creates simple files for (a) plotting profiles along the molecule for a sub-length l (*lmer.out type of output files) and for (b) plotting the length-dependence from bp e to f (*[e:f].out) or (c) from the whole fragment (*plot.out).

   - The region from e to f bp, from which you want to extract avg+-sd as a function of length, if you typed "1" before:
     - If it is linear DNA, then 0 < e < f < N
     - If it is circular DNA, then both e < f or f < e, are valid
     - DEFAULT OPTION, e=f=0, consider the whole fragment.

The program creates different types of outputs:

1. **BPP_plot.out**, which presents parameters with the following order of columns, being averages (first) and standard deviations (second) calculated over all the ensemble. The order of variables is the same as the processed **BPP.out** file:

1	base-pair i
2,3	Shear
4,5	Stretch
6,7	Stagger
8,9	Buckle
10,11	Propeller
12,13	Opening

2. **BSP_plot.out**, as previously, the order of variables is the same as the processed **BSP.out** file:

1	Medium position of bp-step
2,3	Shift
4,5	Slide
6,7	Rise
8,9	Tilt
10,11	Roll
12,13	Twist
14,15	Bending

3. **structural_lmer.out** to extract parameters for a particular sub-length l in the same order as the processed **structural_parameters.out** file:

1	Medium position along the sub-fragment
2,3	Added shift
4,5	Added slide
6,7	Added rise
8,9	End-to-End L
10,11	Contour L
12,13	Twist
14,15	Roll
16,17	Tilt
18,19	Bending
20,21	Bending**2
22,23	D correlation
24	AVSTR B
25	AVSTR B**2
26	AVSTR D C

4. **structural_[e:f].out** to extract length-dependence for a particular molecular part:

1	Sub-length (in bp)
2,3	Added shift
4,5	Added slide
6,7	Added rise
8,9	End-to-End L
10,11	Contour L
12,13	Twist
14,15	Roll
16,17	Tilt
18,19	Bending
20,21	Bending**2
22,23	D correlation
24,25	AVSTR B
26,27	AVSTR B**2
28,29	AVSTR D C

5. **structural_plot.out** if DEFAULT OPTION e=f=0.

6. **elastic_lmer.out** to extract parameters for a particular sub-length l following the same order as the processed **elastic_parameters.out** file:

1	Medium position along the sub-fragment
2	Stretch
3	Twist
4	Roll
5	Tilt
6	Stretch-Twist
7	Stretch-Roll
8	Stretch-Tilt
9	Twist-Roll
10	Twist-Tilt
11	Tilt-Roll
12	Dynamic Persistence Length
13	Variance End-End
14	Partial variance End-End

7. **elastic_[e:f].out** to extract length-dependence for a particular molecular part:

1	Sub-length (in bp)
2,3	Stretch
4,5	Twist
6,7	Roll
8,9	Tilt
10,11	Stretch-Twist
12,13	Stretch-Roll
14,15	Stretch-Tilt
16,17	Twist-Roll
18,19	Twist-Tilt
20,21	Tilt-Roll
22,23	Dynamic Persistence Length
24,25	Variance End-End
26,27	Partial variance End-End

8. **elastic_plot.out**, if DEFAULT OPTION e=f=0.

When a particular sub-length is choosen, then the program places items of x-axis (first column) in the medium position along the fragment. For example:

- At sub-length = 1 bp-step, parameters between residue 1 and 2, will be positioned at 1.5, between residue 2 and 3 at 2.5 etc

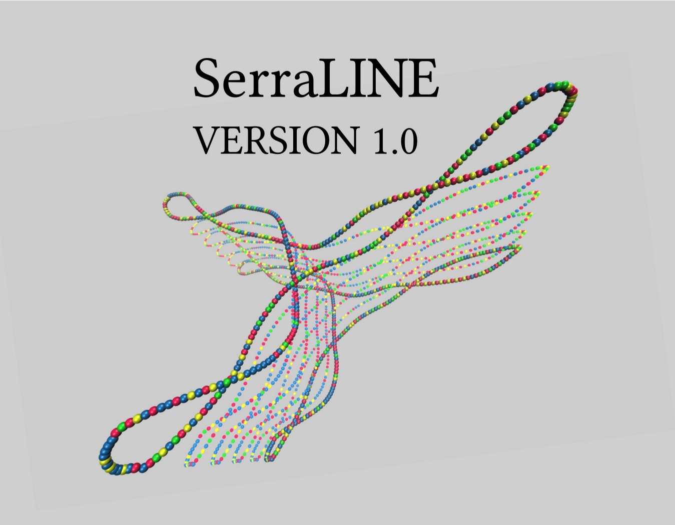
- At sub-length = 2 bp-steps, parameters between residues 1 and 3 will be at 2, between residues 2 and 4 at 3 etc.

- And so on

Note that **a, b, c** and **d** speficied in the executable *Analysis* are totally independent from **e** and **f** specified here in *Extract*, since the first program has the goal to calculate global elastic constants and the second just aims to become an utility for plotting data.

References

[1] Victor Velasco, Matthew Burman, Jack W. Shepherd, Mark C. Leake, Ramin Golestanian and Agnes Noy, "*SerraNA*: a program to infer elastic constants from local to global using nucleic acids simulation data" in BioRxiv, 2020.



by Victor Velasco and Agnes Noy

York, UK, 2020

Introduction

SerraLINE is a software for calculating bending angles, width, height, aspect ratio, and deviation from planarity of DNA molecules using the global molecular contour WrLINE. The molecular contour WrLINE defines a coordinate for each base-pair (bp). Bending angles are measured between two tangent vectors that can be separated by a number of bp (or points, Figure 2). The program can process closed (circular) or opened (linear) trajectories of DNA. The program can project the molecular contour to a global plane that best fits the molecule or to a specific region given by the user. Bending angles are calculated with the same criteria with or without the projection. The projection method mimics imaging experiments where the molecules are visualised in a 2D plane. Global quantities such as width, height, aspect ratio (width/high) and deviation from planarity can only be calculated with the projection method, and are suitable for comparison with experiments (Figures 3 & 4). The software is composed of two executables, ***SerraLINE*** and ***Extract***. The workflow is summarised in Figure 1.

Installation

The only requirement for running *SerraLINE* is a FORTRAN compiler. The program can be compiled on a terminal by typing:

```
$ make all
```

This will produce the two executables. They can also be compiled separately:

```
$ make SerraLINE
$ make Extract
```

SerraLINE

It is the main program that processes the trajectory of the NA molecular contour and calculates the bending angles at all sub-fragment lengths. It is executed by typing:

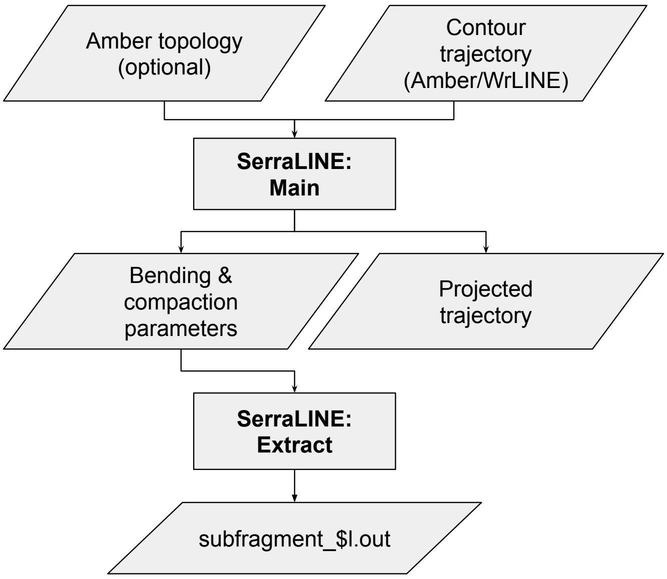

Amber topology (optional)
Contour trajectory (Amber/WrLINE)
SerraLINE: Main
Bending & compaction parameters
Projected trajectory
SerraLINE: Extract
subfragment_$l.out

Figure 1: General workflow of **SerraLINE**.

```
$./SerraLINE < SerraLINE.in
```

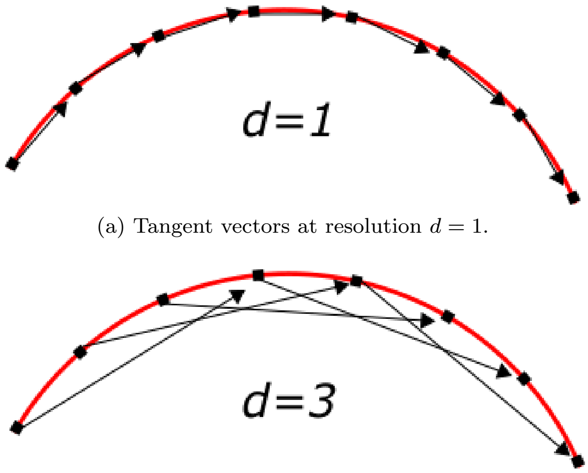
A trajectory file in AMBER (*crd or *x) or WrLINE (*xyz or *3col) style formats is needed. For sequence specificity, a topology file in AMBER format (*prmtop) can be optionally added, where *SerraLINE* will only read the number of bases and sequence, ignoring any other residue.

SerraLINE.in is the input file that indicates:

1. If the structure is opened (typing ”0”) or if it is closed (typing ”1”).
2. In case the topology file is included, if the structure is double-stranded (typing ”2”) or if it is single-stranded (typing ”1”).
3. If a topology file is included (typing ”1”) or not (typing ”0”).
4. If a topology file is not included, the number of bp of the molecule.
5. If the structure is going to be projected to the best fitted plane (typing ”1”, Figure 3), or if it is going to be projected to a region of points (typing a region ’x:y’ greater than 3 points, Figure 4) or if the structure is not going to be projected (typing ”0”).
6. The resolution of the tangent vectors (typing the resolution ”d”, Figure 2).
7. The path for the topology (Optional).
8. The path for the trajectory (Essential).
9. If it is desired to write the coordinates (in *xyz or *crd format) in case the projection method was used (typing ”1”).

The program generates one output:

- **SerraLINE.out** contains bending angles calculated from all possible sub-fragment lengths. If the projection method was used, width, height, aspect ratio and



(a) Tangent vectors at resolution $d = 1$.

(b) Tangent vectors at resolution $d = 3$.

Figure 2: Representation of tangent vectors (black arrows) at different resolutions d, calculated from elements (black squares) that compuse the molecular contour (red curve).

deviation from planarity are printed at the top of the file. *SerraLINE* calculates one of these quantities for each frame and outputs averages and standard deviations. For deviation from planarity parameters, *SerraLINE* prints the distance from the plane averaged along the molecule and the point that is farther apart. It outputs the absolute distance in Å and the relative distance with respect to the height in %.

All distance parameters are printed in Å and angles in degrees.

Extract

Extract program extracts the bending profiles along the molecule at a particular sublength, from the *SerraLINE* output **SerraLINE.out** , creating simple files ready to plot. You can filter a particular sublength to produce plots similar to Figure 5. For running it type:

```
$./Extract < extract.in
```

extract.in is the input file that indicates:

1. Path to SerraLINE output file.
2. The length (l) you want to process, which should be $0 < l < N - 1$

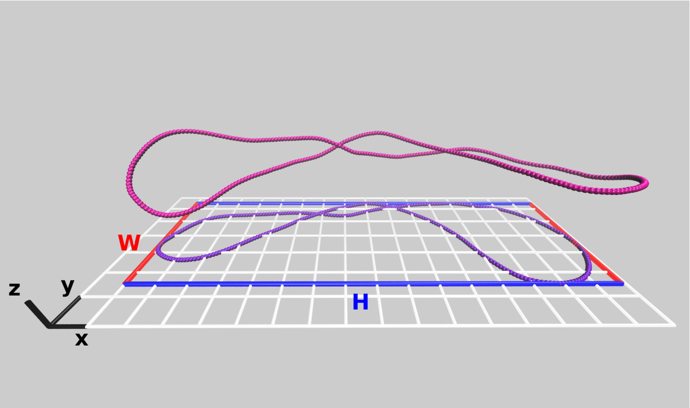


Figure 3: Global plane fitting. A plane is fitted to the whole 3D structure (in magenta) and projected to a plane, where width (H) and height (H) can be calculated.

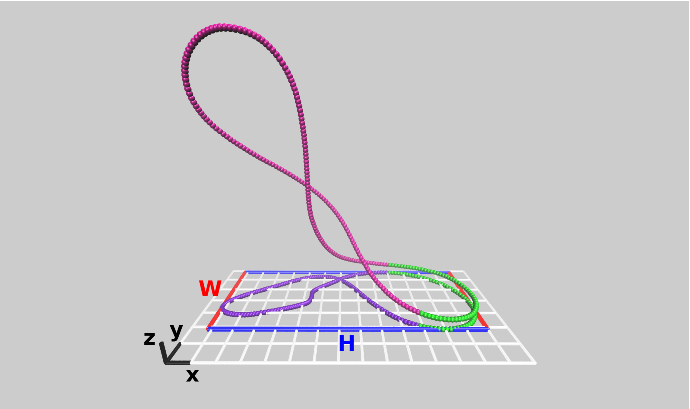


Figure 4: Specific region plane fitting. A plane is fitted to a particular region (green) from the 3D structure.

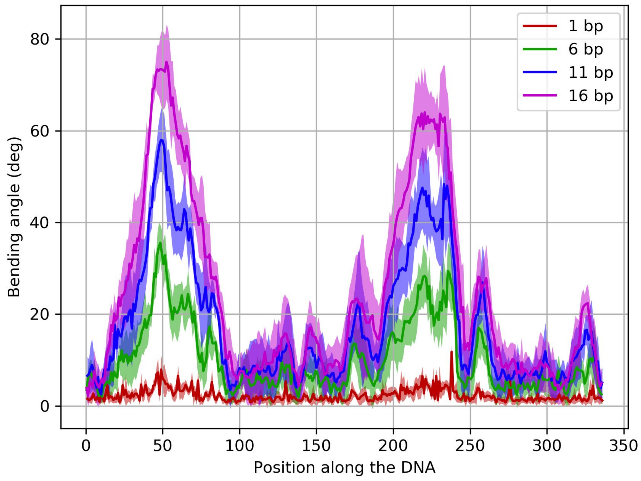


Figure 5: Bending profiles at the lengths of 1, 6, 11 and 16 bp.

Abbreviations

A adenine. 22

ABC Ascona B-DNA. 34

AFM atomic force microscopy. 30

BPP base-pair parameters. 45

BSP base-step parameters. 45

C cytosine. 22

CEHS Cambridge University Engineering Department Helix computation Scheme. 33

DNA deoxyribonucleic acid. 22

FRET fluorescence resonance energy transfer. 29

G guanine. 22

LDEM length-dependent elastic model. 35

Lk linking number. 28

MBT mid-base triad. 47

MC Monte Carlo. 34

MD molecular dynamics. 29

MS Marko and Siggia. 31

MST mid-step triad. 49

NA nucleic acids. 22

NMR nuclear magnetic resonance spectroscopy. 34

PCA principal component analysis. 71

PDB Protein Data Bank. 9, 25

RMSD root-mean square. 72

RNA ribonucleic acid. 22

SAXS small-angle x-ray scattering. 29

SVD singular value decomposition. 69

T thymine. 22

Tw twist. 28

TWLC twistable wormlike chain. 31

U uracil. 22

WLC wormlike chain. 30

Wr writhe. 28

Bibliography

[1] Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., and Harrison, S. C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science 242*, 4880 (1988), 899–907.

[2] Ahn, S. J. *Least squares orthogonal distance fitting of curves and surfaces in space*, vol. 3151. Springer Science & Business Media, 2004.

[3] Altman, D. G., and Bland, J. M. Standard deviations and standard errors. *BMJ 331*, 7521 (2005), 903.

[4] Balaceanu, A., Buitrago, D., Walther, J., Hospital, A., Dans, P. D., and Orozco, M. Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res. 47*, 9 (2019), 4418–4430.

[5] Balaceanu, A., Pérez, A., Dans, P. D., and Orozco, M. Allosterism and signal transfer in DNA. *Nucleic Acids Res. 46*, 15 (2018), 7554–7565.

[6] Balasubramanian, S., Xu, F., and Olson, W. K. DNA sequence-directed organization of chromatin: Structure-based computational analysis of nucleosome-binding sequences. *Biophys. J. 96*, 6 (2009), 2245–2260.

[7] Bao, L., Zhang, X., Shi, Y.-Z., Wu, Y.-Y., and Tan, Z.-J. Understanding the relative flexibility of RNA and DNA duplexes: Stretching and twist-stretch coupling. *Biophys. J. 112*, 6 (2017), 1094 – 1104.

[8] Bates, A. D., Noy, A., Piperakis, M. M., Harris, S. A., and Maxwell, A. Small DNA circles as probes of DNA topology. *Biochemical Society Transactions 41*, 2 (03 2013), 565–570.

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

[10] Bednar, J., Furrer, P., Katritch, V., Stasiak, A., Dubochet, J., and Stasiak, A. Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J. Mol. Biol. 254*, 4 (1995), 579 – 594.

[11] Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem. 91*, 24 (1987), 6269–6271.

[12] Bronner, S., and Shippen, J. Biomechanical metrics of aesthetic perception in dance. *Experimental brain research 233* (2015), 3565–3581.

[13] Bryant, Z., Stone, M. D., Gore, J., Smith, S. B., Cozzarelli, N. R., and Bustamante, C. Structural transitions and elasticity from torque measurements on DNA. *Nature 424*, 6946 (2003), 338–341.

[14] Calladine, C., Drew, H., Luisi, B., and Travers, A. *Understanding DNA: the molecule and how it works*, third ed. San Diego, etc., Elsevier Academic Press, March 2004.

[15] Caraglio, M., Skoruppa, E., and Carlon, E. Overtwisting induces polygonal shapes in bent DNA. *J. Chem. Phys. 150*, 13 (2019), 135101.

[16] Case, D., Betz, R., Cerutti, D., Cheatham, T., Darden, T., Duke, R., Giese, T., Gohlke, H., Götz, A., Homeyer, N., Izadi, S., Janowski, P., Kaus, J., Kovalenko, A., Lee, T.-S., LeGrand, S., Li, P., Lin, C., Luchko, T., and Kollman, P. Amber 16, University of California, San Francisco., 04 2016.

[17] Case, D., Darden, T., Cheatham, T., Simmerling, C., Wang, J., Duke, R., Luo, R., Walker, R., Zhang, W., Merz, K., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K., Paesani, F., Vanícek, J., Liu, J., and Roberts, B. Amber 11, University of California, San Francisco, 03 2010.

[18] Chatterjee, S., and Hadi, A. *Regression Analysis by Example.* Wiley Series in Probability and Statistics. Wiley, 2006.

[19] Chong, S., Chen, C., Ge, H., and Xie, X. Mechanism of transcriptional bursting in bacteria. *Cell 158*, 2 (2014), 314–326.

[20] Corless, S., and Gilbert, N. Effects of DNA supercoiling on chromatin architecture. *Biophysical Reviews 8* (2016), 245–258.

[21] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Kenneth M. Merz, J., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc. 17* (1995), 5179–5197.

[22] Crick, F. Central dogma of molecular biology. *Nature 227* (1970), 561–563.

[23] Da Rosa, G., Grille, L., Calzada, V., Ahmad, K., Arcon, J. P., Battistini, F., Bayarri, G., Bishop, T., Carloni, P., Cheatham Iii, T., et al. Sequence-dependent structural properties of b-dna: what have we learned in 40 years? *Biophysical Reviews*, 1–11.

[24] Dang, L. X. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: A molecular dynamics study. *J. Am. Chem. Soc. 117*, 26 (1995), 6954–6960.

[25] Dans, P. D., Balaceanu, A., Pasi, M., Patelli, A. S., Petkevičiūtė, D., Walther, J., Hospital, A., Bayarri, G., Lavery, R., Maddocks, J. H., and Orozco, M. The static and dynamic structural heterogeneities

of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res. 47*, 21 (10 2019), 11090–11102.

[26] Dans, P. D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R., and Orozco, M. Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res. 42*, 18 (2014), 11304–11320.

[27] Demurtas, D., Amzallag, A., Rawdon, E. J., Maddocks, J. H., Dubochet, J., and Stasiak, A. Bending modes of DNA directly addressed by cryo-electron microscopy of DNA minicircles. *Nucleic Acids Res. 37*, 9 (2009), 2882–2893.

[28] De Bruin, L., and Maddocks, J. H. cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res. 46*, W1 (06 2018), W5–W10.

[29] Dickerson, R. Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Res. 17*, 5 (03 1989), 1797–1803.

[30] Dohnalová, H., Dršata, T., Šponer, J., Zacharias, M., Lipfert, J., and Lankaš, F. Compensatory mechanisms in temperature dependence of DNA double helical structure: Bending and elongation. *J. Chem. Theor. Computation 16* (2020), 2857–2863.

[31] Dorman, C. DNA supercoiling and transcription in bacteria: a two-way street. *BMC Molecular and Cell Biology 20* (07 2019).

[32] Draper, N., and Smith, H. *Applied Regression Analysis.* Wiley Series in Probability and Statistics. Wiley, 1998.

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

[34] Dršata, T., Špačková, N., Jurečka, P., Zgarbová, M., Šponer, J., and Lankaš, F. Mechanical properties of symmetric and asymmetric DNA a-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Res. 42*, 11 (2014), 7383–7394.

[35] Du, Q., Kotlyar, A., and Vologodskii, A. Kinking the double helix by bending deformation. *Nucleic Acids Res. 36*, 4 (2008), 1120–1128.

[36] Dufrêne, Y. F., Ando, T., Garcia, R., Alsteens, D., Martinez-Martin, D., Engel, A., Gerber, C., and Müller, D. J. Imaging modes of atomic force microscopy for application in molecular and cell biology. *Nat. Nanotechnol. 12*, 4 (2017), 295–307.

[37] Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. The gcn4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: Crystal structure of the protein-DNA complex. *Cell 71*, 7 (1992), 1223–1237.

[38] Eslami-Mossallam, B., Schiessel, H., and van Noort, J. Nucleosome dynamics: Sequence matters. *Adv. Colloid Interface Sci. 232* (2016), 101–113. Proceedings from the International Workshop on Polyelectrolytes in Chemistry, Biology and Technology.

[39] F, B., A, H., D, B., D, G., PD, D., JL, G., and M., O. How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. *J Mol Biol 431*, 19 (09 2019), 3845–3859.

[40] Franklin, R. E., and Gosling, R. G. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature 172*, 4369 (1953), 156–157.

[41] FRANKLIN, R. E., and GOSLING, R. G. Molecular configuration in sodium thymonucleate. *Nature 171* (1953), 740–741.

[42] Fujimoto, B. S., and Schurr, J. M. Dependence of the torsional rigidity of DNA on base composition. *Nature 344* (1990), 175–178.

[43] Fuller, W., Wilkins, W., Wilson, H., and Hamilton, L. The molecular configuration of deoxyribonucleic acid: Iv. x-ray diffraction study of the a form. *J. Mol. Biol. 12*, 1 (1965), 60–IN9.

[44] Geggier, S., and Vologodskii, A. Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. U.S.A. 107*, 35 (2010), 15421–15426.

[45] Gibcus, J. H., and Dekker, J. The hierarchy of the 3D genome. *Mol. Cell 49* (2013), 773–782.

[46] Golub, G. H., and Reinsch, C. Singular value decomposition and least squares solutions. In *Linear algebra*. Springer, 1971, pp. 134–151.

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

[48] Gore, J., Bryant, Z., Nöllmann, M., Le, M. U., Cozzarelli, N. R., and Bustamante, C. DNA overwinds when stretched. *Nature 442* (August 2006).

[49] Gross, P., Laurens, N., Oddershede, L., Bockelmann, U., Peterman, E., and Wuite, G. Quantifying how DNA stretches, melts and changes twist under tension. *Nat. Phys. 7*, 9 (2011), 731–736.

[50] Gross, P., Laurens, N., Oddershede, L. B., Bockelmann, U., Peterman, E. J. G., and Wuite, G. J. L. Quantifying how DNA stretches, melts and changes twist under tension. *Nat. Phys. 7*, 9 (2011), 731–736.

[51] Gu, C., Zhang, J., Yang, Y. I., Chen, X., Ge, H., Sun, Y., Su, X., Yang, L., Xie, S., and Gao, Y. Q. DNA structural correlation in short and long ranges. *J. Phys. Chem. B 119*, 44 (2015), 13980–13990.

[52] Guilbaud, S., Salomé, L., Destainville, N., Manghi, M., and Tardin, C. Dependence of DNA persistence length on ionic strength and ion type. *Phys. Rev. Lett. 122* (2019), 028102.

[53] Haeusler, A. R., Goodson, K. A., Lillian, T. D., Wang, X., Goyal, S., Perkins, N. C., and Kahn, J. D. FRET studies of a landscape of Lac repressor-mediated DNA loops. *Nucleic Acids Res. 40*, 10 (2012), 4432–4445.

[54] Heard, W. *Rigid Body Mechanics: Mathematics, Physics and Applications.* Physics textbook. Wiley, 2008.

[55] Heath, P. J., Clendenning, J. B., Fujimoto, B. S., and Schurr, M. J. Effect of bending strain on the torsion elastic constant of DNA. *J. Mol. Biol. 260*, 5 (1996), 718 – 730.

[56] Herrero-Galán, E., Fuentes-Pérez, M. E., Carrasco, C., Valpuesta, J. M., Carrascosa, J. L., Moreno-Herrero, F., and Arias-González, J. R. Mechanical identities of RNA and DNA double helices unveiled at the single-molecule level. *J. Am. Chem. Soc. 135*, 1 (2013), 122–131.

[57] Higgins, N. P., and Vologodskii, A. V. Topological behavior of plasmid DNA. *Microbiol. Spectr. 3*, 2 (2015), 3–2.

[58] Horn, B. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. B 4* (04 1987), 629–642.

[59] Hospital, A., Andrio, P., Cugnasco, C., Codo, L., Becerra, Y., Dans, P. D., Battistini, F., Torres, J., Goñi, R., Orozco, M., and Gelpí, J. L. BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res. 44*, D1 (11 2015), D272–D278.

[60] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering 9*, 3 (2007), 90–95.

[61] III, T. E. C., Cieplak, P., and Kollman, P. A. A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn. 16*, 4 (1999), 845–862. PMID: 10217454.

[62] Irobalieva, R. N., Fogg, J. M., Catanese, D. J., Sutthibutpong, T., Chen, M., Barker, A. K., Ludtke, S. J., Harris, S. A., Schmid, M. F., Chiu, W., and Zechiedrich, L. Structural diversity of supercoiled DNA. *Nat. Commun. 6*, 1 (2015), 8440.

[63] Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., Portella, G., Battistini, F., Gelpí, J. L., González, C., Vendruscolo, M., Laughton, C. A., Harris, S. A., Case, D. A., and Orozco, M. Parmbsc1: A refined force field for DNA simulations. *Nat. Methods 13*, 1 (2015), 55–58.

[64] Janićijević, A., Sugasawa, K., Shimizu, Y., Hanaoka, F., Wijgers, N., Djurica, M., Hoeijmakers, J. H., and Wyman, C. DNA bending by the human damage recognition complex xpc–hr23b. *DNA Repair 2*, 3 (2003), 325–336.

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

[66] Keller, W., König, P., and Richmond, T. J. Crystal structure of a bZIP/DNA complex at 2.2 Å: Determinants of DNA specific recognition. *J. Mol. Biol. 254*, 4 (1995), 657–667.

[67] Kelley, C. T. *Iterative methods for optimization.* SIAM, 1999.

[68] Klug, A. Rosalind franklin and the discovery of the structure of DNA. *Nature 219*, 5156 (1968), 808–810.

[69] Krepl, M., Zgarbová, M., Stadlbauer, P., Otyepka, M., Banáš, P., Koča, J., Cheatham, T. E., Jurečka, P., and Šponer, J. Reference simulations of noncanonical nucleic acids with different χ variants of the amber force field: Quadruplex DNA, quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput. 8*, 7 (2012), 2506–2520. PMID: 23197943.

[70] Ku, H. H., et al. Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards 70*, 4 (1966), 263–273.

[71] Kulkarni, M., and Mukherjee, A. Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition. *Progress in biophysics and molecular biology 128* (2017), 63–73.

[72] König, P., and Richmond, T. J. The x-ray structure of the GCN4-bZIP bound to atf/creb site DNA shows the complex depends on DNA flexibility. *J. Mol. Biol. 233*, 1 (1993), 139–154.

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

[74] Lankaš, F., Šponer, J., Langowski, J., and Cheatham, T. E. DNA base-pair step deformability inferred from molecular dynamics simulations. *Biophys. J. 85* (2003), 2872–2883.

[75] Lankaš, F., Lavery, R., and Maddocks, J. H. Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure 14*, 10 (2006), 1527–1534.

[76] Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D., and Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res. 37* (2014), 5917–5929.

[77] Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, Thomas, I., Dixit, S., Jayaram, B., Lankas, F., Laughton, C., Maddocks, J. H., Michon, A., Osman, R., Orozco, M., Perez, A., Singh, T., Spackova, N., and Sponer, J. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res. 38*, 1 (11 2009), 299–313.

[78] Laxmikanthan, G., Xu, C., Brilot, A. F., Warren, D., Steele, L., Seah, N., Tong, W., Grigorieff, N., Landy, A., and Van Duyne, G. D. Structure of a holliday junction complex reveals mechanisms governing a highly regulated DNA transaction. *Elife 5* (2016), e14313.

[79] Leslie, A., Arnott, S., Chandrasekaran, R., and Ratliff, R. Polymorphism of DNA double helices. *Journal of molecular biology 143*, 1 (1980), 49–72.

[80] Li, J., Sagendorf, J. M., Chiu, T.-P., Pasi, M., Pérez, A., and Rohs, R. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res. 45*, 22 (2017), 12877–12887.

[81] Liebl, K., Drsata, T., Lankas, F., Lipfert, J., and Zacharias, M. Explaining the striking difference in twist-stretch coupling between DNA and rna: A comparative molecular dynamics analysis. *Nucleic Acids Res. 43*, 21 (2015), 10143–10156.

[82] Liebl, K., and Zacharias, M. Unwinding induced melting of double-stranded DNA studied by free energy simulations. *J. Phys. Chem. B 121*, 49 (2017), 11019–11030.

[83] Lionberger, T. A., Demurtas, D., Witz, G., Dorier, J., Lillian, T., Meyhöfer, E., and Stasiak, A. Cooperative kinking at distant sites in mechanically stressed DNA. *Nucleic Acids Res. 39*, 22 (09 2011), 9820–9832.

[84] Lionnet, T., Joubaud, S., Lavery, R., Bensimon, D., and Croquette, V. Wringing out DNA. *Phys. Rev. Lett. 96* (May 2006), 178102.

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

[86] Lipfert, J., Skinner, G. M., Keegstra, J. M., Hensgens, T., Jager, T., Dulin, D., Köber, M., Yu, Z., Donkers, S. P., Chou, F.-C., Das, R., and Dekker, N. H. Double-stranded RNA under force and torque: Similarities to and striking differences from double-stranded DNA. *Proc. Natl. Acad. Sci. 111*, 43 (2014), 15408–15413.

[87] Lipfert, J., Wiggin, M., Kerssemakers, J., Pedaci, F., and Dekker, N. Corrigendum: Freely orbiting magnetic tweezers to directly monitor changes in the twist of nucleic acids. *Nat. Commun. 2* (08 2011), 439.

[88] Liu, Y., Bondarenko, V., Ninfa, A., and Studitsky, V. M. DNA supercoiling allows enhancer action over a large distance. *Proc. Natl. Acad. Sci. 98*, 26 (2001), 14883–14888.

[89] Lodge, J., Kazic, T., and Berg, D. Formation of supercoiling domains in plasmid pBR322. *J. Bacteriol. 171*, 4 (1989), 2181–2187.

[90] Lu, X., and Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three dimensional nucleic acid structures. *Nucleic Acids Res. 31*, 17 (2003), 5108–5121.

[91] Lu, X.-J., Hassan, M. E., and Hunter, C. Structure and conformation of helical nucleic acids: analysis program (schnaap)11edited by k. nagai. *J. Mol. Biol. 273*, 3 (1997), 668 – 680.

[92] Lu, X.-J., Hassan, M. E., and Hunter, C. Structure and conformation of helical nucleic acids: rebuilding program (schnarp)11edited by k. nagai. *J. Mol. Biol. 273*, 3 (1997), 681 – 691.

[93] Maehigashi, T., Hsiao, C., Kruger Woods, K., Moulaei, T., Hud, N. V., and Dean Williams, L. B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res. 40*, 8 (12 2011), 3714–3722.

[94] Marin-Gonzalez, A., Pastrana, C. L., Bocanegra, R., Martín-González, A., Vilhena, J. G., Pérez, R., Ibarra, B., Aicart-Ramos, C., and Moreno-Herrero, F. Understanding the paradoxical mechanical response of in-phase A-tracts at different force regimes. *Nucleic Acids Res.* (04 2020), 5024–5036.

[95] Marín-González, A., Vilhena, J. G., Moreno-Herrero, F., and Pérez, R. DNA crookedness regulates DNA mechanical properties at short length scales. *Phys. Rev. Lett. 122* (2019), 048102.

[96] Marin-Gonzalez, A., Vilhena, J. G., Perez, R., and Moreno-Herrero, F. Understanding the mechanical response of double-stranded DNA and RNA under constant stretching forces using all-atom molecular dynamics. *Proc. Natl. Acad. Sci. 114*, 27 (2017), 7049–7054.

[97] Marko, J. F. Torque and dynamics of linking number relaxation in stretched supercoiled DNA. *Phys. Rev. E 76*, 2 (2007), 021926.

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

[99] Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics 11*, 2 (1963), 431–441.

[100] Matek, C., Ouldridge, T. E., Doye, J. P., and Louis, A. A. Plectoneme tip bubbles: coupled denaturation and writhing in supercoiled DNA. *Sci. Rep. 5*, 1 (2015), 1–4.

[101] Mathew-Fenn, R. S., Das, R., and Harbury, P. Remeasuring the double helix. *Science 322*, 5900 (2008), 446–449.

[102] Mathews, J. H., and Fink, K. D. *Numerical Methods Using Matlab, Third Edition.* Prentice Hall, 1999.

[103] Mazur, A. K., and Maaloum, M. Atomic force microscopy study of DNA flexibility on short length scales: smooth bending versus kinking. *Nucleic Acids Res. 42*, 22 (2014), 14006–14012.

[104] Mitchell, J. S., Glowacki, J., Grandchamp, A. E., Manning, R. S., and Maddocks, J. H. Sequence-dependent persistence lengths of DNA. *J. Chem. Theory Comput. 13*, 4 (2017), 1539–1555.

[105] Mitchell, J. S., Laughton, C. A., and Harris, S. A. Atomistic simulations reveal bubbles, kinks and wrinkles in supercoiled DNA. *Nucleic Acids Res. 39*, 9 (01 2011), 3928–3938.

[106] Mohammad-Rafiee, F., and Golestanian, R. Elastic correlations in nucleosomal DNA structure. *Phys. Rev. Lett. 94* (Jun 2005), 238102.

[107] Mosconi, F., Allemand, J. F. m. c., Bensimon, D., and Croquette, V. Measurement of the torque on a single stretched and twisted DNA using magnetic tweezers. *Phys. Rev. Lett. 102* (2009), 078301.

[108] Nelder, J. A., and Mead, R. A Simplex Method for Function Minimization. *The Computer Journal 7*, 4 (01 1965), 308–313.

[109] Nelson, P. C. *Biological Physics: Energy, Information, Life.* Clancy Marshall, 2014.

[110] Nomidis, S. K., Caraglio, M., Laleman, M., Phillips, K., Skoruppa, E., and Carlon, E. Twist-bend coupling, twist waves, and the shape of DNA loops. *Phys. Rev. E 100* (Aug 2019), 022402.

[111] Nomidis, S. K., Kriegel, F., Vanderlinden, W., Lipfert, J., and Carlon, E. Twist-bend coupling and the torsional response of double-stranded DNA. *Phys. Rev. Lett. 118* (May 2017), 217801.

[112] Norouzi, D., Mohammad-Rafiee, F., and Golestanian, R. Effect of bending anisotropy on the 3d conformation of short DNA loops. *Phys. Rev. Lett. 101* (Oct 2008), 168103.

[113] Noy, A., and Golestanian, R. The Chirality of DNA: Elasticity Cross-Terms at Base-Pair Level Including A-Tracts and the Influence of Ionic Strength. *J. Phys. Chem. B 114* (2010), 8022–8031.

[114] Noy, A., and Golestanian, R. Length scale dependence of DNA mechanical properties. *Phys. Rev. Lett. 109* (Nov 2012), 228101.

[115] Noy, A., Maxwell, A., and Harris, S. A. Interference between triplex and protein binding to distal sites on supercoiled DNA. *Biophys. J. 112*, 3 (2017), 523–531.

[116] Noy, A., Pérez, A., Lankaš, F., Javier Luque, F., and Orozco, M. Relative flexibility of DNA and RNA: a molecular dynamics study. *J. Mol. Biol. 343* (2004), 627–638.

[117] Oberstrass, F. C., Fernandes, L. E., and Bryant, Z. Torque measurements reveal sequence-specific cooperative transitions in supercoiled DNA. *Proc. Natl. Acad. Sci. 109*, 16 (2012), 6106–6111.

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

[119] Pasi, M., Maddocks, J. H., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, III, T., Dans, P. D., Jayaram, B., Lankas, F., Laughton, C., Mitchell, J., Osman, R., Orozco, M., Pérez, A., Petkevičiūtė, D., Spackova, N., Sponer, J., Zakrzewska, K., and Lavery, R. µabc: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in b-DNA. *Nucleic Acids Res. 42*, 19 (2014), 12272–12283.

[120] Pasi, M., Zakrzewska, K., Maddocks, J. H., and Lavery, R. Analyzing DNA curvature and its impact on the ionic environment: application to molecular dynamics simulations of minicircles. *Nucleic Acids Res. 45*, 7 (2017), 4269–4277.

[121] Pérez, A., Lankaš, F., Luque, F. J., and Orozco, M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res. 36* (2008), 2379–2374.

[122] Pérez, A., Marchán, I., Svozil, D., Šponer, J., Cheatham, T. E., Laughton, C. A., and Orozco, M. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J. 92*, 11 (2007), 3817–3829.

[123] Petkevičiūtė, D., Pasi, M., Gonzalez, O., and Maddocks, J. cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res. 42*, 20 (09 2014), e153–e153.

[124] Press, W., Teukolsky, S., Flannery, B., and Vetterling, W. *Numerical Recipes in FORTRAN 77: Volume 1, Volume 1 of Fortran Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 1992.

[125] Pyne, A. L. B., Noy, A., Main, K., Velasco-Berrelleza, V., Piperakis, M. M., Mitchenall, L., Cugliandolo, F., Beton, J. G., Stevenson, C., Hoogenboom, B., Bates, A., Maxwell, A., and Harris, S. Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides. *Nat. Commun. 12* (2021).

[126] Pérez, A., Luque, F. J., and Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc. 129*, 47 (2007), 14739–14745. PMID: 17985896.

[127] Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A., and Segal, E. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet. 44* (2012), 743–750.

[128] Rencher, A., and Schaalje, G. *Linear Models in Statistics.* Wiley, 2008.

[129] Rice, P. A., wei Yang, S., Mizuuchi, K., and Nash, H. A. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell 87*, 7 (1996), 1295–1306.

[130] Riethoven, J.-J. M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Computational biology of transcription factor binding* (2010), 33–42.

[131] Roe, D. R., and Cheatham, T. E. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput. 9*, 7 (2013), 3084–3095. PMID: 26583988.

[132] Roy, R., Hohng, S., and Ha, T. A practical guide to single-molecule fret. *Nat. Methods 5*, 6 (2008), 507–516.

[133] Scott, S., Xu, Z., Kouzine, F., Berard, D. J., Shaheen, C., Gravel, B., Saunders, L., Hofkirchner, A., Leroux, C., Laurin, J., Levens, D., Benham, C. J., and Leslie, S. R. Visualizing structure-mediated interactions in supercoiled DNA molecules. *Nucleic Acids Res. 46*, 9 (04 2018), 4622–4631.

[134] Sheinin, M. Y., and Wang, M. D. Twist-stretch coupling and phase transition during DNA supercoiling. *Phys. Chem. Chem. Phys. 11* (2009), 4800–4803.

[135] Shepherd, J. W., Greenall, R. J., Probert, M., Noy, A., and Leake, M. The emergence of sequence-dependent structural motifs in stretched, torsionally constrained DNA. *Nucleic Acids Res. 48*, 4 (01 2020), 1748–1763.

[136] Shi, X., Herschlag, D., and Harbury, P. A. B. Structural ensemble and microscopic elasticity of freely diffusing DNA by direct measurement of fluctuations. *Proc. Natl. Acad. Sci. U.S.A. 110*, 16 (2013), E1444–E1451.

[137] Shimon, L. J., and Harrison, S. C. The phage 434 or2/r1-69 complex at 2·5 Å resolution. *J. Mol. Biol. 232*, 3 (1993), 826–838.

[138] Shkurti, A., Goni, R., Andrio, P., Breitmoser, E., Bethune, I., Orozco, M., and Laughton, C. A. pypcazip: A pca-based toolkit for compression and analysis of molecular simulation data. *SoftwareX 5* (2016), 44 – 50.

[139] Shrader, T. E., and Crothers, D. M. Artificial nucleosome positioning sequences. *Proc. Natl. Acad. Sci. U.S.A. 86*, 19 (1989), 7418–7422.

[140] Sinden, R. R. *DNA structure and function.* Gulf Professional Publishing, 1994.

[141] Skoruppa, E., Laleman, M., Nomidis, S. K., and Carlon, E. DNA elasticity from coarse-grained simulations: The effect of groove asymmetry. *J. Chem. Phys. 146*, 21 (2017), 214902.

[142] Skoruppa, E., Laleman, M., Nomidis, S. K., and Carlon, E. DNA elasticity from coarse-grained simulations: The effect of groove asymmetry. *J. Chem. Phys. 146*, 21 (2017), 214902.

[143] Skoruppa, E., Nomidis, S. K., Marko, J. F., and Carlon, E. Bend-induced twist waves and the structure of nucleosomal DNA. *Phys. Rev. Lett. 121* (Aug 2018), 088101.

[144] Skoruppa, E., Voorspoels, A., Vreede, J., and Carlon, E. Length-scale-dependent elasticity in DNA from coarse-grained and all-atom models. *Phys. Rev. E 103* (Apr 2021), 042408.

[145] Smale, S. T., and Kadonaga, J. T. The RNA polymerase II core promoter. *Annual review of biochemistry 72*, 1 (2003), 449–479.

[146] Smith, D. E., and Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys. 100* (1994), 3757–3766.

[147] Smith, S. B., Cui, Y., and Bustamante, C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science 271* (1996), 795–798.

[148] Snodin, B. E. K., Randisi, F., Mosayebi, M., Šulc, P., Schreck, J. S., Romano, F., Ouldridge, T. E., Tsukanov, R., Nir, E., Louis, A. A., and Doye, J. P. K. Introducing improved structural properties and salt dependence into a coarse-grained model of DNA. *J. Chem. Phys. 142*, 23 (2015), 234901.

[149] Sutthibutpong, T., Harris, S. A., and Noy, A. Comparison of molecular contours for measuring writhe in atomistic supercoiled DNA. *J. Chem. Theory Comput. 11*, 6 (2015), 2768–2775.

[150] Sutthibutpong, T., Harris, S. A., and Noy, A. Comparison of molecular contours for measuring writhe in atomistic supercoiled DNA. *J. Chem. Theory Comput. 11*, 6 (2015), 2768–2775. PMID: 26575569.

[151] Sutthibutpong, T., Matek, C., Benham, C., Slade, G. G., Noy, A., Laughton, C., K. Doye, J. P., Louis, A. A., and Harris, S. A. Long-range correlations in the mechanics of small DNA circles under topological stress revealed by multi-scale simulation. *Nucleic Acids Res. 44*, 19 (09 2016), 9121–9130.

[152] Sponer, J., Banas, P., Jurecka, P., Zgarbová, M., Kuhrova, P., Havrila, M., Krepl, M., Stadlbauer, P., and Otyepka, M. Molecular dynamics simulations of nucleic acids. from tetranucleotides to the ribosome. *J. Phys. Chem. Lett. 5*, 10 (2014), 1771–1782.

[153] Theodorakopoulos, N. *Statistical Physics Of DNA: An Introduction To Melting, Unzipping And Flexibility Of The Double Helix.* World Scientific Publishing Company, 2019.

[154] Thode, H. *Testing For Normality.* Statistics, textbooks and monographs. CRC Press, 2002.

[155] Tolstorukov, M., Virnik, K., Adhya, S., and Zhurkin, V. A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res. 33* (02 2005), 3907–18.

[156] Tsui, V., and Case, D. A. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers 56*, 4 (2000), 275–291.

[157] Velasco-Berrelleza, V., Burman, M., Shepherd, J. W., Leake, M. C., Golestanian, R., and Noy, A. SerraNA: a program to determine nucleic acids elasticity from simulation data. *Phys. Chem. Chem. Phys. 22* (2020), 19254–19266.

[158] Vilfan, I. D., Lipfert, J., Koster, D., Lemay, S., and Dekker, N. Magnetic tweezers for single-molecule experiments. *Handbook of single-molecule biophysics* (2009), 371–395.

[159] Vinogradov, A. E., and Anatskaya, O. V. DNA helix: the importance of being at-rich. *Mammalian Genome 28* (2017), 455–464.

[160] Virstedt, J., Berge, T., Henderson, R. M., Waring, M. J., and Travers, A. A. The influence of DNA stiffness upon nucleosome formation. *J. Struct. Biol. 148*, 1 (2004), 66–85.

[161] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods 17* (2020), 261–272.

[162] Vologodskaia, M., and Vologodskii, A. Contribution of the intrinsic curvature to measured DNA persistence length11edited by i. tinoco. *J. Mol. Biol. 317*, 2 (2002), 205 – 213.

[163] Walther, J., Dans, P. D., Balaceanu, A., Hospital, A., Bayarri, G., and Orozco, M. A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res. 48*, 5 (01 2020), e29–e29.

[164] WATSON, J. D., and CRICK, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature 171* (April 1953), 737 – 738.

[165] Wenner, J. R., Williams, M. C., Rouzina, I., and Bloomfield, V. A. Salt dependence of the elasticity and overstretching transition of single DNA molecules. *Biophys. J. 82*, 6 (2002), 3160 – 3169.

[166] Whittaker, J. *Graphical Models in Applied Multivariate Statistics.* Wiley Publishing, 2009.

[167] Wiggins, P., van der Heijden, T., F. Moreno-Herrero, A. S., R. Phililips, J. Widom, C. D., and Nelson, P. C. High flexibility of DNA on short length scales probed by atomic force microscopy. *Nat. Nanotechnol. 1*, 137 (2006).

[168] Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. Crystal structure analysis of a complete turn of B-DNA. *Nature 287*, 5784 (1980), 755–758.

[169] Xiao, S., Liang, H., and Wales, D. J. The contribution of backbone electrostatic repulsion to DNA mechanical properties is length-scale-dependent. *J. Phys. Chem. Let. 10*, 17 (2019), 4829–4835.

[170] Yoshua, S. B., Watson, G. D., Howard, J. A., Velasco-Berrelleza, V., Leake, M. C., and Noy, A. Integration host factor bends and bridges DNA in a multiplicity of binding modes with varying specificity. *Nucleic Acids Res. 49*, 15 (2021), 8684–8698.

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

[172] Zechiedrich, E., Khodursky, A., Bachellier, S., Schneider, R., Chen, D., Lilley, D., and Cozzarelli, N. Roles of topoisomerases in maintaining steady-state DNA supercoiling in escherichia coli. *J. Biol. Chem. 275*, 11 (March 2000), 8103—8113.

[173] Zgarbová, M., Luque, F. J., Šponer, J., Cheatham, T. E., Otyepka, M., and Jurečka, P. Toward improved description of DNA backbone: Revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput. 9*, 5 (2013), 2339–2354.

[174] Zgarbová, M., Šponer, J., Otyepka, M., Cheatham, T. E., Galindo-Murillo, R., and Jurečka, P. Refinement of the sugar–phosphate backbone torsion beta for amber force fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput. 11*, 12 (2015), 5723–5736.

[175] Zgarbová, M., Luque, F. J., Šponer, J., Cheatham, T. E., Otyepka, M., and Jurečka, P. Toward improved description of DNA backbone: Revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput. 9*, 5 (2013), 2339–2354. PMID: 24058302.

[176] Zhang, Y., and Crothers, D. M. High-throughput approach for detection of DNA bending and flexibility based on cyclization. *Proc. Natl. Acad. Sci. U.S.A. 100*, 6 (2003), 3161–3166.

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