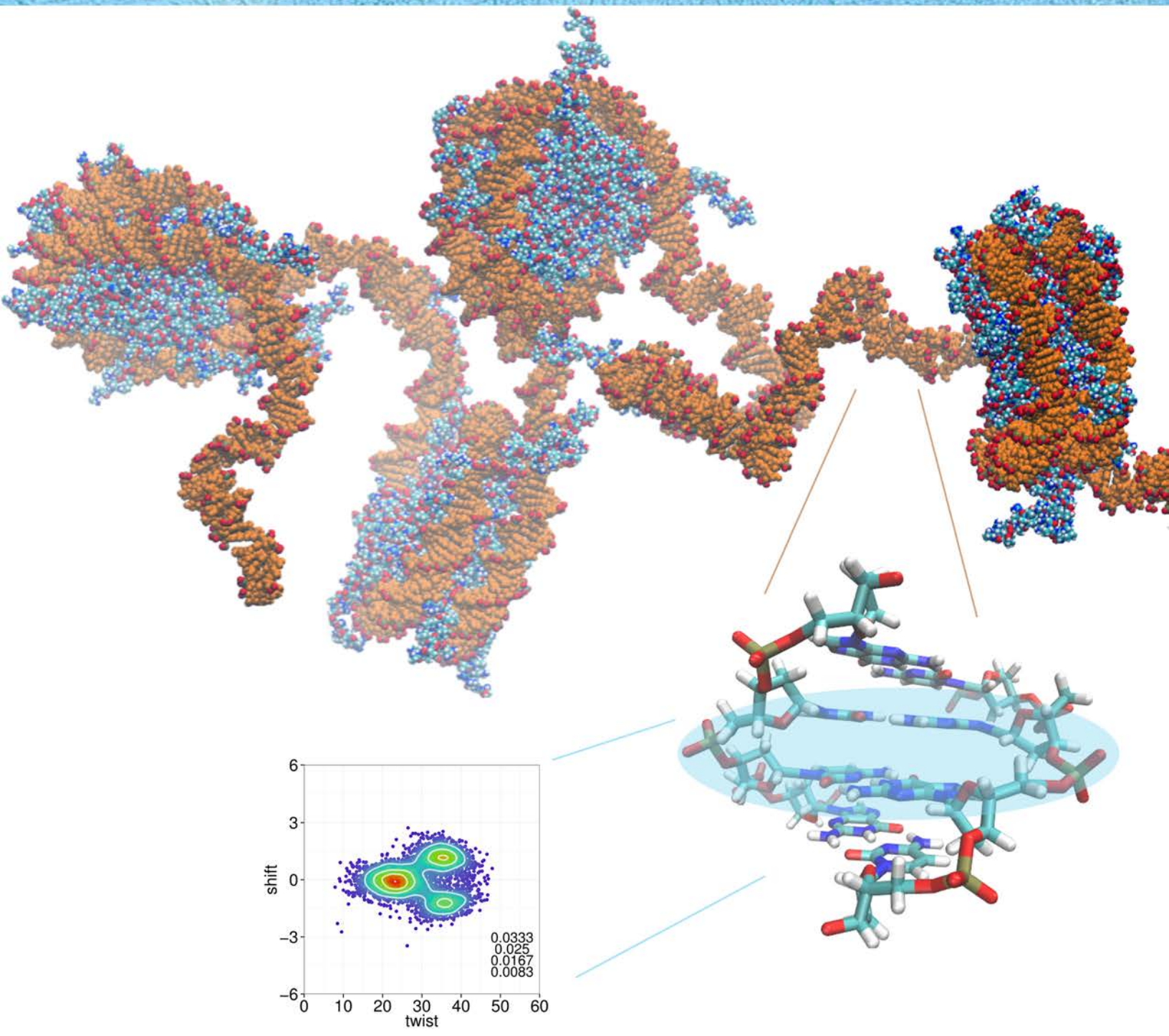
Revealing DNA dynamics from atomistic to genomic level by multiscale computational approaches

Jürgen Walther

Revealing DNA dynamics from atomistic to genomic level by multiscale computational approaches



Revealing DNA dynamics from atomistic to genomic level by multiscale computational approaches

Jürgen Walther

Doctoral programme: Física

Facultat de Física

Departament de Física de la Matèria Condensada

Universitat de Barcelona

UNIVERSITAT DE
BARCELONA

Revealing DNA dynamics from atomistic to genomic level by multiscale computational approaches

Programa de doctorat en Física

Autor/a: Jürgen Walther

Director/a: Dr. Modesto Orozco Lopez

Tutor/a: Dr. Giancarlo Franzese

UNIVERSITAT DE
BARCELONA

Acknowledgements

Many people accompanied me during my PhD years and it is difficult to keep track of all the people I met and who influenced me during my path of becoming a doctor. Of course there is my supervisor, Modesto Orozco, who always led me the right way in my research and from whom I could learn a lot of things. Many thanks to Pablo Dans, my tutor and guide through my PhD from time consuming supervision in the beginning to fruitful discussions of the results the last years. His influence was such that towards the end of my PhD I even started to produce acceptable figures. The multidisciplinary environment in the lab enabled me to learn a lot from the informatics side as well. There was Jose who inspired me a lot about state of the art coding and he was always there when I accidentally deleted some of my files. A special thanks to Adam, a great person from whom I learned so much on how to deal with Virtual Machines, to Genis, a very nice guy with whom I learned to integrate data in a web environment. It was a pleasure to set up many webservers with you. Many thanks to Isabelle, who always had an open ear and gave us lots of biological input from an experimentalist's point of view. Also many thanks to Alexandra and Diana with whom I could share many interesting discussions and fun trips to conferences. Even though the lab changed a lot during those years, the people who influenced me most were the ones who were there from (almost) the beginning…Ricard, my beer brewing "R guru" who patiently solved my for him trivial problems, Francesco, with whom I shared lab and waves, Hansel, who introduced me to the world of data science, Federica, who is still convinced as I am writing this that I will never finish my PhD, Ivan, the movie director whose wedding was as spectacular as his group meeting presentations, Juan, my 'hippie pijo', the shared passion about football made us teammates in our IRB football team all those years, Sanja, with whom I did many outreach projects together, Pedro, who gave me lots of advice in project and career choices even before I started my PhD. To Pablo Romero, Manuel Sarmiento, Felipe Cano, Osama Essarab and Eric Matamoros who I had the pleasure to tutor during their projects and I am sure all of you have a bright future ahead. I would like to thank the entire lab, but it changed so much that it is impossible for me to remember all the names. A special thanks goes to Richard Lavery and Marco Pasi with whom I had the pleasure to stay for one week in Lyon in their research group.

To the IRB football team (Salva, Alex, Juan, Ernest, Craig and Jordi to mention the core of the team) with who I had always lots of fun and success on and off the pitch as we managed to win title and cup one year, thanks to all my surfing and climbing buddies (Francesco, Ricardo, Joel, Sergí, Craig) and Fabian for fun tennis matches, sharing those moments together in all those years fills my heart with joy. A special thanks goes to Carla who was always there for me even in my darkest moments and without her I would still try to figure out how to format a large word document.

My last thanks goes to my parents who were always keen on knowing what I was working on and for the support along my way. Ihr wisst gar nicht wie glücklich ich bin, dass ihr immer für mich da seid. Die regelmäßgen Gespräche haben mich immer wieder auf den Boden der Tatsachen zurückgebracht und mich die Dinge klarer und entspannter sehen lassen.

Table of contents

Figures

Tables

OVERVIEW

The study of DNA from atomistic to mesoscopic level and connecting different resolution levels constitutes a major challenge since the new millennium. In the early 2000s, experiments could resolve for the first time the structure of the nucleosome in high detail or capture physical contacts in the genome of segments far apart in sequence. At around the same time, the force field development for atomistic nucleic acid simulations reached a peak with parmbsc0 in 2007 and coarse grain nucleosome fiber models emerged. The first decade ended with, to my opinion, the most remarkable experimental advance in visualizing the whole genome, Hi-C. In the current decade, almost ten years after Hi-C was invented, the structure of the cell nucleus is still a very hot topic. We can now harvest the fruits of the pioneers in the first decade of multi-scale investigation of DNA and connect the different resolution levels to obtain a complete picture of DNA from electron orbitals to genome folding.

In this work, we use computational approaches to dissect the different resolution levels, from atomistic MD simulations to mesoscopic secondary chromatin structure modeling. We developed a force-field for the accurate description of atomistic DNA dynamics based on quantum mechanical simulations. With the accuracy of parmbsc1, sequence-dependent effects of B-DNA beyond the base pair level were described and used as a starting point to parametrize a novel helical coarse grain model which shows similar accuracy to the DNA dynamics obtained by atomistic MD, but at much lower computational cost. In the nucleosome fiber model the coarse grain DNA algorithm is used for the linker DNA description and alongside with a simple mesoscopic characterization of the nucleosome chromatin dynamics can be probed at kilobase scale with a DNA model whose roots lie in the quantum mechanical regime.

On top of that, to meet current standards of accessibility and usability of tools, the developed coarse grain DNA and nucleosome fiber model are freely available as stand-alone versions or integrated in a single webserver or large-scale online research environment platform.

Thesis Organization

Because of the broad nature of topics covered in this dissertation I will first give a general overview of the topics shared among the studies presented here (**Chapter I**), from the structure of DNA to the organization of chromatin in the cell nucleus both from the experimental and theoretical point of view. In **Chapter II** those general concepts are expanded in more detail for better understanding of the results if required. The results section (**Chapter III**) is compiled of five publications (or in the process of publication). The first part focuses on discoveries on sequence-dependent properties of B-DNA and concomitant structural polymorphisms using molecular dynamics simulations with the state-of-the-art parmbsc1 force field. The second part of the results chapter takes advantage of all the information gathered about B-DNA and describes a new helical coarse grain model of B-DNA alongside with its implementation in a webserver. The last section of the results chapter deals with the development of a nucleosome fiber model, an extension of the helical coarse grain model, and the prediction of realistic chromatin conformations as they could possibly appear in the cell nucleus. A summary and a discussion of each result section is presented in **Chapter IV**, with the main conclusions at the end of this work.

CHAPTER I - INTRODUCTION

1. Basic principles of History and Structure of DNA

1.1 History of DNA

Far before the discovery of the DNA, Charles Darwin who already could describe the evolution of life after it started wondered *What is the essence of life?*. The famous physicist Erwin Schrödinger got one step closer to the solution by asking *How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?*. He assumed that essence of life had to be the information stored in a molecule, an "aperiodic crystal", where different molecular entities are connected by covalent chemical bonds. The discovery of the structure of DNA by Watson and Crick (1) was a scientific breakthrough for which they jointly received the Nobel Prize in 1962.

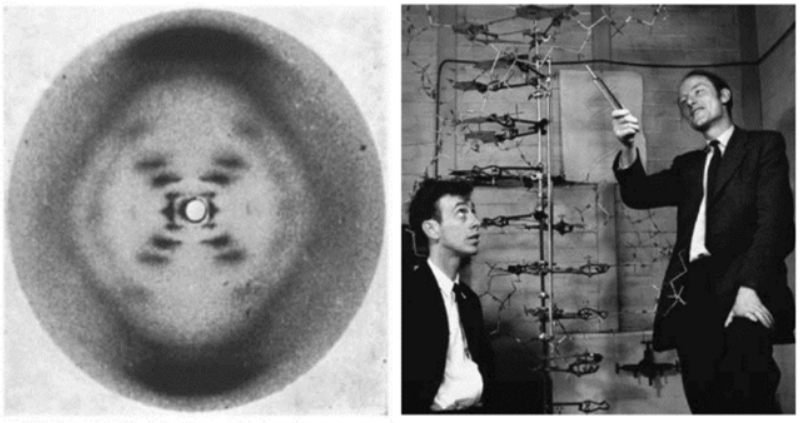


Figure 1. Discovery of DNA structure. X-ray diffraction image of DNA used for constructing the model (left) and Watson and Crick next to their model (right).

Based on the ideas ruled out by Schrödinger, they managed to construct the correct model of the DNA double-helix from X-ray diffraction images collected by Rosalind Franklin and Maurice Wilkins (see Figure 1). Their experimental findings satisfied previous experimental work by Erwin Chargaff who found 1:1 molar ratios of adenine:thymine and cytosine:guanine in DNA. Chargaff

also discovered different proportions of base composition among different species, which confirmed Schrödinger's hypothesis of the non-repetitive nature of DNA. Both Chargaff's experiments and the Watson-Crick structural model combined could explain basic principles of base-pairing and the genetic code alongside its replication and transcription. Later on, many more biological functions of DNA were unveiled (2): the process of DNA replication, the process of transcription into messenger RNA and subsequent translation into protein sequences, the compaction of DNA into chromatin and more recently the influence of sequence-dependent DNA properties and epigenetic marks on the dynamics of chromatin architecture (3, 4).

1.2 Structure of DNA

DNA is a long polymer which is comprised of repeating units called nucleotides coiled around each other to form two complementary strands. Each nucleotide consists of one of the four nitrogenous bases (adenine (A), guanine (G), cytosine (C) or thymine (T)) and a phosphate-deoxyribose segment comprising the backbone. The bases are planar aromatic heterocyclic molecules which are divided into two groups: purines (A,G) and pyrimidines (C,T). In the natural Watson-Crick base-pairing, a purine and a pyrimidine from each strand are held together by specific hydrogen bonds: adenine pairs with thymine and guanine pairs with cytosine (see Figure 2). The A-T base-pair is kept together by two hydrogen bonds while the G-C base-pair has three hydrogen bonds. The additional hydrogen bond makes the G-C base-pair more stable (around 1.5 kcal/mol, (5)). The nucleotides within one strand are connected via the backbone by phosphodiester bonds. Base-stacking interactions among aromatic nucleobases and hydrogen bonding between them are the main drivers of the stability of the DNA where stacking preferences and the physical properties of the sugar-phosphate backbone give each base pair step a slight twist of 35-36°, with bases nearly parallel to each other and an inter-base distance of 3.3-3.4 Å resulting in a double-helical staircase where ten base pairs form a helical turn.

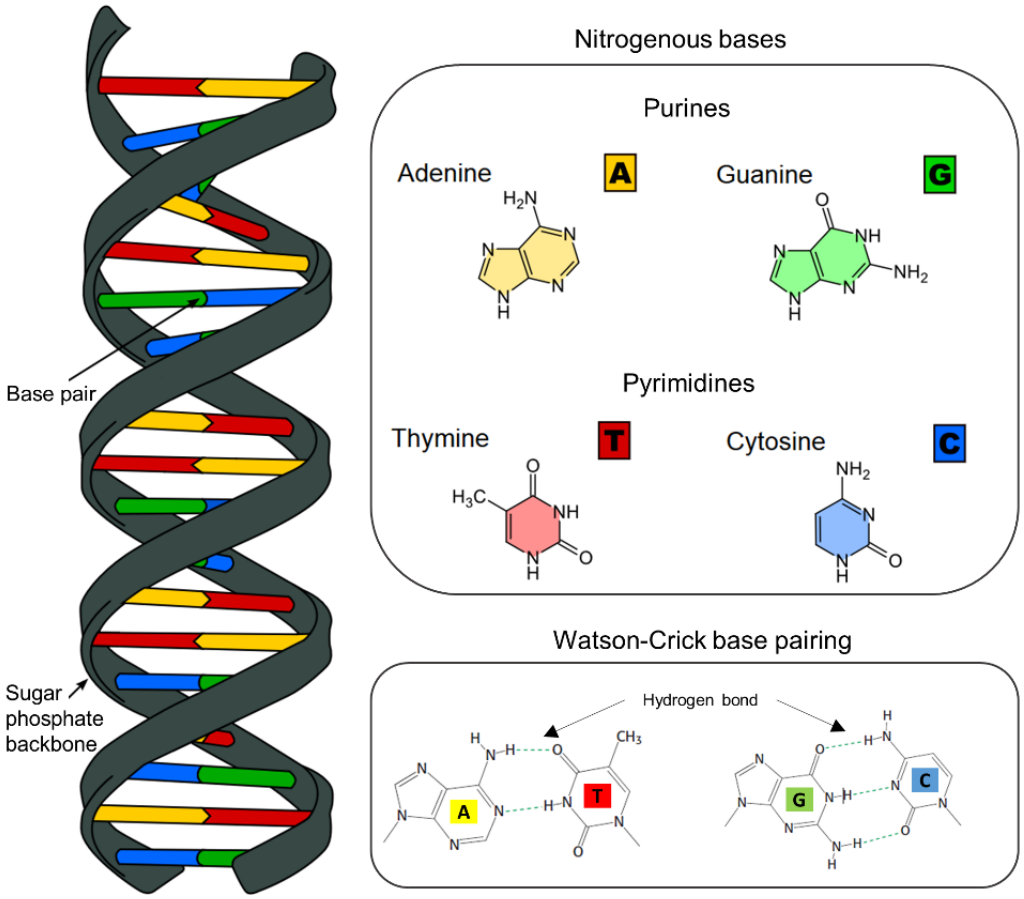


Figure 2. Structure of DNA double-helix. DNA double-helix with highlighted nucleotide units (left), the structure of the four nucleobases (top right) and the Watson-Crick base pairing between the bases (bottom right).

1.2.1 Helical parameters

From a structural point of view the canonical model for DNA allows an elegant description of local DNA dynamics by two types of movements at the base-pair level (intra base pair dynamics): translations and rotations with respect to the previous base pair in the helix. This simplicity of DNA dynamics gives rise to a set of geometrical descriptors of base morphology to describe DNA conformation. This set of rotational and translational parameters between bases and base-pairs was developed at the EMBO meeting in Cambridge in 1988 ("Cambridge Accord") and standardized at the Tsukuba Workshop in Nucleic Acid Structure and Interactions (6) by choosing a single reference frame to calculate base morphology parameters. Parameters are defined either

locally with respect to a local coordinate system attached to each individual base pair, or with respect to a global curvilinear helical axis (Figure 3).

Thus, to fully describe the orientation and position of the two rigid body bases of a base-pair, 3 translational and 3 rotational parameters are defined, referred to as intra base pair parameters:

- **Shear**, **stretch** and **stagger** are called the relative displacements of the bases along their x-, y- and helical axis (z-axis)

- **Buckle**, **propeller** twist and **opening** accordingly are the relative torsions of the base planes around their x-, y- and helical axis (z-axis)

The degrees of freedom of a base-pair modeled as a rigid body is characterized with 10 coordinates, 6 of which are defined relative to the previous base-pair in a dimer reference frame (inter base pair parameters):

- **Rise** is the relative displacement of one base pair to another in the direction of the helical axis (z-axis) while **slide** is the displacement of one base pair compared to another in the direction of the long axis (y-axis), measured between the midpoints of each C6-C8 vector. Similarly, **shift** describes the relative position of two neighboring base pairs along their short axis (x-axis).

- **Twist** is the angle between successive base pairs about the helical axis (z-axis). More practically, it is measured as the change in orientation of the C1'-C1' vectors going from one base pair to the next. Corresponding to slide in the translational parameters, **roll** is the dihedral angle for torsion of one base pair with its neighbor about the y-axis. A positive roll value opens the base pair towards the minor groove while negative roll indicates opening towards the major groove (groove definitions see section 1.2.3). **Tilt** (as shift for the translational inter base pair parameters) is the corresponding dihedral angle for rotation of one base pair with respect to its neighbor about the short axis (x-axis).

The remaining 4 parameters describe the geometry of a rigid base pair with respect to the helical axis:

- **X-displacement** and **Y-displacement** define the distance, along the x- or y-axis respectively, of the midpoint of the base pair mean plane with the helical axis. For example, a base pair with positive X displacement is translated towards the major groove.

- **Inclination** is the angle between the long axis (y-axis) of the rigid base pair and a plane perpendicular to the helical axis. **Tip** is the angle between the short axis (x-axis) of the base pair and a plane perpendicular to the helical axis.

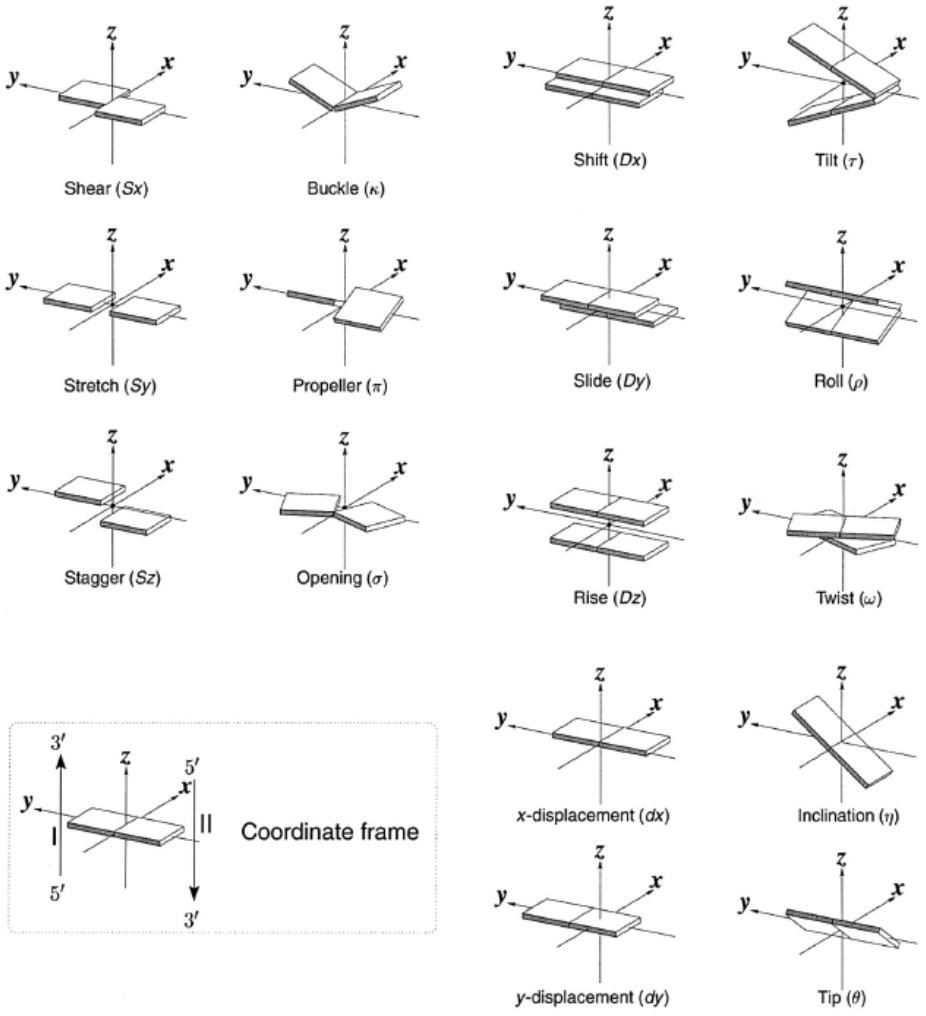


Figure 3. Base-pair geometry. Intra base pair parameters (top left), reference coordinate frame (bottom left), inter base pair parameters (right top) and global base pair parameters (right bottom).

In summary, the helical parameters are a complete and intuitive toolset for the description of the helix at base resolution level. Assuming rigid planar base pairs, the six inter base pair parameters fully characterize the structure of DNA. The assumption of rigid planar bases is a good approximation when global DNA dynamics at base pair level are studied since the displacement of individual bases is usually small compared to the rigid base-pair movement. However, if the relative orientation of bases within a base pair is wished to be considered, the six intra base pair degrees of freedom must be additionally introduced. Even though the helical parameters are mathematically independent, some of the parameters exhibit coupled behavior, for example slide, roll and twist change simultaneously with overall bending. Similarly, inter base pair parameters such as shift and twist are tightly connected to some of the dihedral angles (ε and ζ) defining backbone geometry.

1.2.2 Backbone geometry

The backbone configuration of DNA is best described by its torsional degrees of freedom. The torsion angles of a nucleotide consist of 6 main chain, 5 sugar and one glycosidic torsion angle.

The rotation of the base relative to the sugar is described by the glycosidic torsional angle χ (O4'-C1'-N9-C4 in purines and O4'-C1'-N1-C2 in pyrimidines). The base can adopt two major orientations about the C1'-N9 bond: syn and anti, and a minor one: high-anti. The angle ranges for the three conformations are 30°-90° for syn, 180°-300° for anti and around 270° for high-anti. In syn conformation the Watson-Crick hydrogen bonding groups are oriented towards the sugar while in the anti conformation these groups are directed away from the sugar ring. Purine bases can both be oriented in syn and anti with a slight preference for the anti configuration while pyrimidine nucleotides are found mostly in anti due to unfavorable electrostatic contacts in the syn configuration (O2 and phosphate group along the 5' direction). In canonical double-helical DNA, syn orientation of the nucleotides is almost never observed since Watson-Crick base pairing requires, in general, nucleotides to adopt anti conformation; only in some exotic DNA conformations such as Z-DNA, quadruplexes or triplex DNAs the syn orientation can play a significant role.

The sugar ring is the flexible link between the nucleobase and phosphate backbone, with different puckering modes influencing their relative orientation. The conformation of the five-membered

furanose ring of the backbone can be described by 5 endocyclic torsional angles ($\nu_0, \nu_1, \nu_2, \nu_3$ and ν_4, for definitions see Figure 4). The ring is usually not planar and atoms deviating from the ring coplanarity lead to pucker conformations. Sugar pucker states are called after cardinal directions, where C3'-endo-C2'-exo is called North, O4'-endo is called East, C3'-exo-C2'-endo is called South and O4'-exo is called West. These states can be summarized into a more elegant representation of the degree of the pucker by using a pseudorotation concept where $\tau_{\max}$ is the degree of the pucker and P the pseudorotation phase angle (7, 8). The parameters are calculated as:

$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2 \cdot \nu_2 \cdot (\sin 36° + \sin 72°)} \quad (1)$$

where P=0 corresponds to a maximally positive ν_2 torsional angle which is the standard conformation for nucleic acids. The puckering amplitude τ_m describes the maximum out-of-plane pucker and is given by:

$$\tau_m = \frac{\nu_2}{\cos P} \quad (2)$$

In practice, while the North conformation is predominant in RNA, DNA favors the South conformation with a P angle of 140° to 185° (9).

The *main chain torsions* of a nucleotide are comprised of six dihedral angles α, β, γ, δ, ε, ζ and are described in Figure 4. The torsion is defined by four consecutive atoms about the bond between the two central atoms (for example α is the rotation about the P-O5' bond). A common convention for describing these backbone angles is to define three major ranges as gauche+ (g+) around 60°, gauche- (g-) around 300° and trans (t) around 180°. Even though the six torsional angles represent six degrees of freedom, some main chain motions are correlated, forming torsional couples. Coupling of α and γ angle is responsible for the orientation of the phosphate group to the furanose since the β torsion adopts values mostly in the trans region. In the canonical form of B-DNA the α/γ couple adopts g-/g+. Another torsional couple is ε/ζ around the O3'. Their concerted rotations influence the motion of the O3' atoms and the phosphate group of the following nucleotide (on the 3' side). Two major regions in the ε/ζ conformational landscape, namely the BI and BII

backbone states, are characterized by the torsion difference (ε – ζ). The BI state is defined by ε and ζ adopting values of 120°-210° (t) and 235°-295° (g-) respectively.

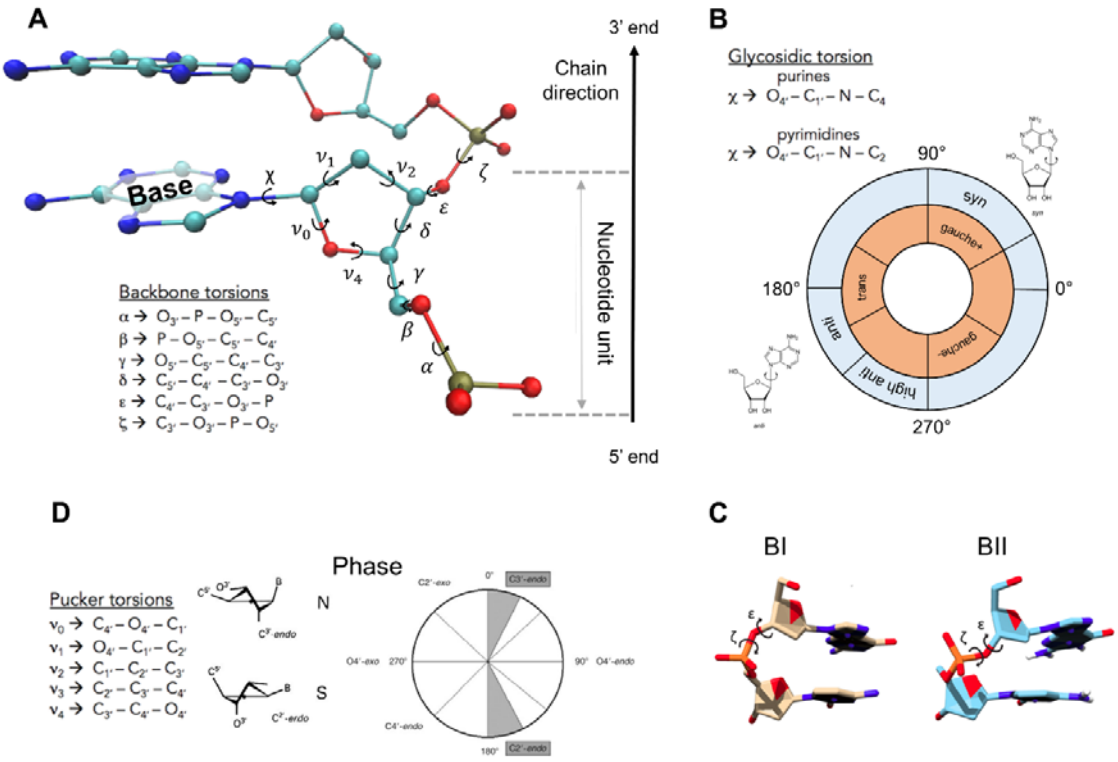


Figure 4. Definition of DNA backbone torsions. A: Main chain torsions. B: Glycosidic torsions. C: BI/BII transitions in the main chain. D: Puckering types.

In the transition from BI to BII state the phosphate of the following nucleotide is pushed towards the minor groove, narrowing it, and ε and ζ lie in the ranges of 210°-300° (g-) and 150°-210° (t) (10, 11). Backbone angles not only are coupled among each other, their motion correlates well with helical parameters. BI and BII states are associated with the inter base pair parameters, low twist/low shift corresponds to a BII conformation while high twist/high shift corresponds to the more common BI state. Moreover, as the helical parameters, the relative population of BI/BII states in DNA dynamics is sequence-dependent.

1.2.3 Helices

Several characteristics arise from the double helical structure of DNA. A dominant feature are the grooves which are spaces between the two strands. Due to the asymmetry in the base pairs two

parallel types of grooves exist: the *major groove* and the *minor groove* (Figure 5). Their dimensions are related to the distances and orientation of base pairs from and to the helical axis, respectively, and are characterized by two parameters. The groove width is defined as the perpendicular distance between phosphate groups on opposite strands with respect to the helical axis, the groove depth is calculated as the difference in polar radii between the phosphorous and N6 adenine or N2 guanine atoms, for major and minor groove respectively. The grooves can serve as binding pocket for different molecules, bigger molecules such as proteins preferably bind to the major groove while smaller ligands tend to bind to the minor groove.

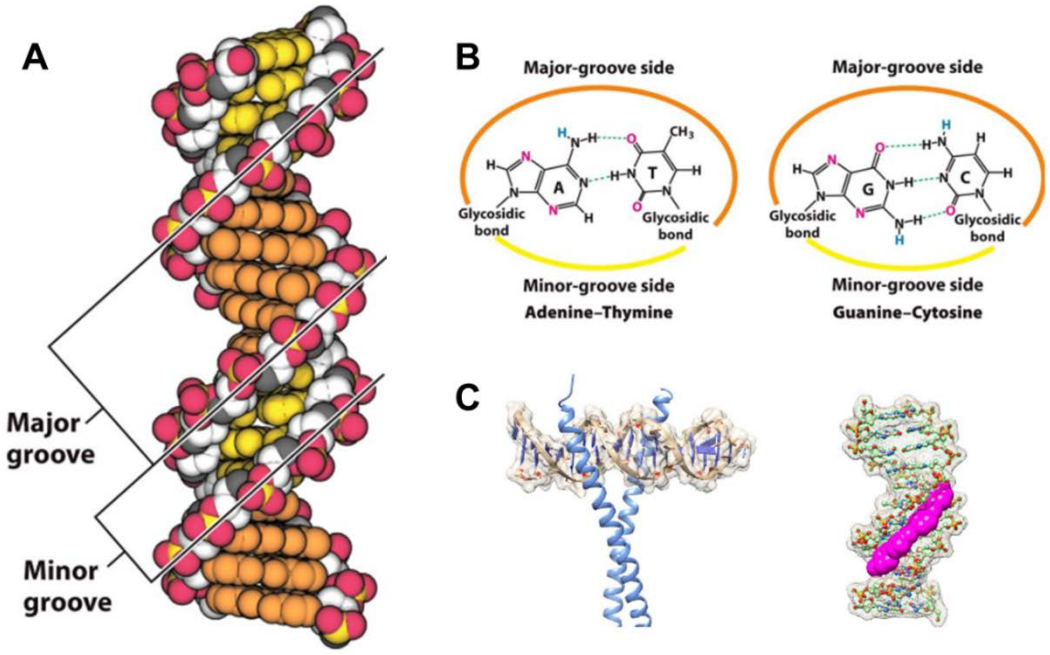


Figure 5. Groove geometry. A: Definition of major and minor groove. B: Orientations of base pairs towards the grooves (taken from Biochemistry: A Short Course (Second Edition), 2013). C: DNA protein complex with Leucine zipper bound major groove (left; PDB:1YSA) and DNA-ligand complex (ligand in magenta) bound to minor groove (right; PDB:264D).

1.2.4 Structural families

The most common structural type of DNA is its right-handed B-form (also called **B-DNA**). The B-DNA helix is characterized by a right-handed spiral formed by two anti-parallel polynucleotide chains with approximately 10 base pairs per complete helical turn, the sugar pucker in C2'-endo, an anti conformation of the glycosidic torsion and well defined major and minor grooves. The wide major groove is richer in H-bonding capabilities (O6, N6, N7 of purines and N4, O4 of

pyrimidines) than the minor groove (N3, N2 of purines and O2 of pyrimidines). For certain stretches of purines (e.g. GAGGGA) and under non-physiological conditions (low humidity) DNA can adopt the right-handed A-form. Compared to B-DNA, **A-DNA** has a wider spiral and a more compact form with over 11 base pairs per helical turn with smaller distance between them.

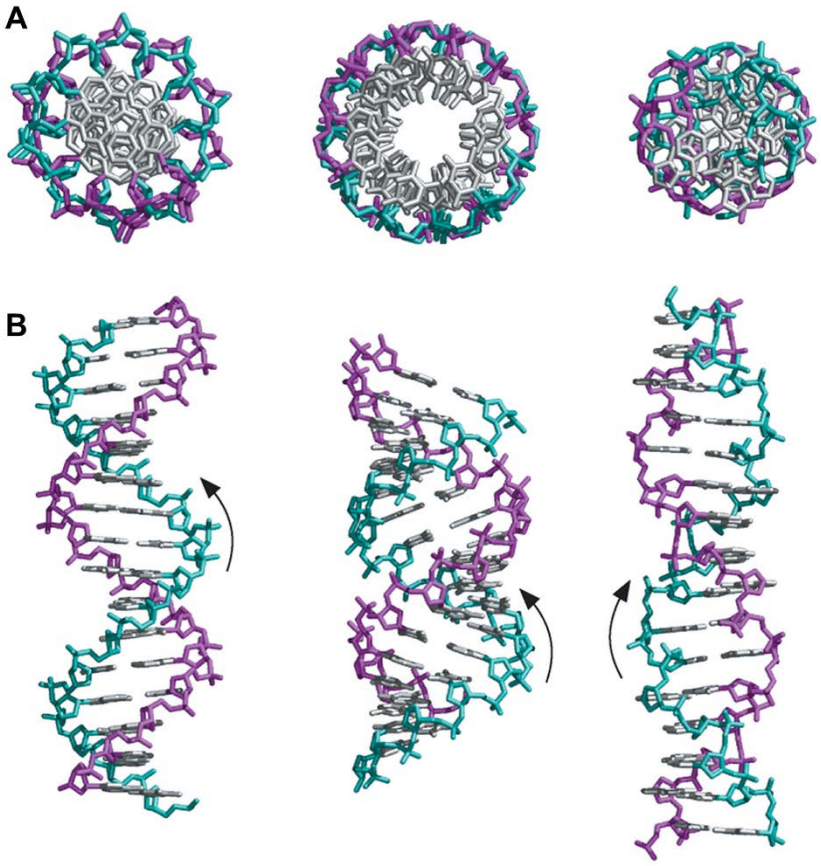


Figure 6. Three major forms of DNA double-helix. From left to right: B-DNA, A-DNA, Z-DNA. A: Top view. B: Side view.

The C3'-endo sugar pucker lowers the distance between consecutive phosphate groups which forces the displacement of the base pairs with respect to the helical axis by nearly 5Å. The last major family of DNA helices is Z-DNA. Z-DNA is a left-handed helix favored by alternating pyrimidine-purine steps (e.g. CGCGCG) at high ionic strength (above 4M NaCl). The purines of the left-handed double helix are in syn conformation which results in a "zig zag" arrangement of the phosphate groups. Even though the structure can exist only at high ionic strength it maybe is

involved in regulation of transcription (12). For completeness, the characteristics of the major helix forms of DNA are shown in Table 1.

Geometry attribute	A-DNA	B-DNA	Z-DNA
Helix sense	right-handed	right-handed	left-handed
Repeat unit	1 bp	1 bp	2 bp
Helical twist	32.7°	36.0°	C/G: -49.3°/-10.3°
Roll	0°	0°	C/G: 5.6°/-5.6°
bp/turn	11	10	6
Inclination	22.6°	2.8°	0.1°
Rise	2.54 Å	3.38 Å	7.25 Å
Pitch	28.2 Å	33.2 Å	45.6 Å
Propeller twist	-10.5°	-15.1°	8.3°
Glycosyl angle	anti	anti	C/G: anti/syn
Sugar pucker	C3'-endo	C2'-endo	C/G: C2'-endo/C2'-exo
Diameter	23 Å	20 Å	18 Å
Major groove			
Width	2.2 Å	11.6 Å	8.8 Å
Depth	13.0 Å	8.5 Å	3.7 Å
Minor groove			
Width	11.1 Å	6.0 Å	2.0 Å
Depth	2.6 Å	8.2 Å	13.8 Å

Table 1. Geometrical characteristics of the three major DNA double helices (data taken from (13)).

The studies presented in this work consider DNA in its B-form, other double helical conformers were only used for parametrizing and testing the newly developed parmbsc1 force field.

1.2.5 Constrained DNA

Even though the study of free DNA is essential to understand basic principles of DNA flexibility, DNA in nature is not always present in its naked B-form. In a cellular environment several factors can influence the structure of B-DNA, some of those reduce the DNA's freedom significantly. In this section I want to briefly discuss two types of B-DNA in a restrained environment, namely supercoiled DNA and protein-bound DNA (Figure 7). Simulations of those two types of constrained DNA are available via the MCDNA webserver I developed in my thesis to simulate DNA dynamics via a coarse grain model.

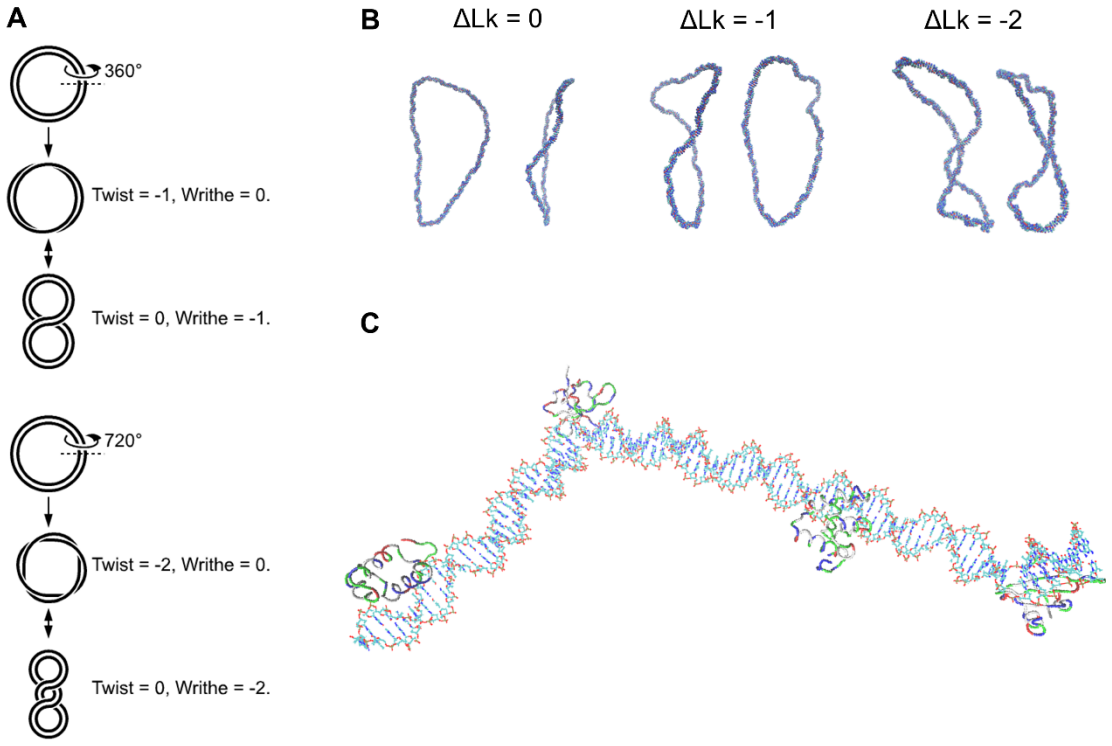


Figure 7. Constrained DNA. A: Schematic view of Twist and Writhe of a supercoiled circle. B: Representative structures of a DNA minicircle of 260bp in length with different changes in linking number (structure is shown from the front and rotated by 90°). C: Structure of protein-coated DNA.

DNA supercoiling

DNA supercoiling is a cellular strategy for packing the genetic material efficiently into a small nuclear space, but it is also implicated in genetic control. The over- or underwinding of DNA emerges from several cellular processes that induce torsional stress. In prokaryotes and eukaryotes the DNA is slightly negatively supercoiled (14). In such a constrained environment, for example, DNA supercoiling can enhance contact of DNA fragments which lie far apart in the linear genomic sequence. From a mathematical point of view, supercoiled DNA can be described as if the extremes of a DNA fragment were fused together and form a circle. An important variable for constrained circular structures is the linking number. The relaxed structure of an unconstrained DNA helix is characterized by the number of helical turns (the sum of the twist values of every base pair step divided by 360° (Tw)) and is called the default linking number Lk_0 of the relaxed circle (Lk_0=Tw). When DNA is over- or underwound and topologically constrained, the resultant torsional stress is relieved either by the introduction of writhe (Wr) which is the number of times the double helix crosses over on itself (supercoils) or by a change in the number of helical turns (Tw). The total linking number Lk then changes from its relaxed state Lk_0 ($\Delta\mathrm{Lk} = \mathrm{Lk} - \mathrm{Lk}_0$) and is distributed among Tw and Wr by satisfying the topological condition $\mathrm{Lk} = \mathrm{Tw} + \mathrm{Wr}$.

Protein-coated DNA

The interaction of regulatory proteins with DNA is crucial for several cellular processes ranging from gene expression regulation to DNA replication, repair and compaction.

Proteins can bind the DNA in two ways. In the non-specific binding the overall electrostatic attraction between protein and DNA and the overall DNA geometry are the main factors. In specific binding proteins recognize specific DNA sequences by either direct or indirect readout. In a direct readout the DNA sequence is read through specific contacts between amino acid side-chains and base functional groups exposed at the protein–DNA interface. In an indirect readout, proteins recognize DNA sequences through sequence-dependent variations in flexibility and structural parameters such as the groove width, the twist between base pairs, or the backbone conformation. In most cases both direct and indirect readouts work in a complementary way for

specific protein binding (15). Once the protein is bound to the DNA it alters the geometry of the nucleic acid at the binding region forcing it to deviate from its equilibrium conformation. However, this altered geometry is relatively stable (an example is the nucleosome (see next section)), as a result, DNA in protein-DNA complexes behaves like a rigid object compared to naked DNA. For example, for a DNA structure of 1000 base pairs in length without any proteins bound to it long-range contacts between distal DNA sites which can be essential for gene regulation are very rare. In contrast, when regulatory proteins bind to DNA, alter its path and constrain its flexibility more long-range contacts can emerge. This mechanism is the basis of gene regulation where DNA is compacted into chromatin inside the cell nucleus.

2. Chromatin structure – a multi-scale problem

In eukaryotes, the higher order structure of DNA inside the cell nucleus is called chromatin - the nucleoprotein complex that stores the genetic material. Chromatin is present in a highly compact form with its 3D arrangement resembling a tightly packed “ball of wool”. In humans, for example, DNA of 2m in length is compressed into a nucleus of around 6µm in diameter, which corresponds to a compression ratio of up to 10000. Due to such a high folding ratio, different levels of compaction have to be considered to thoroughly understand the multi-scale nature of the chromatin fiber. The first level of compaction is achieved by wrapping 147 base pairs (bp) of B-DNA ~1.7 times around an octamer of histone proteins forming the fundamental repeat unit of eukaryotic DNA: the nucleosome. Nucleosomes are connected via DNA linkers of 20-80 bp in length (depending on the organism) forming a “beads-on-a-string” form of chromatin. In the next level of compaction, the nucleosome string condenses into a polymer-like form. Early *in-vitro* (16, 17) and *in-silico* (18) experiments suggested a regular compaction into a fiber of 30nm in diameter, however *in-vivo* the situation is more complex due to many parameters such as DNA linker length, the linker histone concentration, the cellular ionic environment and the effect of chromatin remodelers (19, 20). In the last level of compaction, the dense chromatin chain is super-coiled forming chromosomes (see Figure 8). With the emergence of new experimental techniques in the last decades such as STORM (21), cryo-EM (22), FISH (23, 24) and 3C-based techniques (25, 26), the dynamic three-dimensional structure of chromatin among different resolution scales

from base pair to sub-chromosomal megabase level with its implications for gene regulation and diseases is now broadly studied (27–32).

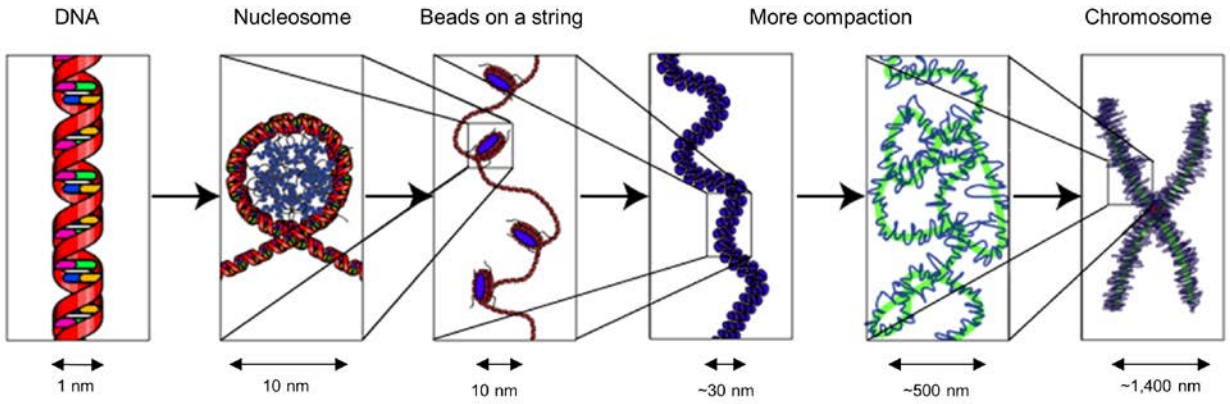


Figure 8. Multi-scale nature of chromatin structure.

2.1 Nucleosome

The nucleosome core particle consists of two copies of histones H2A, H2B, H3 and H4. In higher eukaryotes the accommodation of an additional linker histone is possible (H1 or H5 depending on the organism). It binds to the nucleosome core in the region close to the DNA edges of the DNA binding regions modulating the entry-exit angles of linker DNA which in turn influences the higher order structure of chromatin (33).

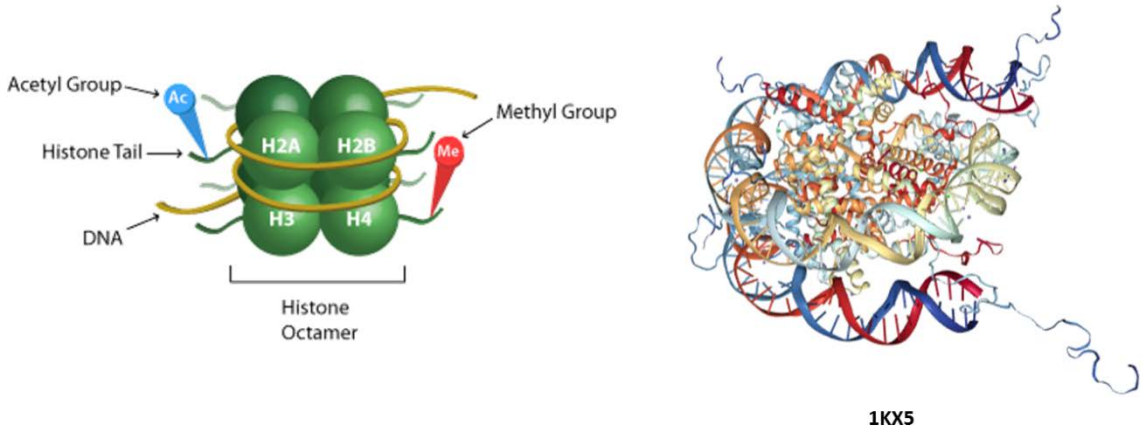


Figure 9. Nucleosome structure. A: Schematic view of the nucleosome structure with DNA, histone octamer, histone tails and a selection of possible epigenetic modifications. B: High-resolution X-ray structure (PDB:1KX5).

The sequence and molecular structure of histones is highly conserved among different species, with alpha helices allowing the polymerization and N-terminal tails unstructured and exposed to the environment serving as a target for many posttranslational covalent modification processes in the nucleus such as acetylation, methylation and phosphorylation. These modifications in turn are connected to the transcription state of the nearby genes (34) making them a strong indicator for the determination of active and inactive chromatin domains. The high-resolution X-ray crystal structure of the nucleosome ((35), PDB ID 1KX5, 1.9Å resolution, see Figure 9) reveals 147 base pairs of DNA wrapped 1.65 times around the cylindrical nucleosome core particle with the histone tails protruding out of the core. The shape of the nucleosomal DNA is far from equilibrium: the high DNA curvature is reflected by large absolute inter base pair step values of slide and roll and it kinks the minor groove in favor of the major groove. Due to this unusual shape the underlying sequence plays a significant role in wrapping around the histone complex, favoring or disfavoring nucleosome formation which has implications on nucleosome positioning along the genomic sequence (36, 37).

2.2 Chromatin secondary structure

In the next step, nucleosomes are connected via DNA linkers and chromatin secondary structure is then defined as the arrangement of the 'beads-on-a-string' fiber. Several regular topologies of chromatin secondary structure have been proposed to exist based on *in-vitro* data (38), the most popular among them are 'solenoid' and 'zigzag' arrangement of nucleosomes (Figure 10A). The one-start 'solenoid' model is an interdigitated one-start helix where consecutive nucleosomes interact with each other and follow a helical trajectory with bending of linker DNA. In the two-start 'zigzag' model straight linker DNA connects two opposing nucleosome cores which gives rise to a two-start helix. Due to advances in experiments and computational modeling in the last few years, it is now assumed that chromatin *in-vivo* adopts more dynamic and heterogeneous conformations (39) which depend on DNA linker length, linker histone concentration, epigenetic modifications and the effect of chromatin remodelers which altogether influence the local geometry and transcriptional state of chromatin. Recent experiments using super resolution STORM microcopy (21) (Figure 10B) suggest that chromatin in human and mouse is organized in nucleosomes assembled in heterogeneous groups of varying sizes called "clutches" interspersed

with nucleosome depleted regions where the clutch size and compaction correlates well with active and inactive chromatin.

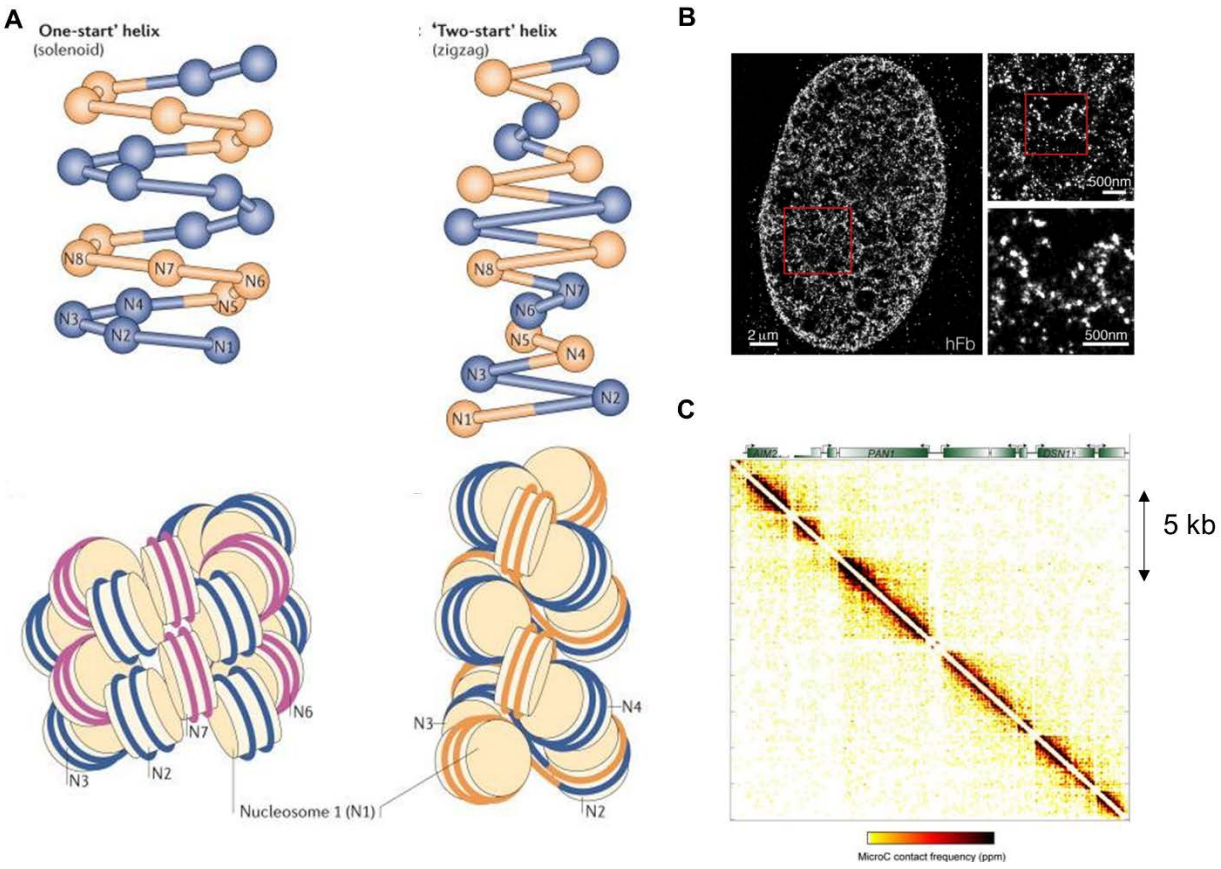


Figure 10. Secondary chromatin structure. A: Theoretical secondary structure motifs based on in-vitro experiments (taken from (32)). B: STORM microscopy of nucleosome occupancy in cells (taken from (21)). C: Micro-C contact matrix (taken from (40)).

In yeast, Micro-C (40) experiments revealed that nucleosomes form self-associating domains of 1-5 genes in size (ca. 2-10kb) where domain boundaries are enriched in nucleosome depleted regions (Figure 10C). The length of the DNA linker connecting two adjacent nucleosomes might play a decisive role in chromatin compaction (41), and the distribution of linker sizes can vary between different cells, even when they are perfectly synchronized. Unfortunately, most of experimental data are determined by population-based approaches, where thousands to millions of cells are needed to determine the 1D nucleosome positioning along the genomic sequence ((42); now also single cell MNase-seq exists (43), but the noise/signal ratio in these experiments is still too large). The conversion of nucleosome positions of a population of cells into a possible

3D conformation of chromatin inside a single cell is a promising approach to be able to represent the population-based dynamics in nucleosome positioning by a set of 3D structures derived by computational models.

2.3 Chromatin tertiary structures

The tertiary structure of chromatin in the interphase nucleus needs to be explained over a wide range of length and resolution scales (44, 45). It can be best understood 'top-down' – beginning with its biggest unit, the chromosomes (in the Mb scale), and ending with specific geometrical arrangements with biological relevance of the chromatin fiber in the kb regime (Figure 11). In the cell nucleus, chromosomes (in the Mb scale) are isolated and occupy distinct territories where inter-chromosomal interactions are rare compared to intra-chromosomal contacts. Large, gene-poor chromosomes are commonly located on the periphery near the nuclear membrane while gene-rich chromosomes are generally found inside the nuclear core. In the multi-Mb scale (in humans), chromosomes are organized into two spatial compartments labeled A and B, with A having a more open structure and being expression-active while B is more closed and expression-inactive (26). Beyond compartmentalization, chromatin is found to form self-associating domains called TADs. These self-interacting regions can range in the low Mb scale in humans down to around 5kb in smaller organisms such as yeast (40). Proteins attached to the boundary of TADs such as CTCF or cohesin are key factors of the dynamic remodeling of chromatin such as chromatin looping (low to high kb scale depending on the organism) where DNA regions which are far apart in the linear genome are brought into close contact. Looping events can regulate gene expression by influencing physical enhancer-promoter contacts. Approximately 50% of human genes are believed to be involved in long range chromatin interactions through DNA looping (46).

Most of the findings of 3D genome organization were achieved by a new experimental method. Since the emergence of Chromosome Conformation Capture techniques (3C) in 2002 (25) the study of global genome structure inside the cell nucleus has flourished in the last decade. Several 3C-based techniques exist, the most popular amongst them being called Hi-C (26). By cross-linking genomic segments of few hundred to a few thousand base pairs in length, Hi-C experiments can estimate genome-wide the frequency of interaction between genome loci in the nucleus. The measured contacts are statistical averages over a population of cells. A variety of polymer models

of the tertiary structure of chromatin have been developed to accommodate the large set of restraints which arise due to the contacts into a physically realistic geometrical structure (see section 3.4).

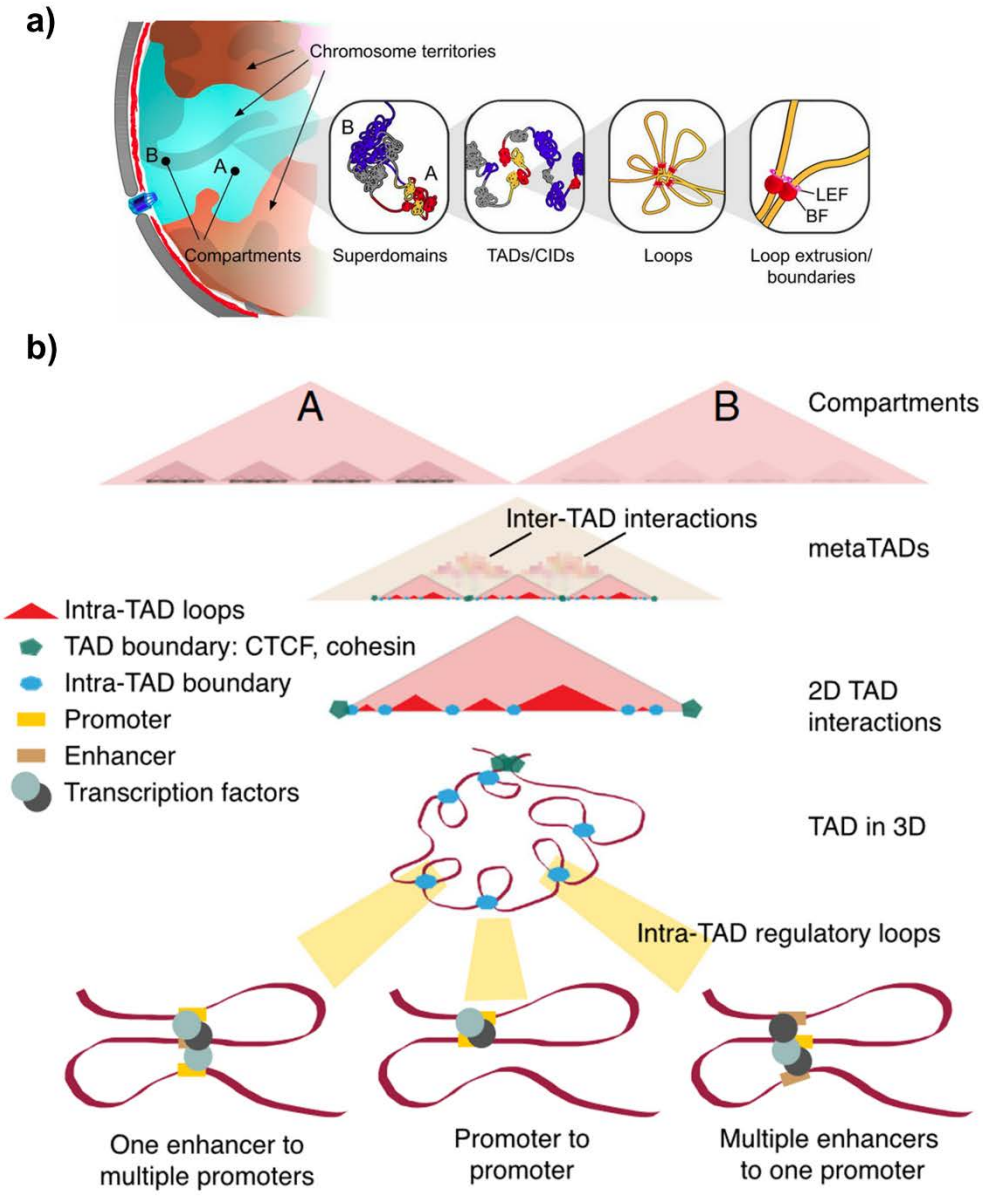


Figure 11. Chromatin tertiary structure. a) Schematic view of different levels of compaction (45). b) Single compaction states illustrated by means of contact matrices (44).

Using the Hi-C technique combined with FISH microscopy the structural genome variability and its biological implications are currently studied in different environments and conditions related

to cellular development and disease (for a more detailed description, see (29, 30)). Again, worries exist on the value of the 3D models obtained by imposing Hi-C restraints as they correspond to average contacts obtained in a pool of cells, and deconvolution of the cell-pool signal or alternatively single cell data obtained by ultra-resolution microscopy of high sensitivity or single cell Hi-C are required to provide a real representation of chromatin structure.

Until recently, the different levels of chromatin compaction were treated separately where in distinct resolution scales (chromatin secondary structure in the low kb scale and chromatin tertiary structure in the Mb scale) different tools for experimental and theoretical approaches are used. With the improvement in Hi-C resolution and cost-efficiency (1kb (47)) and the emergence of STORM microscopy, recent efforts are being made (48) where Hi-C and STORM microscopy attempt to connect both resolution levels by combining different experimental techniques and novel computational modeling which in my opinion will be the future of this field of research in the next decade.

3. Theoretical multi-scale modeling of DNA

In an ideal scenario, a single theoretical framework could describe the dynamic properties of DNA. However, the study of DNA covers a broad range of different scales. The nuclear DNA in a human cell measures more than 2m while the distance between two base pairs lies in the Å-scale. Some dynamic structural changes like chromatin reorganization along the cell cycle happen in the day time-scale while electronic rearrangements occur in the sub-femtosecond scale. It is then impossible for one single theoretical model to cover such a broad time and size range and therefore multiscale approaches relying on different levels of simplifications are necessary (Figure 12). The theoretical models applied to the study of DNA include (from small to large sizes or short to long time-scale) quantum mechanical (QM) ab initio approaches, classical atomistic molecular dynamics (MD), coarse grain (CG) and mesoscale modeling (49, 50).

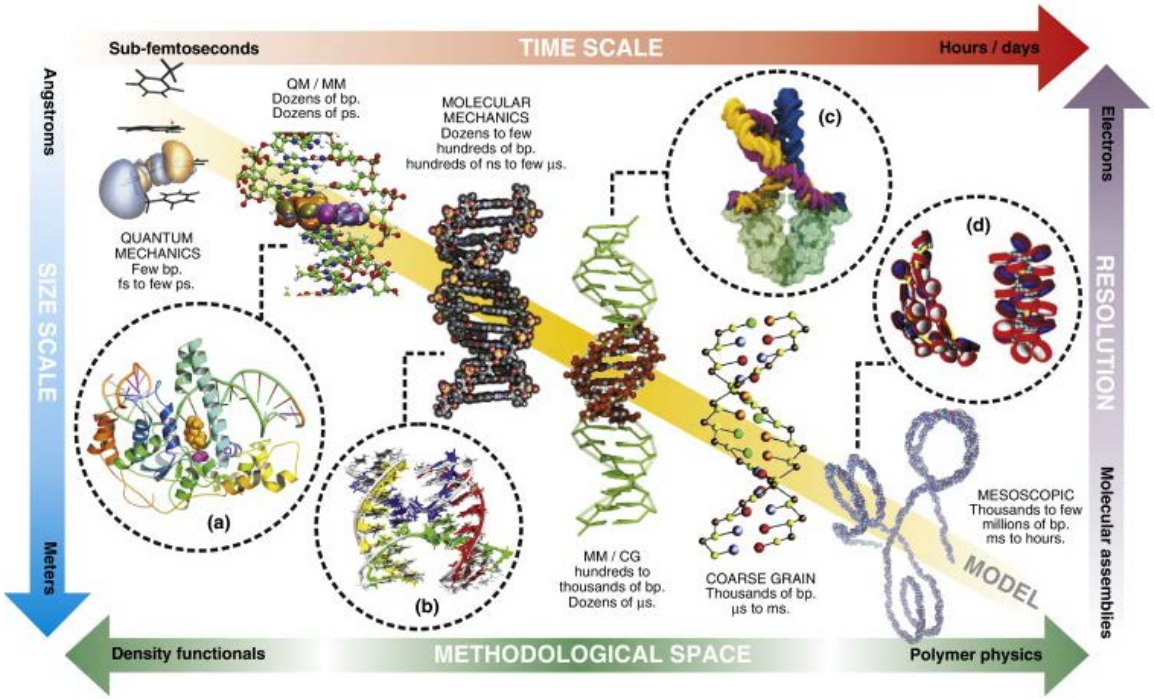


Figure 12. Multi-scale simulations of DNA.

3.1 Ab initio approaches

The highest level of detail of DNA simulations is achieved by QM approaches. These 'first principle' (ab initio) methods are used to study changes in electronic structure, including catalytic, photophysical or spectroscopic properties. QM models most commonly use the Born-Oppenheimer approximation where nuclei (treated as classical particles) and electron movements are disconnected. Average and more accurate representations are used to depict the correlation between electron densities. QM methods, even for the highest level of simplification, require an immense computational power which limits them for the study of small model systems (one or a few nucleotide units) at short time scales (sub femtosecond). Combining quantum mechanical and efficient molecular mechanical methods (QM/MM) makes it possible to study a larger system where only the region of interest is modeled in QM description while the surroundings are treated classically. QM/MM methods constitute a perfect theoretical framework for systems where the region requiring QM level can be precisely localized. For more information on QM and QM/MM methods see (51, 52).

3.2 Classical approaches

Instead of explicitly treating electronic densities as in the QM approaches, classical models represent atoms as deformable and charged balls of a given radius joined by springs. The energy terms modulating local (bonded) and remote (non-bonded) interactions are simplified as classical terms defining the force field, a classical Hamiltonian which can be used to derive forces which in turn is the fundament to derive trajectories by simple integration of Newton's equations (the molecular dynamics; MD approach). The force field is then the heart of Molecular Dynamics (MD) simulations and its parametrization is tightly connected to the accuracy of MD. Classical atomistic studies of DNA are usually done on duplexes of dozens (linear DNA) to a few hundreds of base pairs (circular DNA) and time scales of up to a few µs can be reached. A more detailed explanation of MD simulations is given in the following chapter (for a comprehensive review see (49)).

3.3 Coarse grain approaches

In coarse grain (CG) models the complexity of the system is reduced to achieve longer time and length scales (thousands of base pairs) than those accessible to MD simulations. In CG models, chemical groups or even entire residues are represented as single interacting centers, which decreases the number of pairwise interactions in the calculations of potential energies and forces (Figure 13). Two types of coarse graining exist to accurately describe the DNA dynamic properties (49, 50, 53, 54). Firstly, in Cartesian particle-based CG methods 3 to 8 beads represent one nucleotide and the beads are chosen to reproduce the connectivity between backbone, sugar puckering and base, as well as hydrogen bonds between bases (55–57). The energy functional of particle-based CG methods can be derived in a 'top-down' manner where the set of interactions is empirically parametrized by a trial-and-error procedure to fit experimentally determined thermodynamic properties or structural and dynamic features of double- and single-stranded DNA. In the 'bottom-up' approach, reference MD simulations are mapped into a CG system via the many-body potential of mean force (PMF). Most particle CG models use implicit Langevin dynamics with the solvent treated as continuum.

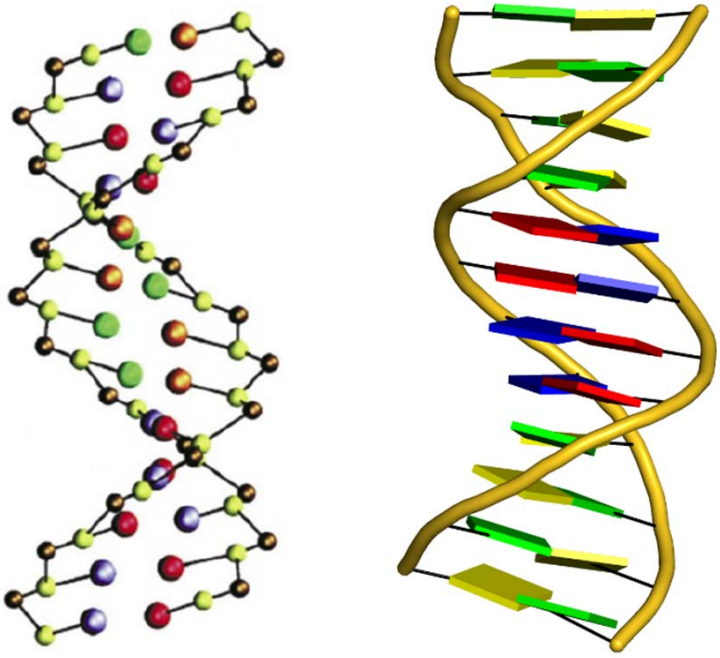


Figure 13. Coarse grain DNA models. Left: Cartesian particle-based coarse grain model by De Pablo group (Knotts et al. 2007). Right: Internal coarse grain model by Maddocks group (cgDNA Daiva et al. 2014 NAR).

The second coarse graining method is called 'internal CG model' and uses rigid bases or rigid base pairs, the relative movements between them being described by helical parameters. In the case of rigid bases an oligomer of n base pairs is represented by 12n-6 internal coordinates (58) while for rigid base pairs 6n-6 inter base pair parameters have to be considered (59, 60). The internal energy of the system is the sum of the local nearest-neighbor interactions and is typically represented in a quadratic form (harmonic approach) (49, 50, 53, 59–61). DNA conformations are sampled in the internal space usually via a Monte Carlo algorithm and are then subject to back mapping into Cartesian space. The internal CG 'force field' contains ground state and stiffness matrices that depend on the underlying DNA sequence and are parametrized from MD simulations. Most commonly, internal CG models use the rigid base pair approach with the base pair step (bps)-dependent parametrization. Nearest-neighbor representations of DNA have been traditionally used, which means that dynamics of DNA can be derived from the parameters of the 10 unique bps (59). However, recent studies (62, 63) revealed that the flanking base pairs of a bps influence the dynamics of the central bps so that this approach is not sufficient to describe DNA dynamics, and the tetranucleotide environment has to be considered (nearest neighbor). A

compendium of different particle-based and internal coordinate-based CG methods can be found in (49, 50, 64–68).

3.4 Mesoscopic approaches

For the description of secondary and tertiary chromatin structures (see description above) CG models become computationally too expensive forcing the use of even more simplified methods. Two types of models exist covering chromatin properties of either secondary or tertiary structure (Figure 14): (i) nucleosome fiber resolution working in the kb range and (ii) chromosome level resolution working in the Mb‑Gb range.

The 'bottom-up' approach of nucleosome fibers (27, 49, 50) takes the accurate atomistic description of the constituents of chromatin and transfers it into a coarse grain model. Known physical properties of the nucleosome core particle and linker DNA (both from experiment and simulations) are used to derive DNA flexibility and CG potentials. The representation of the nucleosome core is usually based on its experimentally determined X-ray structure while it is common to summarize several base pairs (6 in (69), 10 in (18)) into one bead and to use average bending and stretching properties derived from worm-like chain models for the linker DNA representation. Model parameters such as the ionic environment, the DNA linker length, the presence and absence of linker histones, or the existence of posttranslational histone modifications can be included to resemble the *in vivo* situation as close as possible (70, 71).

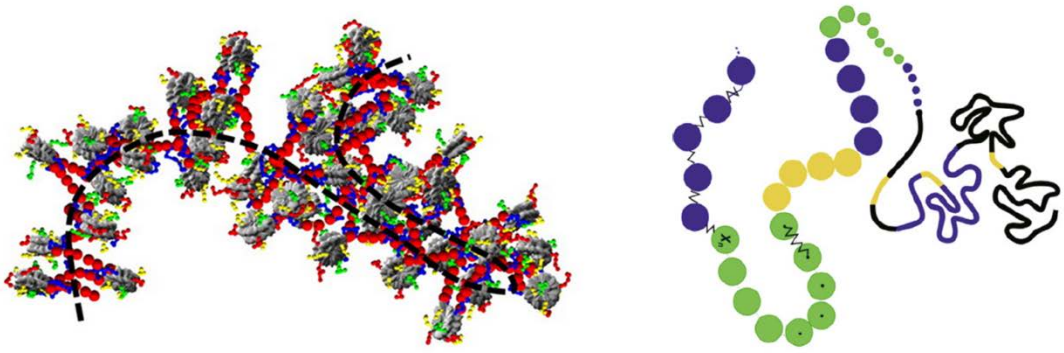


Figure 14. Mesoscopic models. Left: Nucleosome fiber model by Schlick group (39). Right: Block co-polymer model by Jost group (72).

Chromosome simulations usually make use of polymer models (one monomer can comprise from less than one kb up to several kb's (49)) and introduce additional interactions to fulfill experimental restraints such as the contact probabilities between genomic fragments derived by Hi-C or ultra-resolution microscopy experiments. By tuning specific inter-chain interactions these simple physical models can be very valuable to describe dynamic chromatin rearrangements in the Mb to Gb scale by simple mechanisms, for example specific attractive inter-chain interactions between monomers in certain regions with the same epigenetic mark could capture the nature of the epigenomic domains (72). A review of mesoscopic approaches of chromatin tertiary structure prediction can be found in (49, 50, 73, 74).

4. Programs for multiscale DNA modeling and analysis

Several programs were developed to simulate DNA from atomistic to chromosomal level, most of them are freely available for the user to download and compile to use the program on a local machine or cluster. The Amber suite of programs (75) is a prominent example of a complex toolkit to set up and perform MD simulations of biomolecules. Amber provides a set of in-house analysis programs via Ambertools and there also exist external programs such as Curves+ (76) or 3DNA (77) which provide information on the helical parameters, groove geometry and backbone conformation of the simulated DNA trajectory. DNA simulations via CG models can be performed for example with oxDNA (78) (Cartesian CG model) or cgDNAmc (79) (internal CG model). To predict the tertiary chromatin structure programs are provided which integrate experimental Hi-C data to convert into spatial restraints to build a polymer-like model of the three dimensional chromatin structure in an interactive way (80) (TADbit). Results of the 3D structure can be visualized alongside the experimental data via TADkit locally or in a web-based service (http://sgt.cnag.cat/3dg/tadkit/).

In recent years, the community for web-based services to facilitate simulation and analysis of nucleic acids has been growing steadily. Web services are used to make computational tools developed in-house freely available to experts as well as to non-experts without the sometimes laborious compilation of the source code on the local machine. Web services have the advantage that they can make use of a graphical interface for simple data input by the user and to show

directly the output of the program in a well understandable interactive way. They can comprise a single program or even a pipeline of several already existing tools so that the user can perform a multi-step process in a single interface. In the next section I will focus on web services for DNA structure generation and structure analysis.

4.1 Webservers for DNA structure generation and analysis

A plethora of web services exist to build, analyze and visualize DNA structures in atomistic or coarse grain representation. To build a DNA structure from scratch the user usually just has to provide a sequence as input and the webserver creates a DNA structure in a geometrical configuration specified by the user. There is the possibility to create straight or bent DNA (81, 82) conformations based on user specified input of helical parameters (77, 82) or based on precalculated equilibrium helical parameter values (83) (cgdna web). Those web services offer in-house analysis in the internal helical space and other geometrical parameters of the generated structure (82, 83). The user can also upload his own DNA structure in PDB format (76, 77, 81, 82) to use the analysis tools of the web services which usually comprise Cartesian and internal descriptors of the DNA geometry. Some web services not only allow structure generation and analysis, they can also setup and analyze atomistic MD simulations. NAFlex (81) (Figure 15) for example offers a variety of methods to explore nucleic acids flexibility, from base pair resolution elastic model of flexibility to MD simulations. Within the MD-framework NAFlex uses the MDWeb platform (84) to set-up the simulation by a multi-step procedure. Following preparation simulations can be launched using common molecular modeling programs such as Amber (75). Trajectories, obtained either in situ or provided by the user, can be visualized and analyzed by a variety of tools in helical and Cartesian space which makes NAFlex a tool for DNA simulation and analysis across multiple length scales. The trend nowadays is to follow the example of NAFlex, but at a much bigger scale offering the user more possibilities for biomolecular simulations and to analyze experimental data in the same online environment.

4.2 Online research environments

Another type of web environment emerged, the Virtual Research Environment (VRE; http://vre.multiscalegenomics.eu/home/; see Figure 15). It was developed by the Multiscale

Complex Genomics (MuG) consortium with participation of several known research groups in Europe.

Figure 15. Webservers and online research environments. Left: NAFlex webserver (*https://mmb.irbbarcelona.org/NAFlex*). Right: Virtual Research environment of the MuG consortium (*https://vre.multiscalegenomics.eu*) with a selection of integrated tools (top) and a snapshot of the virtual workspace (bottom).

In the VRE many individual programs to simulate DNA at different resolution and length scales and tools to analyze experimental data from 1D to 3D chromatin organization are integrated into a single online research environment. At the current state (as of February 2019) it comprises a set of tools for studying DNA and chromatin flexibility, determining *in-vivo* nucleosome positions along the genomic sequence (Mnase-seq) and analysis of Hi-C, ChIP- and DNase-seq data. The VRE is open for every tool developer to integrate their tools in a straight forward manner. Input files can be created directly in the VRE or uploaded by the user. All input and generated output of the integrated tools are visible in a single user workspace which makes it feasible for the user to execute different tools in the same environment. The existence of a single workspace per user allows interconnectivity between tools, which means the output of one tool can serve as input of another tool without any problems creating a pipeline of tool executions. A possible way to interconnect tools in the VRE is to first determine *in-vivo* nucleosome positions from MNase-seq data using nucleR (85) and use the predicted nucleosome positions of a genomic segment to construct a three-dimensional structure with the integrated nucleosome fiber model.

5. Parmbsc1

The balancing between the strong electrostatic repulsion between the phosphates and attractive forces such as stacking and hydrogen bonding between the nucleobases dictate DNA's overall stability. On top of that, the solvent environment influences this balance by screening phosphate repulsion and indirectly affects the fine shape of the DNA double helix. Accounting for a correct balance between strong and opposite interactions is a great challenge for simple molecular mechanics force fields.

The MMB group developed along many years strong knowledge in examining physical properties of nucleic acids by molecular dynamics simulations, part of the efforts can be contemplated in the parmbsc0 force field that, after its publication in 2007, has been the gold standard for force-fields until recently. However, as simulation time extended, several shortcomings of parmbsc0 simulations arose (86–92), among them excessive terminal base fraying (87), a too stiff χ torsion which led to difficulties in representing exotic DNA structures and significant deviations of helical parameter averages (twist and roll) related to the underestimation of the BI/BII equilibrium coming from wrong sampling of ε/ζ coupled distributions (87, 91).

Once identified those problems efforts in the group were undertaken to reparametrize the parmbsc0 force field with regard to the torsional backbone angles, most notably sugar puckering, ε/ζ and χ torsions using high-level QM calculations. In the meantime, several efforts by other groups were undertaken in parallel to improve known inconsistencies of parmbsc0. Specific corrections to parmbsc0 involved modifying the χ distribution (χOL4) for the simulation of DNA quadruplexes and the ε/ζ distribution (ε/ζOL1) (90, 93). Recently, the Czech group published the latest force field (OL15) which incorporated all previous OL corrections for DNA and included additional improvement on the β torsion angle estimation (94).

In our case, after more than four years of extensive testing, the new parameter set named parmbsc1 was released in 2014 and it was shown to outperform previous molecular mechanics force fields at that time (see Section 1.3 in Chapter II for an overview of the force field development).

By correcting the coupled ε/ζ profile, parmbsc1 (95) improved backbone sub-state populations BI/BII as ε and ζ in trans/gauche- (gauche-/trans) represent the canonical BI (BII) state (95). Additionally, by correcting the backbone dihedrals, known deviations from normality of helical parameters (twist and roll) of certain base pair steps which are tightly correlated to the backbone state and experimental values of helical parameters were well reproduced. The correction of the glycosidic angle accounted for the syn/anti equilibrium of the base orientation which reduced terminal base fraying and allowed for accurate simulation of non-canonical DNA structures. The resulting trajectories showed better conserved terminal hydrogen bonding and low RMSd of terminal bases compared to experimental data. Small imperfections with the puckering profiles appeared in a pilot simulation of the small duplex d(CpGpApTpCpG). The puckering profile was corrected by reparametrization of puckering torsions which correct an excessive bias of parmbsc0 towards East conformations.

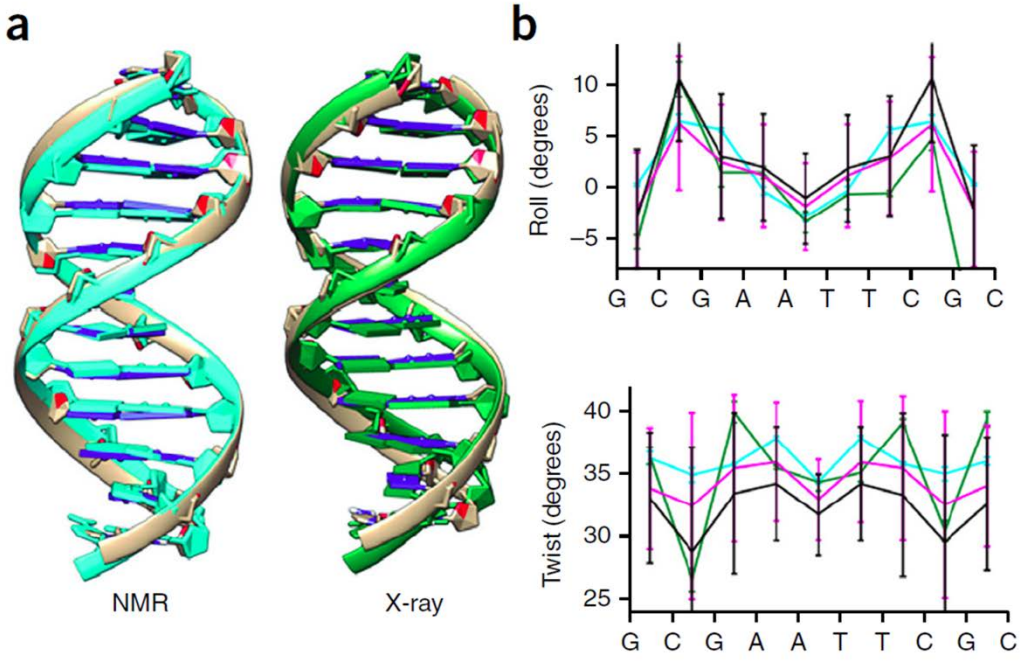


Figure 16. Analysis of DDD. (a) Comparison of the MD average structure (light brown) with the NMR structure (light blue) (PDB ID 1NAJ) and the X-ray structure (green) (PDB ID 1BNA). (b) Comparison of average values of helical rotational parameters (twist, roll and shift) per base-pair step coming from NMR (cyan), X-ray (green), 1-µs parmbsc0 trajectory (black) and 1.2-µs parmbsc1 trajectory (magenta) data. Error bars denote ±s.d. (adapted from *(95)*)

Incorporating these three corrections into one force field (95), we proceeded to assess its performance by validating it on more than a hundred DNA structures with a large variety of DNA motifs. The simulations with an overall accumulated time of ~140 μs all showed good agreement with known structural properties both from experiment and theory, the motifs ranging from unrestrained canonical B-DNA, canonical B-DNA restrained to a circle, various non-canonical forms, other unusual DNA conformations such as triplexes and quadruplexes, to complexes where DNA is bound to a protein or a ligand.

One of the validation studies was a simulation of the most known B-DNA duplex, the Drew-Dickerson dodecamer (DDD). Parmbsc1 simulations lead to significant improvements compared to parmbsc0 simulations preserving hydrogen bonds and helical parameters at the terminal residues (see Figure 16) sampling now correctly twist and roll profiles with the average much closer to experimental values as well as an increased BII population (by 7%).

Other tests involved simulations of DNA minicircles of 106 bp in length at different superhelical stress. While without supercoiling no denatured regions were observed (only one kink out of all replicas) negatively supercoiled minicircles formed distortions as a result of superhelical stress, a phenomenon known experimentally (96, 97). Looking at the DNA flexibility parmbsc1 predicted persistence lengths obtained from simulations of long (up to 56 bp) canonical B-DNA duplexes in the range of 40–57 nm, close to the generally accepted value of 50 nm. Besides the universal experimental validation by structural comparison also direct experimental observables could be computed for the structures where data was available. In the case of DDD, RDCs and NOEs are similar to those obtained in the NMR-refined structures. Additionally, simulations using parmbsc1 provide violations statistics equivalent to those determined from “de novo” NMR-derived ensembles.

In summary, parmbsc1 undoubtedly constitutes a major improvement to previous force fields in the study of dynamic properties of DNA. Two hundred citations at present time (more than 100 predicted in 2019) indicates the impact that this force-field is having in the field.

Bibliography for Chapter I

1. WATSON,J.D. and CRICK,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–8.
http://www.ncbi.nlm.nih.gov/pubmed/13054692

2. CRICK,F. (1970) Central Dogma of Molecular Biology. *Nature*, **227**, 561–563.
https://doi.org/10.1038/227561a0

3. Watson,J.D. (2008) Molecular biology of the gene Pearson/Benjamin Cummings.

4. Boyle,J. (2008) Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Biochem. Mol. Biol. Educ.*, **36**, 317–318.
https://doi.org/10.1002/bmb.20192

5. E. Stofer, C. Chipot,† and and Lavery*,R. (1999) Free Energy Calculations of Watson−Crick Base Pairing in Aqueous Solution. 10.1021/JA991092Z.
https://doi.org/10.1021/JA991092Z

6. Olson,W.K., Bansal,M., Burley,S.K., Dickerson,R.E., Gerstein,M., Harvey,S.C., Heinemann,U., Lu,X.-J., Neidle,S., Shakked,Z., *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry 1 1Edited by P. E. Wright 2 2This is a document of the Nomenclature Committee of IUBMB (NC-IUBMB)/IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN), whose members are R. Cammack (chairman), A. Bairoch, H.M. Berman, S. Boyce, C.R. Cantor, K. Elliott, D. Horton, M. Kanehisa, A. Kotyk, G.P. Moss, N. Sharon and K.F. Tipton. *J. Mol. Biol.*, **313**, 229–237.
https://doi.org/10.1006/jmbi.2001.4987
http://www.ncbi.nlm.nih.gov/pubmed/11601858

7. Westhof,E. and Sundaralingam,M. (1983) A method for the analysis of puckering disorder in five-membered rings: the relative mobilities of furanose and proline rings and their effects on polynucleotide and polypeptide backbone flexibility. *J. Am. Chem. Soc.*, **105**, 970–976.
https://doi.org/10.1021/ja00342a054

8. Altona,C. and Sundaralingam,M. (1972) Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *J. Am. Chem. Soc.*, **94**, 8205–8212.
https://doi.org/10.1021/ja00778a043

9. Levitt,M. and Warshel,A. (1978) Extreme conformational flexibility of the furanose ring in DNA and RNA. *J. Am. Chem. Soc.*, **100**, 2607–2613.
https://doi.org/10.1021/ja00477a004

10. Trieb,M., Rauch,C., Wellenzohn,B., Wibowo,F., Loerting,T. and Liedl,K.R. (2004) Dynamics of DNA: B_I and B_{II} Phosphate Backbone Transitions. *J. Phys. Chem. B*, **108**, 2470–2476.
https://doi.org/10.1021/jp037079p

11. Hartmann,B., Piazzola,D. and Lavery,R. (1993) BI-BII transitions in B-DNA. *Nucleic Acids Res.*, **21**, 561–8.
https://doi.org/10.1093/nar/21.3.561
http://www.ncbi.nlm.nih.gov/pubmed/8441668

12. Oh,D.-B., Kim,Y.-G. and Rich,A. (2002) Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proc. Natl. Acad. Sci.*, **99**, 16666–16671.
https://doi.org/10.1073/pnas.262672699
http://www.ncbi.nlm.nih.gov/pubmed/12486233

13. Neidle,S. (2008) Principles of nucleic acid structure Elsevier.

14. Noy,A., Maxwell,A. and Harris,S.A. (2017) Interference between Triplex and Protein Binding to Distal Sites on Supercoiled DNA. *Biophys. J.*, **112**, 523–531.
https://doi.org/10.1016/j.bpj.2016.12.034
http://www.ncbi.nlm.nih.gov/pubmed/28108011

15. Siggers,T. and Gordan,R. (2014) Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.*, **42**, 2099–2111.
https://doi.org/10.1093/nar/gkt1112
http://www.ncbi.nlm.nih.gov/pubmed/24243859

16. Grigoryev,S.A. (2018) Chromatin Higher-Order Folding: A Perspective with Linker DNA Angles. *Biophys. J.*, **114**, 2290–2297.
https://doi.org/10.1016/j.bpj.2018.03.009
http://www.ncbi.nlm.nih.gov/pubmed/29628212

17. Horowitz-Scherer,R.A. and Woodcock,C.L. (2006) Organization of interphase chromatin. *Chromosoma*, **115**, 1–14.
https://doi.org/10.1007/s00412-005-0035-3

18. Schlick,T. and Perisić,O. (2009) Mesoscale simulations of two nucleosome-repeat length oligonucleosomes. *Phys. Chem. Chem. Phys.*, **11**, 10729–37.
https://doi.org/10.1039/b918629h
http://www.ncbi.nlm.nih.gov/pubmed/20145817

19. Collepardo-Guevara,R. and Schlick,T. (2014) Chromatin fiber polymorphism triggered by variations of DNA linker lengths. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 8061–6.
https://doi.org/10.1073/pnas.1315872111
http://www.ncbi.nlm.nih.gov/pubmed/24847063

20. Fan,Y., Korolev,N., Lyubartsev,A.P. and Nordenskiöld,L. (2013) An Advanced Coarse-Grained Nucleosome Core Particle Model for Computer Simulations of Nucleosome-Nucleosome Interactions under Varying Ionic Conditions. *PLoS One*, **8**, e54228.
https://doi.org/10.1371/journal.pone.0054228

21. Ricci,M.A., Manzo,C., García-Parajo,M.F., Lakadamyali,M. and Cosma,M.P. (2015) Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*, **160**, 1145–58.
https://doi.org/10.1016/j.cell.2015.01.054
http://www.ncbi.nlm.nih.gov/pubmed/25768910

22. Song,F., Chen,P., Sun,D., Wang,M., Dong,L., Liang,D., Xu,R.-M., Zhu,P. and Li,G. (2014) Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units. *Science (80-.).*, **344**, 376–380.
https://doi.org/10.1126/science.1251413
http://www.ncbi.nlm.nih.gov/pubmed/24763583

23. Rouquette,J., Cremer,C., Cremer,T. and Fakan,S. (2010) Functional nuclear architecture studied by microscopy: present and future. *Int. Rev. Cell Mol. Biol.*, **282**, 1–90.
https://doi.org/10.1016/S1937-6448(10)82001-5
http://www.ncbi.nlm.nih.gov/pubmed/20630466

24. Cremer,M., Grasser,F., Lanctôt,C., Müller,S., Neusser,M., Zinner,R., Solovei,I. and Cremer,T. (2012) Multicolor 3D Fluorescence In Situ Hybridization for Imaging Interphase Chromosomes. In *Methods in molecular biology (Clifton, N.J.)*.Vol. 463, pp. 205–239.
https://doi.org/10.1007/978-1-59745-406-3_15
http://www.ncbi.nlm.nih.gov/pubmed/18951171

25. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing Chromosome Conformation. *Science (80-.).*, **295**, 1306–1311.
https://doi.org/10.1126/science.1067799
http://www.ncbi.nlm.nih.gov/pubmed/11847345

26. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93.
https://doi.org/10.1126/science.1181369
http://www.ncbi.nlm.nih.gov/pubmed/19815776

27. Ozer,G., Luque,A. and Schlick,T. (2015) The chromatin fiber: multiscale problems and approaches. *Curr. Opin. Struct. Biol.*, **31**, 124–139.
https://doi.org/10.1016/j.sbi.2015.04.002
http://www.ncbi.nlm.nih.gov/pubmed/26057099

28. Pombo,A. and Dillon,N. (2015) Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.*, **16**, 245–257.
https://doi.org/10.1038/nrm3965

29. Spielmann,M., Lupiáñez,D.G. and Mundlos,S. (2018) Structural variation in the 3D genome. *Nat. Rev. Genet.*, **19**, 453–467.
https://doi.org/10.1038/s41576-018-0007-0
http://www.ncbi.nlm.nih.gov/pubmed/29692413

30. Bonev,B. and Cavalli,G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661–678.
https://doi.org/10.1038/nrg.2016.112

31. Marti-Renom,M.A., Almouzni,G., Bickmore,W.A., Bystricky,K., Cavalli,G., Fraser,P., Gasser,S.M., Giorgetti,L., Heard,E., Nicodemi,M., *et al.* (2018) Challenges and guidelines toward 4D nucleome data and model standards. *Nat. Genet.*, **50**, 1352–1358.
https://doi.org/10.1038/s41588-018-0236-3

32. Luger,K., Dechassa,M.L. and Tremethick,D.J. (2012) New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.*, **13**, 436–47.
https://doi.org/10.1038/nrm3382
http://www.ncbi.nlm.nih.gov/pubmed/22722606

33. Zhou,B.-R., Feng,H., Kato,H., Dai,L., Yang,Y., Zhou,Y. and Bai,Y. (2013) Structural insights into the histone H1-nucleosome complex. *Proc. Natl. Acad. Sci.*, **110**, 19390–19395.
https://doi.org/10.1073/pnas.1314905110
http://www.ncbi.nlm.nih.gov/pubmed/24218562

34. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
https://doi.org/10.1038/nbt.1662

35. Davey,C.A., Sargent,D.F., Luger,K., Maeder,A.W. and Richmond,T.J. (2002) Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution. *J. Mol. Biol.*, **319**, 1097–1113.
https://doi.org/10.1016/S0022-2836(02)00386-8
http://www.ncbi.nlm.nih.gov/pubmed/12079350

36. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thåström,A., Field,Y., Moore,I.K., Wang,J.-P.Z. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
https://doi.org/10.1038/nature04979

37. Olson,W.K. and Zhurkin,V.B. (2011) Working the kinks out of nucleosomal DNA. *Curr. Opin. Struct. Biol.*, **21**, 348–357.
https://doi.org/10.1016/j.sbi.2011.03.006

38. Tremethick,D.J. (2007) Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. *Cell*, **128**, 651–654.
https://doi.org/10.1016/j.cell.2007.02.008
http://www.ncbi.nlm.nih.gov/pubmed/17320503

39. Grigoryev,S.A., Arya,G., Correll,S., Woodcock,C.L. and Schlick,T. (2009) Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci.*, **106**, 13317–13322.
https://doi.org/10.1073/pnas.0903280106

40. Hsieh,T.-H.S., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, **162**, 108–119.
https://doi.org/10.1016/j.cell.2015.05.048
http://www.ncbi.nlm.nih.gov/pubmed/26119342

41. Wiese,O., Marenduzzo,D. and Brackley,C.A. (2018) Nucleosome positions alone determine micro-domains in yeast chromosomes. *bioRxiv*, 10.1101/456202.
https://doi.org/10.1101/456202

42. Teng,Y., Yu,S. and Waters,R. (2001) The mapping of nucleosomes and regulatory protein binding sites at the Saccharomyces cerevisiae MFA2 gene: a high resolution approach. *Nucleic Acids Res.*, **29**, 64e – 64.
https://doi.org/10.1093/nar/29.13.e64

43. Lai,B., Gao,W., Cui,K., Xie,W., Tang,Q., Jin,W., Hu,G., Ni,B. and Zhao,K. (2018) Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature*, **562**, 281–285.
https://doi.org/10.1038/s41586-018-0567-3
http://www.ncbi.nlm.nih.gov/pubmed/30258225

44. Mishra,A. and Hawkins,R.D. (2017) Three-dimensional genome architecture and emerging technologies: looping in disease. *Genome Med.*, **9**, 87.
https://doi.org/10.1186/s13073-017-0477-2
http://www.ncbi.nlm.nih.gov/pubmed/28964259

45. Pueschel,R., Coraggio,F. and Meister,P. (2016) From single genes to entire genomes: the search for a function of nuclear organization. *Development*, **143**, 910–923.
https://doi.org/10.1242/dev.129007
http://www.ncbi.nlm.nih.gov/pubmed/26980791

46. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.-A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–4.
https://doi.org/10.1038/nature12644
http://www.ncbi.nlm.nih.gov/pubmed/24141950

47. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–80.
https://doi.org/10.1016/j.cell.2014.11.021
http://www.ncbi.nlm.nih.gov/pubmed/25497547

48. Bintu,B., Mateo,L.J., Su,J.-H., Sinnott-Armstrong,N.A., Parker,M., Kinrot,S., Yamaya,K., Boettiger,A.N. and Zhuang,X. (2018) Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science (80-.).*, **362**, eaau1783.
https://doi.org/10.1126/science.aau1783
http://www.ncbi.nlm.nih.gov/pubmed/30361340

49. Dans,P.D., Walther,J. and Gómez,H. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45.
https://doi.org/10.1016/J.SBI.2015.11.011

50. Gómez,H., Walther,J., Darré,L., Ivani,I., Dans,P.D. and Orozco,M. (2017) Chapter 7. Molecular Modelling of Nucleic Acids. In.pp. 165–197.
https://doi.org/10.1039/9781788010139-00165

51. Šponer,J., Riley,K.E. and Hobza,P. (2008) Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys. Chem. Chem. Phys.*, **10**, 2595.
https://doi.org/10.1039/b719370j
http://www.ncbi.nlm.nih.gov/pubmed/18464974

52. Banáš,P., Jurečka,P., Walter,N.G., Šponer,J. and Otyepka,M. (2009) Theoretical studies of RNA catalysis: Hybrid QM/MM methods and their comparison with MD and QM. *Methods*, **49**, 202–216.
https://doi.org/10.1016/j.ymeth.2009.04.007
http://www.ncbi.nlm.nih.gov/pubmed/19398008

53. Orozco,M., Pérez,A., Noy,A. and Luque,F.J. (2003) Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.*, **32**, 350–364.
https://doi.org/10.1039/B207226M

54. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193.
https://doi.org/10.1016/j.sbi.2008.01.005
http://www.ncbi.nlm.nih.gov/pubmed/18304803

55. Knotts,T.A., Rathore,N., Schwartz,D.C. and de Pablo,J.J. (2007) A coarse grain model for DNA. *J. Chem. Phys.*, **126**, 084901.
https://doi.org/10.1063/1.2431804
http://www.ncbi.nlm.nih.gov/pubmed/17343470

56. Machado,M.R. and Pantano,S. (2015) Exploring LacI–DNA Dynamics by Multiscale Simulations Using the SIRAH Force Field. *J. Chem. Theory Comput.*, **11**, 5012–5023.
https://doi.org/10.1021/acs.jctc.5b00575

57. Uusitalo,J.J., Ingólfsson,H.I., Akhshi,P., Tieleman,D.P. and Marrink,S.J. (2015) Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.*, **11**, 3932–3945.
https://doi.org/10.1021/acs.jctc.5b00286

58. Petkevičiūtė,D., Pasi,M., Gonzalez,O. and Maddocks,J.H. (2014) cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.*, **42**, e153.
https://doi.org/10.1093/nar/gku825
http://www.ncbi.nlm.nih.gov/pubmed/25228467

59. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad.*

Sci. U. S. A., **95**, 11163–8.
https://doi.org/10.1073/pnas.95.19.11163
http://www.ncbi.nlm.nih.gov/pubmed/9736707

60. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.*, **85**, 2872–2883.
https://doi.org/10.1016/S0006-3495(03)74710-9
http://www.ncbi.nlm.nih.gov/pubmed/14581192

61. Lankaš,F., Gonzalez,O., Heffler,L.M., Stoll,G., Moakher,M. and Maddocks,J.H. (2009) On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, **11**, 10565.
https://doi.org/10.1039/b919565n
http://www.ncbi.nlm.nih.gov/pubmed/20145802

62. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
https://doi.org/10.1093/nar/gku855
http://www.ncbi.nlm.nih.gov/pubmed/25260586

63. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H., *et al.* (2005) Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps. *Biophys. J.*, **89**, 3721–3740.
https://doi.org/10.1529/biophysj.105.067397
http://www.ncbi.nlm.nih.gov/pubmed/16169978

64. Geršata,T. and Lankaš,F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102.
https://doi.org/10.1088/0953-8984/27/32/323102

65. Sim,A.Y., Minary,P. and Levitt,M. (2012) Modeling nucleic acids. *Curr. Opin. Struct. Biol.*, **22**, 273–278.
https://doi.org/10.1016/J.SBI.2012.03.012

66. Potoyan,D.A., Savelyev,A. and Papoian,G.A. (2013) Recent successes in coarse-grained modeling of DNA. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 69–83.
https://doi.org/10.1002/wcms.1114

67. Ingólfsson,H.I., Lopez,C.A., Uusitalo,J.J., de Jong,D.H., Gopal,S.M., Periole,X. and Marrink,S.J. (2014) The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **4**, 225–248.
https://doi.org/10.1002/wcms.1169

68. Noid,W.G. (2013) Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, **139**, 090901.
https://doi.org/10.1063/1.4818908

69. Kimura,H., Shimooka,Y., Nishikawa,J., Miura,O., Sugiyama,S., Yamada,S. and Ohyama,T. (2013) The genome folding mechanism in yeast. *J. Biochem.*, **154**, 137–147.
https://doi.org/10.1093/jb/mvt033

70. Luque,A., Ozer,G. and Schlick,T. (2016) Correlation among DNA Linker Length, Linker Histone Concentration, and Histone Tails in Chromatin. *Biophys. J.*, **110**, 2309–2319.
https://doi.org/10.1016/j.bpj.2016.04.024
http://www.ncbi.nlm.nih.gov/pubmed/27276249

71. Collepardo-Guevara,R., Portella,G., Vendruscolo,M., Frenkel,D., Schlick,T. and Orozco,M. (2015) Chromatin Unfolding by Epigenetic Modifications Explained by Dramatic Impairment of Internucleosome Interactions: A Multiscale Computational Study. *J. Am. Chem. Soc.*, **137**, 10205–10215.
https://doi.org/10.1021/jacs.5b04086

72. Jost,D., Carrivain,P., Cavalli,G. and Vaillant,C. (2014) Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.*, **42**, 9553–9561.
https://doi.org/10.1093/nar/gku698
http://www.ncbi.nlm.nih.gov/pubmed/25092923

73. .Tiana,G. and Giorgetti,L. (2018) Integrating experiment, theory and simulation to determine the structure and dynamics of mammalian chromosomes. *Curr. Opin. Struct. Biol.*, **49**, 11–17.
https://doi.org/10.1016/j.sbi.2017.10.016
http://www.ncbi.nlm.nih.gov/pubmed/29128709

74. Sugawara,T. and Kimura,A. (2017) Physical properties of the chromosomes and implications for development. *Dev. Growth Differ.*, **59**, 405–414.
https://doi.org/10.1111/dgd.12363

75. D.A.Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R

Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGra,D.M.Y. and P.A.K. (2018) AMBER 2018.

76. Blanchet,C., Pasi,M., Zakrzewska,K. and Lavery,R. (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.*, **39**, W68-73.
https://doi.org/10.1093/nar/gkr316
http://www.ncbi.nlm.nih.gov/pubmed/21558323

77. Lu,X.-J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
https://doi.org/10.1093/nar/gkg680
http://www.ncbi.nlm.nih.gov/pubmed/12930962

78. Ouldridge,T.E., Louis,A.A. and Doye,J.P.K. (2010) DNA Nanotweezers Studied with a Coarse-Grained Model of DNA. *Phys. Rev. Lett.*, **104**, 178101.
https://doi.org/10.1103/PhysRevLett.104.178101

79. Mitchell,J.S., Glowacki,J., Grandchamp,A.E., Manning,R.S. and Maddocks,J.H. (2017) Sequence-Dependent Persistence Lengths of DNA. *J. Chem. Theory Comput.*, **13**, 1539–1555.
https://doi.org/10.1021/acs.jctc.6b00904

80. Serra,F., Baù,D., Goodstadt,M., Castillo,D., Filion,G.J. and Marti-Renom,M.A. (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLOS Comput. Biol.*, **13**, e1005665.
https://doi.org/10.1371/journal.pcbi.1005665

81. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47-55.
https://doi.org/10.1093/nar/gkt378
http://www.ncbi.nlm.nih.gov/pubmed/23685436

82. van Dijk,M. and Bonvin,A.M.J.J. (2009) 3D-DART: a DNA structure modelling server. *Nucleic Acids Res.*, **37**, W235-9.
https://doi.org/10.1093/nar/gkp287
http://www.ncbi.nlm.nih.gov/pubmed/19417072

83. De Bruin,L. and Maddocks,J.H. (2018) cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.*, **46**,

W5–W10.
https://doi.org/10.1093/nar/gky351

84. Hospital,A., Andrio,P., Fenollosa,C., Cicin-Sain,D., Orozco,M. and Gelpí,J.L. (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics*, **28**, 1278–1279.
https://doi.org/10.1093/bioinformatics/bts139
http://www.ncbi.nlm.nih.gov/pubmed/22437851

85. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.
https://doi.org/10.1093/bioinformatics/btr345
http://www.ncbi.nlm.nih.gov/pubmed/21653521

86. Lane,A.N., Chaires,J.B., Gray,R.D. and Trent,J.O. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482–5515.
https://doi.org/10.1093/nar/gkn517
http://www.ncbi.nlm.nih.gov/pubmed/18718931

87. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V., Šponer,J. and Lankaš,F. (2013) Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
https://doi.org/10.1021/ct300671y

88. Pérez,A., Lankas,F., Luque,F.J. and Orozco,M. (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.*, **36**, 2379–94.
https://doi.org/10.1093/nar/gkn082
http://www.ncbi.nlm.nih.gov/pubmed/18299282

89. Fadrná,E., Špačková,N., Sarzyñska,J., Koča,J., Orozco,M., Cheatham,T.E., Kulinski,T. and Šponer,J. (2009) Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J. Chem. Theory Comput.*, **5**, 2514–2530.
https://doi.org/10.1021/ct900200k

90. Krepl,M., Zgarbová,M., Stadlbauer,P., Otyepka,M., Banáš,P., Koča,J., Cheatham,T.E., Jurečka,P. and Šponer,J. (2012) 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**, 2506–2520.
https://doi.org/10.1021/ct300275s
http://www.ncbi.nlm.nih.gov/pubmed/23197943

91. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring

polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
https://doi.org/10.1093/nar/gks884
http://www.ncbi.nlm.nih.gov/pubmed/23012264

92. Heddi,B., Foloppe,N., Oguey,C. and Hartmann,B. (2008) Importance of Accurate DNA Structures in Solution: The Jun–Fos Model. *J. Mol. Biol.*, **382**, 956–970.
https://doi.org/10.1016/j.jmb.2008.07.047

93. Zgarbová,M., Luque,F.J., Šponer,J., Cheatham,T.E., Otyepka,M. and Jurečka,P. (2013) Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.*, **9**, 2339–2354.
https://doi.org/10.1021/ct400154j

94. Zgarbová,M., Šponer,J., Otyepka,M., Cheatham,T.E., Galindo-Murillo,R. and Jurečka,P. (2015) 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**, 5723–5736.
https://doi.org/10.1021/acs.jctc.5b00716

95. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
https://doi.org/10.1038/nmeth.3658

96. Moroz,J.D. and Nelson,P. (1997) Torsional directed walks, entropic elasticity, and DNA twist stiffness. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 14418–22.
https://doi.org/10.1073/pnas.94.26.14418
http://www.ncbi.nlm.nih.gov/pubmed/9405627

97. Du,Q., Kotlyar,A. and Vologodskii,A. (2008) Kinking the double helix by bending deformation. *Nucleic Acids Res.*, **36**, 1120–8.
https://doi.org/10.1093/nar/gkm1125
http://www.ncbi.nlm.nih.gov/pubmed/18096619

OBJECTIVES

The global objective of this thesis is to advance in the development of methods for the multiscale modeling of DNA from atomistic to sub-chromosomal level. The works presented here represent each a rung in description of DNA dynamics along the resolution ladder underlining the connectivity of each of the simulation methods with its neighbors in terms of model resolution which simplifies defining specific objectives that run like a golden thread through the different topics covered in this thesis.

- **Elucidation of sequence-dependent effects of B-DNA beyond the base pair level**. Using MD simulations with the parmbsc1 force-field, we aimed to develop a complete set of rules at the tetranucleotide level to describe complex polymorphisms in helical space and the correlations between helical conformations and the backbone sub-state at the base, base pair and base pair step level. To examine higher-than-tetranucleotide effects on DNA dynamics we studied the d(CpTpApG) tetranucleotide in different hexa- and octamer environments to uncover the potential influence of specific sequence patterns on long-range conformational changes.

- **Development of a mesoscopic B-DNA model** based on the set of rules derived for DNA dynamics at the tetranucleotide level. Our purpose was to decipher complex structural polymorphisms with the help of machine learning tools to parametrize an extended nearest neighbor helical coarse grain model. We compared the similarity of atomistic reconstituted coarse grain ensembles with MD trajectories and experimentally resolved structures in Cartesian and helical space to test whether the developed model can potentially replace atomistic MD in the study of certain systems at atomistic detail.

- **Development of a webserver using the mesoscopic B-DNA model** to provide simulations of B-DNA - in linear form or in a constrained environment such as supercoiled and protein-coated DNA – in a web environment easily manageable for non-expert users. For additional user friendliness we aimed to provide direct online analysis so that the

generated set of structures can be subject to a large variety of analysis tools within the webserver.

- **Development of a nucleosome fiber model** with base pair resolution which is used to probe chromatin dynamics of fibers with different linker sequence distribution. The flexibility in the choice of the model parameters is targeted to be such that realistic fiber conformations can be deduced directly from experimentally determined *in-vivo* nucleosome positioning data and that additional restraints can be directly applied.

- **Implementation of the mesoscopic B-DNA model and the nucleosome fiber model in the Virtual Research Environment** (VRE) for easy user access and for simplifying communication among the integrated tools which comprise computational modeling and analysis of experiments related to genome architecture from base pair to chromosome level.

Additional to our research objectives we decided to compile current knowledge on **computational approaches in multi-scale DNA modeling** from quantum mechanics to chromosome simulations. Two works were done on this topic, one focusing on providing a compendium of recent advances in computational modeling of DNA while the other one, in form of a book chapter, centralized more in the basic methodological description of the different simulation methods.

CHAPTER II - METHODS

The focus of this thesis is the theoretical description of DNA in the multi-scale regime – from atomistic simulations up to kb long mesoscale modeling (see Figure 17). In this chapter the computational methods used in this thesis are explained in more detail, namely the molecular dynamics (MD) algorithm for atomistic simulations and the Monte Carlos sampling method for the coarse grain helical DNA model and the mesoscopic chromatin model. The nature and the parametrization scheme of the helical coarse grain model developed here is also briefly explained in this chapter. Different strategies to build mesoscale chromatin models are already discussed in Section 3.4 in Chapter I, however I will give a more detailed overview of what has to be taken into account to transfer from a coarse grain to a mesoscale model. The chapter is complemented by the analysis methods I used to bridge the resolution gap between atomistic and mesoscopic modeling.

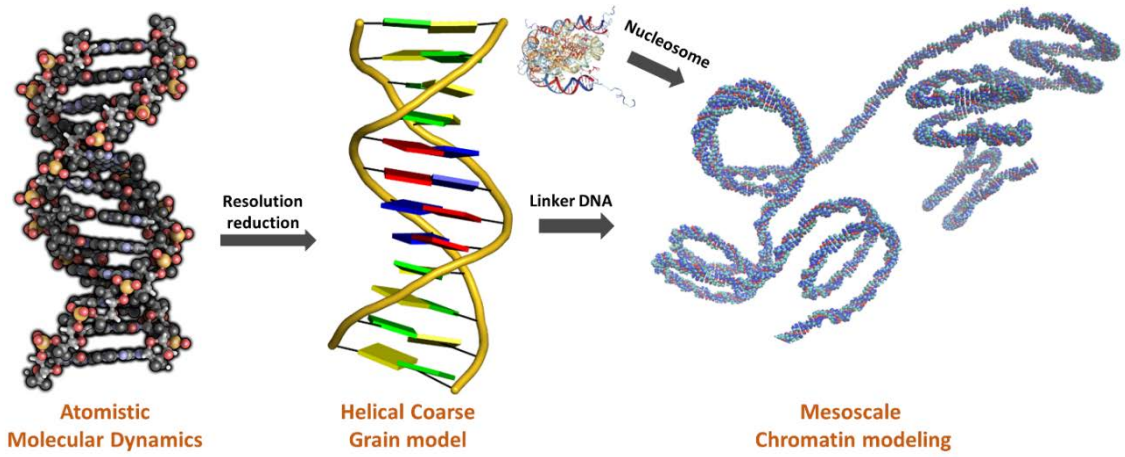


Figure 17. Multi-scale nature of DNA modeling. The pathway from atomistic MD simulations to mesoscopic nucleosome fiber modeling.

1. Molecular Dynamics

1.1 Classical mechanics and force fields

Among the different existing techniques to obtain macromolecular dynamic information, the most popular one is atomistic Molecular Dynamics (MD). Classical mechanics is used to represent atoms as spheres of a given radius, hardness, charge and mass (1). The energy functional used by force-fields is usually composed of two terms: bonded and non-bonded components (Figure 18).

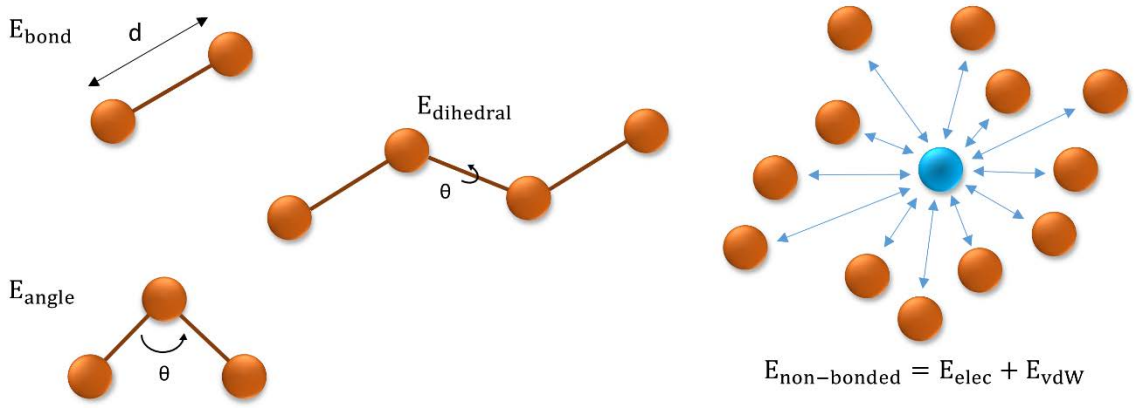


Figure 18. Schematic illustration of the terms in a classical fixed-charge force field, i.e. bond stretching (E_{bond}), bond-angle bending (E_{angle}) and dihedral-angle torsion ($E_{dihedral}$), as well as van der Waals (E_{vdW}) and electrostatic (E_{elec}) interactions

Bonded terms are associated with chemical bond lengths, bond angles and bond dihedrals and non-bonded terms describe electrostatic and van der Waals interactions. The potential energy can be thus written as

$$E_{pot} = E_{bonded} + E_{non-bonded} \quad (3)$$

where

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} \quad (4)$$

and

$$E_{non-bonded} = E_{elec} + E_{vdW} \quad (5)$$

Dissecting the different bonded terms:

- Chemical bonds:

Energy related to the bond length between two atoms. Harmonic potentials are used to approximate the small vibration of covalent bonds around its equilibrium bond length

$$E_{bond} = \sum_{bonds} k_r(r - r_0)^2 \quad (6)$$

where k_r is the bond force constant, r is the observed bond length and r_0 is the reference equilibrium bond length of the atom pair.

- Bond angles:

Energy associated to the angle between two adjacent bonds in a molecule. A harmonic oscillator is used to estimate the energy

$$E_{angle} = \sum_{angle} k_\theta(\theta - \theta_0)^2 \quad (7)$$

where k_θ is the angular force constant, θ is the observed angle, θ_0 is the reference equilibrium bond angle.

- Torsions:

Energy describing the rotation energetic barriers of an atom pair bond (four adjacent atoms define dihedral angle ω). Due to their periodicity dihedral angle potentials cannot be described by harmonic terms, a truncated cosine Fourier expansion is used instead.

$$E_{dihedral} = \sum_{dihedral} \sum_{n} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) \quad (8)$$

where V_n is the height of the barrier (for every term), n the periodicity (usually truncated to 3), ω is the observed dihedral angle and γ the phase angle. Closely related to the torsional interaction are out-of-plane distortions, i.e. the capacity of an atom to be out of the plane formed by the other three atoms involved in the dihedral. These improper dihedral angles can also be accounted for in force fields.

Within the non-bonded terms:

- Electrostatic term:

Energy term associated with the point charges of atoms in a molecule. The electrostatic interaction between two atoms i and j are modeled by a Coulomb potential:

$$E_{elec} = \sum_{i,j} \frac{1}{4\pi\epsilon\epsilon_0} \frac{q_i q_j}{r_{ij}} \quad (9)$$

where ϵ is the dielectric constant of the medium, ϵ_0 the vacuum permittivity and r_{ij} the distance between the two point charges q_i and q_j.

- Van der Waals energy term:

The van der Waals interactions describe the behavior of two atoms when they are approaching each other without forming a covalent bond leading to Pauli repulsion and when they are close to an ideal inter-nuclear distance leading to attraction. Both attraction and repulsion are summarized in one potential, usually described through the Lennard-Jones potential:

$$E_{vdW} = \sum_{i,j} \epsilon^* \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^{6} \right] \quad (10)$$

where ϵ^* is the depth of the potential well, r_m is the distance at which the potential reaches its minimum for the given atom pair i,j and r_{ij} is the distance between atom i and j.

The bonded interactions are computationally inexpensive since they all occur among neighboring atoms. The most expensive part of energy evaluation are the non-bonded terms represented by the Lennard-Jones and Coulomb potential since they theoretically involve all particles in the system. The van der Waals term decays rapidly with distance justifying the use of "cut-offs", however the Coulomb potential falls off slowly, with r^{-1}, and would suffer from major truncation artifacts if a cut-off was imposed (2). To overcome this problem, methods for long-range corrections of the electrostatic potential have been developed, the most common approach is the particle-mesh Ewald (PME) method (3) which shows a good balance between accuracy and computational efficiency. Therefore, the electrostatic energy is divided in two terms, a short-range potential, calculated in the real space, and a long-range potential, which becomes short-ranged when calculated in the Fourier space. Consequently, both terms of PME converge rapidly and an inclusion of a cut-off distance does not impair accuracy. The PME scales with the number of particles N in the order of $O(N \cdot log\, N)$ compared to $O(N^2)$ for direct calculations which facilitates the simulation of larger systems using the PME method.

The PME assumes periodic symmetry of the system in order to perform Fourier transformation. In MD simulations this is achieved by using periodic boundaries to mimic an infinite system where an infinite array of identical copies of the simulation region (unit cell; its shape can be a cube, dodecahedron or truncated octahedron) extends around the unit cell in every direction. The size of the unit cell is chosen to be big enough to avoid that the biomolecule gets too close to the edge. The periodic boundary conditions keep the total particle number constant since any particle that passes through one side of the unit cell reappears on the opposite side.

The fine tuning of the different constants appearing in the above mentioned energy terms is a complex process that requires detailed evaluation of each functional to yield an accurate representation of the DNA's conformational space and thermodynamics.

1.2 Molecular Dynamics algorithm

The time evolution of the system can be generated by using the laws of classical mechanics (Newton's second law of motion). Forces acting on a particle with coordinates X are proportional to the negative gradient of its potential energy U

$$\vec{F}(X) = -\nabla U(X) \tag{11}$$

which leads to that mass m and acceleration a of a particle are proportional to the negative gradient of its potential energy:

$$\vec{F}(X) = m\vec{a}(X) = m\frac{d\vec{v}}{dt} = m\frac{d^2\vec{x}}{dt^2} = -\nabla U(X) \tag{12}$$

Theoretically, system particle coordinates during time could be obtained analytically solving equ. (12). However due to complex particle couplings, the equation needs to be solved numerically. The verlet algorithm and its variants leap-frog and velocity verlet are the most common in MD simulation programs. *Verlet* algorithm (4) calculates the atomic positions r at time t + Δt from the actual positions r(t), the positions from the previous step r(t – Δt) and the accelerations a(t):

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \Delta t^2 a(t) \tag{13}$$

The *Verlet* algorithm does not calculate explicitly velocities, but they can be extracted using different simple approaches.

Leap-frog, a variation of the *Verlet* algorithm, calculates the velocities alongside with the new positions (5):

$$r(t + \Delta t) = r(t) + \Delta t \cdot v\left(t + \frac{1}{2}\Delta t\right) \tag{14}$$

$$v\left(t + \frac{1}{2}\,\Delta t\right) = v\left(t - \frac{1}{2}\Delta t\right) + \Delta t \cdot a(t) \tag{15}$$

Velocities and positions are not synchronized, as they are calculated at time $t + {}^{1}/_{2}\ \Delta t$ and $t + \Delta t$, respectively. The *velocity verlet algorithm* allows calculating velocities and positions at the same time:

$$r(t + \Delta t) = r(t) + \Delta t \cdot v(t) + \frac{1}{2}\Delta t^2 a(t) \quad (16)$$

$$v(t + \Delta t) = v(t) + \frac{1}{2}\Delta t\,[a(t) + a(t + \Delta t)] \quad (17)$$

In all the algorithms the integration time step Δt is crucial since the acceleration is assumed to be constant during that time. Choosing a too large time step might cause instabilities of the macromolecular system while too small integration time steps increase the computational time of sampling the movement. In atomistic MD simulations the integration time step is chosen not to be bigger than the smallest motion in the system which for biological systems is the bond stretching involving hydrogen atoms occurring on the 1 fs timescale. Seen from the biological level, these vibrations are irrelevant for the final results of the simulation and special algorithms such as SHAKE (6), LINCS (7) or RATTLE (8) exist to constrain the smallest vibrational movements to allow longer integration time steps (from 1 fs to 2 fs) which in turn speeds up the simulation.

As mentioned above, MD simulations rely on Newton's equation of motion where the energy of the system E, the volume V and the number of particles N is conserved, known as microcanonical ensemble (NVE ensemble). However, to capture conditions closer to experiments, pressure and temperature need to be kept constant and ensembles such as canonical or isothermal-isobaric give a more appropriate description. In the canonical ensemble (NVT ensemble) the total energy is allowed to vary, but the system is maintained at constant temperature by means of a thermostat. Popular techniques to control temperature include weakly coupled algorithms like velocity rescaling, Berendsen thermostat, Nose-Hoover thermostat and Langevin dynamics (9–12). Additionally to the condition of constant temperature, most experiments are performed in an environment where pressure is invariable with varying volume, so experimental conditions can be best reproduced with an isothermal-isobaric ensemble (NPT). Similarly to the thermostat the pressure can be controlled by the Berendsen, Nose-Hoover or Parrinello-Rahman barostat.

Generally, while stochastic models lack reproducibility of the trajectory, deterministic algorithms tend to lose ergodicity and can equilibrate very slowly.

The environment of most biomolecular MD simulations is made of water and ions with their representations being as important as the representation of the studied biomolecule. On the one hand, the solvent can be represented implicitly as a continuous medium by approximating the mean force exerted by the media on the solute while on the other hand water and ions can be explicitly included in the system. Implicit water models have substantial advantages in computing time, however they neglect specific important features such as hydrogen bond fluctuations at the solute surface, water dipole reorientation in response to conformational changes and bridging water molecules. Therefore it is common to use explicit water models for simulations of biomolecules up to a certain size, among the most popular ones are TIP3P (13) and SPC/E (14). Together with different models of monovalent ions (15, 16) the impact of using distinct water and ion representations in MD simulations on local and global fiber properties is a current field of study (see section 5 in Chapter I).

1.3 DNA Force-field

To correctly evaluate the generic force-field formula

$$\mathrm{E_{total}} = \mathrm{E_{bond}} + \mathrm{E_{angle}} + \mathrm{E_{dihedral}} + \mathrm{E_{elec}} + \mathrm{E_{vdW}} \qquad (18)$$

parameters for each energy term have to be carefully evaluated for each atom type to reproduce realistic DNA dynamics. Atom type assignment depends on the functional group the atom is part of and/or hybridization state.

The first simulation of a DNA molecule took place in the 80’ and since then MD simulations have been improving rapidly with the first microsecond simulation of a Drew-Dickerson dodecamer (17). This fast development goes along with the increase in computational power where faster processors, bigger supercomputers and advances in GPU technologies allow the study of larger biomolecular systems at a longer time scale. This in turn can create new issues in force-field parametrization never observed before on shorter time scales (18–21) which are ought to be improved.

The basic parametrization scheme relies on the transferability approximation where the force-field is parametrized on a small set of molecules and then applied to a wider group of molecules with similar chemical groups. This assumes that parameters are not dependent on the local environment and that, in case of DNA, the parametrization of the four nucleotide units should be enough to simulate any DNA molecule. Current parametrization efforts respond to errors detected in previous force-fields, most commonly a QM study on a small system is directly compared with the same system simulated with MD to refit the parameters. This approach is used in the AMBER or CHARMM families of force-fields (20, 22–27). While the advantage of this strategy is its accuracy due to the QM study, potential neighboring effects could be neglected due to system size. In other approaches experimental data is used as additional restraints to fit the force-field for different macromolecules (28–30) which by definition results in good reproducibility of the experimental data at the expense of universality of the description of different forms of the macromolecule.

Although being a stable biomolecule, DNA's flexibility and charged nature makes it difficult to simulate. In the physics point of view, two forces balance the DNA structure: strong electrostatic repulsion between the phosphates in the backbone and attractive stacking and hydrogen bonding between nucleobases. Changes in solvent environment can additionally affect these two forces and thus influence the shape of DNA.

Different force-fields emerged since the first MD simulation of a DNA molecule by Levitt, with parm99 (24) being in our opinion the first reliable force-field for DNA which could correctly simulate DNA dynamics of timescales up to 50 ns. However, with the increase in computational power problems at longer simulation timescales arose. Distortions in the structure were related to disproportionate α/γ populations from the canonical *gauche-/gauche+* state towards *gauche+/trans*. Those problems were corrected by the parmbsc0 force-field (25) which allowed stable DNA simulations in the multi-nanosecond regime. For a decade, parmbsc0 became the 'gold-standard' for DNA simulations with over 1500 citations (up to 02/2019). Nevertheless, some issues still remained and with the steady increase in computational power, challenges in multi-microsecond simulations came up. Experimental values of helical parameters, especially those of roll and twist, were misestimated (parmbsc0 undertwists the structure by, in average, 3° (17)). The BI/BII-equilibrium, which has been shown to correlate with the bimodality of the twist

distribution, especially for RpY steps such as d(CpG), is biased towards the canonical BI state (17, 21, 31). The fraying of terminal bases was very large, generating unrealistic configurations at the ends of the DNA (32).

Corrections to the parmbsc0 force-field involved reparametrization for exotic DNA forms or the attempt to create a new standard for DNA simulations. For example, the OL1 parameter set (26) tackled the ε/ζ representation for more accurate BI/BII populations, OL4 (20) aimed to correct the χ distribution, followed by OL15 which included all previous OL corrections for DNA and incorporated additional adjustments of the β torsion (33). Efforts from MacKerell's group incorporate a Drude's oscillator term accounting for polarisation effects (the energy contribution arising from the mutual relaxation of electron distribution of interacting particles (34)). The resulting force field (CHARMM36pol) is the first polarizable force field able to reproduce some nucleic acids structures (like the B-DNA duplex) in the 100 ns regime. However, when simulations are extended to the microsecond regime even canonical DNA structures are corrupted (35), so more work is still needed to recalibrate all the terms to the incorporation of polarization.

Even though the modifications for parmbsc0 could correct some issues, not all the problems of DNA simulations were addressed yet, which motivated the development of a universal force-field for DNA simulations. The efforts resulted in a new force-field, parmbsc1, which constitutes a major improvement to previous force-fields in the study of dynamic properties of DNA (see section 5 in Chapter I).

2. Parametrization of helical coarse grain model

Coarse grain (CG) models which reduce the complexity of the system are used to achieve longer time and length scales in DNA simulation than those accessible to MD. The 'internal CG model' (see Section 3.3 in Chapter I for a summary of different approaches to construct a CG model) uses internal degrees of freedom to describe DNA dynamics, usually rigid base pairs being the smallest unit of the model (36, 37) (some models assume rigid bases (38–40) which implies a slightly different parametrization procedure). The motion between two rigid base pairs consists of the six inter base pair parameters (three translational (shift, slide, rise) and three rotational (tilt, roll,

twist) parameters are considered in this model for each base pair step (bps)). The DNA deformations can be approximated as a sum of harmonic distortions (36) from equilibrium bps geometries and the Hamiltonian can be described as:

$$\mathrm{E} = \sum_{\mathrm{j}=1}^{\mathrm{N}-1} \mathrm{E}^{\mathrm{j}} \qquad (19)$$

where E is the total energy of DNA deformations, N the number of base pairs in the DNA oligomer (the sum expands over all bps) and E^j is the individual deformation energy of bps j calculated as

$$\mathrm{E}^{\mathrm{j}} = \Xi^{\mathrm{j}} \left(\mathrm{X}^{\mathrm{j}} - \mathrm{X}_0^{\mathrm{j}}\right)^2 \text{ with } \Xi = \mathrm{k_B TC^{-1}} = \begin{pmatrix} \mathrm{k_f} & \mathrm{k_{lf}} & \mathrm{k_{sf}} & \mathrm{k_{tf}} & \mathrm{k_{rf}} & \mathrm{k_{wf}} \\ \mathrm{k_{fl}} & \mathrm{k_l} & \mathrm{k_{sl}} & \mathrm{k_{tl}} & \mathrm{k_{rl}} & \mathrm{k_{wl}} \\ \mathrm{k_{fs}} & \mathrm{k_{ls}} & \mathrm{k_s} & \mathrm{k_{ts}} & \mathrm{k_{rs}} & \mathrm{k_{ws}} \\ \mathrm{k_{ft}} & \mathrm{k_{lt}} & \mathrm{k_{st}} & \mathrm{k_t} & \mathrm{k_{rt}} & \mathrm{k_{wt}} \\ \mathrm{k_{fr}} & \mathrm{k_{lr}} & \mathrm{k_{sr}} & \mathrm{k_{tr}} & \mathrm{k_r} & \mathrm{k_{wr}} \\ \mathrm{k_{fw}} & \mathrm{k_{lw}} & \mathrm{k_{sw}} & \mathrm{k_{tw}} & \mathrm{k_{rw}} & \mathrm{k_w} \end{pmatrix} \qquad (20)$$

where k_B is the Boltzmann constant, T is the absolute temperature, X^{j} is the vector of inter base pair parameters of bps j, X_0 is the equilibrium inter base pair geometry, E^j is the energy of base pair step j associated with the deformation $\mathrm{X}^{\mathrm{j}} - \mathrm{X}_0^{\mathrm{j}}$ and k_{XY} stands for the different stiffness constants defined by the 36 elements of the stiffness matrix (Ξ) (shift (f), slide (l), rise (s), tilt (t), roll (r), twist (w)). The stiffness matrix can be calculated (see equ (20)) by inversion of the helical covariance matrix C obtained from either the analysis of MD simulations at dinucleotide or tetranucleotide level (18, 41) or from the analysis of dinucleotide step variability in crystal structures of DNAs and DNA-protein complexes (31, 36). Early elastic models rely on the use of a nearest neighbor representation of DNA (36, 37, 41) (10 stiffness matrices and equilibrium conformations based on the unique bps). However, recent works revealed that the flanking base pairs of a bps influence the dynamics of the central bps so that a bps approach is not sufficient to describe DNA dynamics and the bps in its tetranucleotide environment has to be considered (nearest neighbor) (41–47). Recent studies have also raised concerns on the use of the purely harmonic approach (21, 31, 48, 49), as DNA samples different conformational sub-states which motivates the development of an extension of the nearest neighbor model that uses multi-state harmonicity in each tetranucleotide.

3. Parametrization of nucleosome fiber model

The transition from a CG DNA model to a mesoscopic nucleosome fiber model brings several challenges (Figure 19). CG DNA models in most cases do not consider long-range interactions since it is highly unlikely that distant DNA segments in sequence get close in three dimensional space for the DNA length scales accessible to a CG model. In nucleosome fiber models there are two crucial differences compared to CG models of naked DNA: first, the size of the model is much bigger (low to high kb-scale) which makes it necessary to represent fiber parts in a more coarse fashion, and secondly, with the nucleosome a complicated DNA-protein complex is involved, so long-range potentials become important to avoid physical overlap of fiber constituents (50, 51).

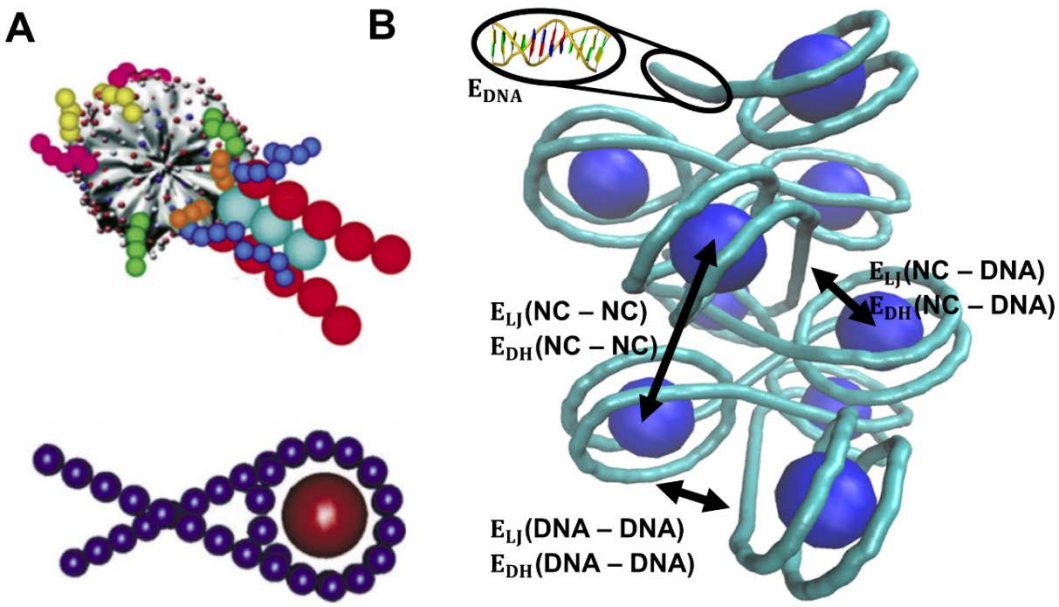


Figure 19 Parametrization of nucleosome fiber model A: Different representation of the nucleosome. Top: Detailed DISCO model from Schlick group (52). Bottom: Spherical nucleosome core (55) B: Energy terms to consider in the nucleosome fiber model of DNA and nucleosome core (NC).

To tackle the first issue - as pointed out in Section 3.4 of Chapter I - nucleosome fiber models are built based on the 'bottom-up' approach which means that known physical properties of DNA and nucleosome core particle are mapped onto a more coarse level to derive appropriate geometry

and potentials of the constituents. Two choices have to be taken when building a nucleosome fiber model: (1) the level of detail of representing the nucleosome and (2) the representation of the linker DNA. In most models the coarse grain nucleosome representation is based on its experimentally derived structure (more details on the nucleosome see Section 2.1 of Chapter I). Pioneering work such as Schlick's model uses an irregular surface with a set of Debye-Hückel charges of the nucleosome core defined by a discrete surface charge optimization algorithm (52), together with flexible histone tails and different linker histones (53) while in other models the whole nucleosome core is represented by simple objects such as a cylinder (six angle model; (54)) or as a sphere ((55) and the model presented here). Appropriate inter-nucleosomal potentials have to be developed depending on the present nucleosome geometry. Most nucleosome fiber models choose to model the linker DNA as a worm-like chain (WLC) (55, 56).

In this treatment, a DNA segment is represented as an elastic chain of N beads, each bead comprising M base pairs, connected by N-1 inter-bead segments of average length l. The WLC model for linker DNA usually comprises stretching, bending and torsional energy between two consecutive beads, the associated stretching, bending and torsional constants for those quadratic energy terms are derived based on average DNA fiber properties. This means that with a WLC model sequence-dependent description of linker DNA flexibility is typically not considered. Other approaches choose a 'helical CG' model with base pair resolution to represent the linker DNA to be able to accurately describe its sequence-dependent dynamic behavior (57).

To prevent physical overlap between the constituents of a nucleosome fiber (assuming they are spherical) a CG potential is applied usually in the form of a Lennard-Jones potential:

$$\mathrm{E_{LJ}} = \sum_{\mathrm{i,j}} \epsilon^* \left[\left(\frac{\mathrm{r_m}}{\mathrm{r_{ij}}} \right)^{12} - 2 \left(\frac{\mathrm{r_m}}{\mathrm{r_{ij}}} \right)^{6} \right] \qquad (21)$$

where ϵ^* is the depth of the potential well, r_m is the distance at which the potential reaches its minimum for the given constituent pair i,j and r_{ij} is the distance between the geometric center of constituent i and j.

In normal environmental conditions DNA and the nucleosome cores are charged (the DNA has a total negative charge due to the phosphate and the nucleosome core is positively charged). To

account for electrostatic interactions within the nucleosome fiber a Debye-Hückel potential (58) is usually applied:

$$\mathrm{E_{DH}} = \frac{1}{4\pi\epsilon\epsilon_0}\sum_{\mathrm{i,j}} \mathrm{q_i q_j}\frac{\mathrm{e^{-\kappa r_{ij}}}}{\mathrm{r_{ij}}} \qquad (22)$$

where ϵ_0 is the electric permeability of vacuum, ϵ is the dielectric constant (set to 80), r_{ij} is the distance between the geometric centers of fiber constituents i and j, q_i and q_j are the associated charges and κ is the inverse Debye length (the decay length of electrostatic interactions in a solution with physiological concentration of monovalent salt). In the case of interaction between nucleosomes more complex potentials combining prevention of physical overlap and electrostatics can be used depending on the geometric representation of the nucleosome in the model (52, 54, 56). The total energy of a general nucleosome fiber model is then

$$\mathrm{E} = \mathrm{E_{DNA}} + \mathrm{E_{LJ}} + \mathrm{E_{DH}} \qquad (23)$$

where E_{DNA} represents the internal energy associated to linker DNA (see equ. (19)), E_{LJ} the total excluded volume energy and E_{DH} the total electrostatic contribution by DNA-DNA, DNA-nucleosome and nucleosome-nucleosome interactions.

Nucleosome fiber models are usually parameterized to match compaction levels found in experiments for example by *in-vitro* experimental data such as compaction of medium-sized chromatin fibers by sedimentation coefficient measurements (59), EMANIC (60) or X-ray (61).

4. Monte Carlo algorithm

The sampling of internal CG models and many nucleosome fiber models is done via Markov Chain Monte Carlo algorithm (only a few use Langevin MD (50)). This type of algorithm predicts a new configuration based solely in its present state (Markov process). In a Markov Chain Monte Carlo move a set of parameters of the current fiber configuration is changed in a random way and the suggested configuration is accepted or not based on its energy (see below). The set of coordinates

subject to change can be of different nature (Cartesian, internal, helical). The magnitude of the change attempted is also random, but constrained within reasonable limits to avoid extremely high rejection rates. Determination of these limits is non-trivial when local helical coordinates are used (see Section 2.1 and 3 in Chapter III). In models using non Cartesian coordinates, like those presented in Section 2.1 and 3 in Chapter III (based on local helical coordinates), the internal move can be mapped back to Cartesian coordinates for the possibility to simultaneously evaluate energy potentials in internal and Cartesian space after the Monte Carlo move.

Irrespectively of the coordinate system and the energy functional used, current Monte Carlo approaches follow the Metropolis approach (62) and the attempted new configuration is accepted or rejected based on the energy difference $\Delta E = E_{new} - E_{previous}$ between the new and the previous configuration. In case of $\Delta E < 0$, the new configuration is accepted and used to generate a new potential Monte Carlo move. If the potential energy of the new configuration is higher than the one of the previous configuration two possibilities exist: either the new configuration gets rejected and another Monte Carlo move on the original configuration is carried out or the new configuration can still get accepted based on the Metropolis criterion

$$u \leq e^{-\frac{\Delta E}{k_B T}} \qquad (24)$$

where u is a random number between 0 and 1, k_B the Boltzmann constant and T the temperature. The Metropolis criterion is usually applied in Markov Chain Monte Carlo models since it allows to escape local potential energy minima in the sampling space to reach an extensively sampling of the whole conformational space.

5. Analysis

In the previous section the methodology and the algorithm to generate a series of structures (trajectory) of MD and CG/mesoscopic simulations were ruled out. In this section I will introduce the main analysis tools I used to analyze time ensembles from MD simulations or probabilistic ensembles from CG/mesoscopic models. When coarse grain DNA structures can be mapped back to an 'atomistic' representation equivalent analysis at the atomistic level can be performed for

MD and coarse grain trajectories. Due to the ergodic theorems, time and probabilistic ensembles behave the same way in terms of ensemble average properties.

5.1 RMSd – Root Mean Square Deviation

RMSd is the value that quantifies the minimum deviation of atomic positions of a given structure X from those of a reference structure Y.

$$\mathrm{RMSd} = \min\left(\sqrt{\sum_{i=1}^{N}(\vec{x}_i - \vec{y}_i)^2}\right) \quad (25)$$

where $\vec{x}_i$ and $\vec{y}_i$ are the coordinates of each of the N selected equivalent atoms in structure X and reference structure Y. The RMSd between two structures is minimized by a simple least squares fitting algorithm and the reference structure is usually an experimental structure or the first structure of the trajectory. Consequently, RMSd can serve as an indicator of structural stability along simulation time and as a similarity measure with experimental data.

5.2 Radius of gyration

The radius of gyration measures the compactness of a system. It is defined as the mass-weighted distance of each particle from the center-of-mass of the structure:

$$R_g = \sqrt{\frac{\sum_{i=1}^{n} m_i d_i^2}{\sum_{i=1}^{n} m_i}} \quad (26)$$

where m_i is the mass of particle i and d_i is the Euclidian distance of particle i to the center-of-mass. The radius of gyration has many different fields of application, for example it can be used to compare the compactness of chromatin of fibers with different linker length distributions.

5.3 Principal component analysis

To describe the major movements of DNA in a simulation the principal component analysis (PCA) method is able to separate major motions from thermal fluctuations. The natural motions of a structure, the principal components, explain most of the variance of the trajectory. Since PCA helps extracting *essential* motions of DNA it is often called *'Essential Dynamics'.* The principal components are derived from the covariance matrix C of the atomic positional fluctuations:

$$\mathrm{C} = \mathrm{cov}(\mathrm{X}) = \langle \Delta\mathrm{X}\Delta\mathrm{X}^{\mathrm{T}} \rangle \tag{27}$$

with

$$\Delta\mathrm{X} = \mathrm{X} - \mathrm{X}_{\mathrm{ref}} \tag{28}$$

where X is the set of atomic coordinates in matrix form of a given structure, X_{ref} is a reference value (usually the average structure of the trajectory), ΔX^T is the transpose of ΔX and the angle brackets $\langle \cdot \rangle$ represent averaging over the distribution.

The principal components can be obtained by diagonalization of the covariance matrix C

$$\Lambda = A^T C A \tag{29}$$

where Λ is the diagonalized covariance matrix with eigenvalues λ, the n-th column of the transformation matrix A corresponds to the eigenvector with eigenvalue λ_n. The eigenvectors are the principal components and the corresponding eigenvalues represent the percentage of variance explained by each eigenvector.

Essential dynamics can be used to determine similarity between trajectories (63). For example, Hess' metrics accumulate the dot products between a reduced set of eigenvectors of two trajectories A and B and the absolute similarity index γ_{AB} can be defined as

$$\gamma_{AB} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(v_i^A v_j^B\right)^2 \tag{30}$$

where n is the number of eigenvectors used and v_i^X stands for the i-th unitary eigenvector of trajectory X. A more sophisticated similarity measure is Perez's metrics which account for the

different energy contributions of the essential movements by weighting the accumulated inner products between two eigenvectors (63). Essential dynamics can also be used to reduce the complexity of the system by only considering a certain number of the most significant eigenvectors. PCA analysis can also be used in internal space, for example PCA on the inter base pair parameter space (6-dimensional) of a given base pair step helps to uncover complex correlations among the parameters.

5.4 Distance matrix

A distance matrix shows the Cartesian distance of the chosen constituents of a macromolecular system. The diagonal entries in the distance matrix comprise self-interaction and have zero distance by definition while the off-diagonal entries can show complex three dimensional structural arrangements in two dimensional matrix form. This dimension reduction method is used for example for long-range contacts in DNA fibers or regions with higher nucleosome content in chromatin fibers. If an ensemble of structures is subject to analysis, different parameters such as the mean, minimum or maximum distance in matrix form can shed light on structural arrangements of the ensemble.

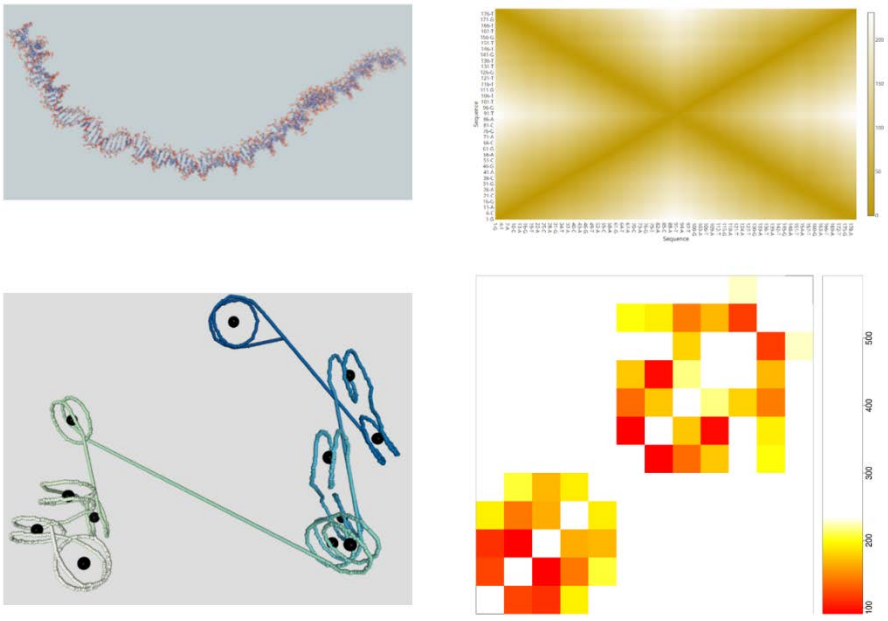


Figure 20. Distance matrices. Top: DNA structure (left) and its corresponding distance matrix (right) counting from the center of each nucleobase. Bottom: Nucleosome fiber structure (left) and distance matrix of the center of each nucleosome core (right).

5.5 Solvent accessible surface area

Solvent accessible surface area (SASA) is designated as the region of the surface of a molecule exposed enough to be able to interact with solvent molecules. It is usually defined as a surface built by the delineation drawn by the center of a sphere (rough presentation of a solvent molecule, usually radius of 1.4 Å, approximating the size of a water molecule) rolling over the molecular surface (see Figure 21). SASA values are obtained in this thesis using the well-known software called NACCESS (64). SASA can be used for nucleic acid fiber to allow an easy identification of the fragments of DNA affected by the attached proteins, not only in the regions where they are docked, but also in protein-free regions that see their accessibility hindered by these proteins.

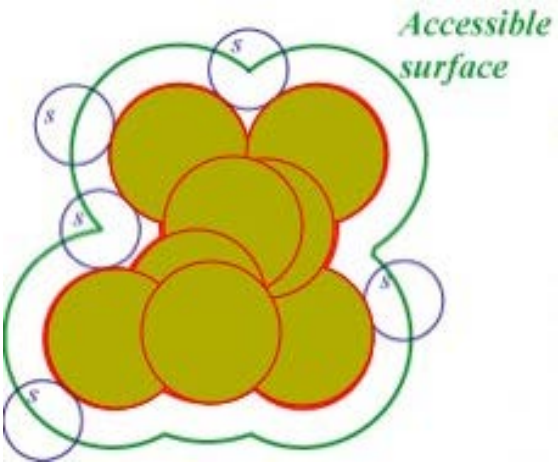


Figure 21. Solvent Accessible Surface Area (SASA). Atoms of the molecule are represented as van der Waals spheres, and the solvent accessible surface area is defined by the center of a rolling ball (S) representing the solvent molecule while going over the molecule surface.

5.6 Hydrogen bonds

The electrostatic attractive interaction between a proton in one molecule (acceptor) and an electronegative atom (donor) in another atom is a hydrogen bond (HB). The strength of interaction of HBs is higher than for van der Waals interaction, but weaker than covalent bonds. HBs are crucial for the 3D arrangement adopted by a macromolecular system since its intra-molecular HBs define and maintain the structure. As the strength of HBs is sensitive to orientation and distance between donor and acceptor, HBs are determined by a defined cut-off distance

between donor and acceptor, 3.5 Å in our case, and a cut-off angle, 120° in our case. Studies of HB dynamics (breaking/formation) give quantitative information about conformational reshaping of the secondary structure.

5.7 Helical analysis

DNA trajectories can also be analyzed by means of its internal coordinates in the helical space. A helical axis is fitted as a curvilinear line in the direction of the propagation of the helix and the position and orientation of the bases within a base pair can be defined by ten internal coordinates (see Section 1.2.1 in Chapter I for details). The motion between two base pairs is assumed by only three translations and three rotations. This set of parameters can be obtained from a MD or 'atomistic' reconstituted coarse grain trajectory using Curves+ software (65) which provides information on the helical parameters, groove geometry and backbone conformation (in terms of dihedral torsion angles). Curves+ can be used to assess the quality of trajectories in atomistic resolution and compare important parameters among different trajectories either in the internal inter base pair space or by evaluating backbone properties such as torsional angles, BI/BII state population or groove widths and groove depths. From the set of helical parameters obtained by trajectory analysis via Curves+ flexibility constants of a base pair step depending on their sequence environment can be derived and used in lower resolution coarse grain models (for more details see Section 2 in Chapter II). Similar analysis programs such as 3DNA exist (66), nevertheless we use Curves+ as standard analysis program due to its versatility.

Additionally, CANION (67), a new module of Curves+, can be used to calculate the positions of any atom in curvilinear helicoidal coordinates with respect to the helical axis. Using the CANION extension, several studies can be performed, for example to examine the impact of different ion concentrations in the grooves on DNA flexibility or to determine the position of the phosphate in the backbone relative to the helical axis, which turned out to be crucial for mapping helical models to the Cartesian atomistic level.

5.8 Bending

The bending of a segment of DNA quantifies the curvature of the fiber. The total bending angle of a segment of k base pairs starting from base pair n is obtained by (68)

$$B_n^{tot}(k) = \sqrt{b_n^x(k)^2 + b_n^y(k)^2} \qquad (31)$$

where $b_n^x(k)$ and $b_n^y(k)$ are the bending contributions in the xz- and yz-plane (the directions of x-,y- and z-axis are determined by the triad of the first base pair in the segment; for definitions of the axes see Section 1.2.1 in Chapter I). The individual bending contributions can be calculated based on the rotational inter base pair parameters tilt, roll and twist as

$$b_n^x(k) = \sum_{i=n}^{k+n-1} \rho_i \cos \Phi_i + \sum_{i=n}^{k+n-1} \tau_i \sin \Phi_i \qquad (32)$$

$$b_n^y(k) = \sum_{i=n}^{k+n-1} \rho_i \sin \Phi_i + \sum_{i=n}^{k+n-1} \tau_i \cos \Phi_i \qquad (33)$$

where ρ_i, τ_i and Φ_i are the values of tilt and roll of base pair step i and cumulative twist summed from base pair step n to i. Parameter k is usually chosen to be 5 base pairs (half turn) or 10 base pairs (full turn) while to calculate the total bending of a DNA fiber of N base pairs in length the parameters n and k have to be set to n=1 and k=N. Bending calculations can be used to compare DNA distortions along the sequence in a free or constrained environment such as supercoiled or protein-bound.

5.9 Persistence length

The persistence length is the decay length through which the memory of the initial orientation of a polymer chain persists. DNA and chromatin fibers are assumed to behave similar to theoretical polymeric chains, so their persistence length can be calculated by different polymer physics formulas. In general the persistence length of a polymer chain is calculated as

$$\langle u(s) \cdot u(s') \rangle = e^{-\frac{l}{l_p}}, \qquad \forall\, s, s' \tag{34}$$

where $\langle \cdot \rangle$ stands for averaging over thermal realizations, $u(s)$ and $u(s')$ are the tangent unit vectors to the chain separated by a curvilinear distance $l = |s' - s|$, and l_p denoting the persistence length (69). The persistence length can also be calculated by the mean-square end-to-end distance $\langle R^2 \rangle$ of a polymer (70)

$$\langle R^2 \rangle = 2 l_p L \left[1 - \frac{l_p}{L} \left(1 - e^{-\frac{L}{l_p}} \right) \right] \tag{35}$$

where L is the length of the polymer and l_p the apparent persistence length. Formulas (34) and (35) are usually used to calculate the apparent persistence length of structures without taking into account their potentially bent intrinsic shape.

For DNA due to its non-linear intrinsic shape the persistence length can be split in two parts, the dynamic and the static persistence length l_d and l_s (71). The static persistence length represents the intrinsic distorted shape of DNA, the dynamic persistence length the DNA's thermal fluctuations. l_s is calculated by using formula (34) for the relaxed ground state of the DNA fiber, so no averaging over thermal realizations is needed. Knowing l_s and the apparent persistence length l_p, the dynamic persistence length can be obtained via

$$\frac{1}{l_p} = \frac{1}{l_s} + \frac{1}{l_d} \tag{36}$$

Further expansions of formula (36) lead to a sequence-averaged description of persistence length (69). The apparent persistence length of DNA is roughly 50 nm or 150 bp. To calculate the persistence length the length of the underlying chain is usually in the order or several multiples of the order of the persistence length, however some persistence length calculations were done on DNA fibers of 30-50 base pairs in length (72).

Bibliography for Chapter II

1. Lifson,S. and Warshel,A. (1968) Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and *n* -Alkane Molecules. *J. Chem. Phys.*, **49**, 5116–5129.
https://doi.org/10.1063/1.1670007

2. Saito,M. (1994) Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation. *J. Chem. Phys.*, **101**, 4055–4061.
https://doi.org/10.1063/1.468411

3. Darden,T., York,D. and Pedersen,L. (1993) Particle mesh Ewald: An *N* ·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
https://doi.org/10.1063/1.464397

4. Verlet,L. (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, **159**, 98–103.
https://doi.org/10.1103/PhysRev.159.98

5. Hockney and W.,R. (1970) The potential calculation and some applications.

6. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J.. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
https://doi.org/10.1016/0021-9991(77)90098-5

7. Hess,B., Bekker,H., Berendsen,H.J.C. and Fraaije,J.G.E.M. (1997) LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, **18**, 1463–1472.

https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

8. Andersen,H.C. (1983) Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, **52**, 24–34.
https://doi.org/10.1016/0021-9991(83)90014-1

9. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
https://doi.org/10.1063/1.448118

10. Nosé,S. and Shuichi (1984) A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, **81**, 511–519.
https://doi.org/10.1063/1.447334

11. Hoover (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A, Gen. Phys.*, **31**, 1695–1697.
http://www.ncbi.nlm.nih.gov/pubmed/9895674

12. Brooks,C.L., Brünger,A. and Karplus,M. (1985) Active site dynamics in protein molecules: A stochastic boundary molecular-dynamics approach. *Biopolymers*, **24**, 843–865.
https://doi.org/10.1002/bip.360240509
http://www.ncbi.nlm.nih.gov/pubmed/2410050

13. Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
https://doi.org/10.1063/1.445869

14. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
https://doi.org/10.1021/j100308a038

15. Smith,D.E. and Dang,L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
https://doi.org/10.1063/1.466363

16. Joung,I.S. and Cheatham,T.E. (2008) Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B*, **112**, 9020–9041.
https://doi.org/10.1021/jp8001614

17. Pérez,A., Luque,F.J. and Orozco,M. (2007) Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.*, **129**, 14739–14745.
https://doi.org/10.1021/ja0753546
http://www.ncbi.nlm.nih.gov/pubmed/17985896

18. Pérez,A., Lankas,F., Luque,F.J. and Orozco,M. (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.*, **36**, 2379–94.
https://doi.org/10.1093/nar/gkn082
http://www.ncbi.nlm.nih.gov/pubmed/18299282

19. Fadrná,E., Špačková,N., Sarzyñska,J., Koča,J., Orozco,M., Cheatham,T.E., Kulinski,T. and Šponer,J. (2009) Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J. Chem. Theory Comput.*, **5**, 2514–2530.
https://doi.org/10.1021/ct900200k

20. Krepl,M., Zgarbová,M., Stadlbauer,P., Otyepka,M., Banáš,P., Koča,J., Cheatham,T.E.,

Jurečka,P. and Šponer,J. (2012) 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**, 2506–2520.
https://doi.org/10.1021/ct300275s
http://www.ncbi.nlm.nih.gov/pubmed/23197943

21. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V., Šponer,J. and Lankaš,F. (2013) Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
https://doi.org/10.1021/ct300671y

22. MacKerell,A.D., Bashford,D., Bellott,M., Dunbrack,R.L., Evanseck,J.D., Field,M.J., Fischer,S., Gao,J., Guo,H., Ha,S., *et al.* (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J. Phys. Chem. B*, **102**, 3586–3616.
https://doi.org/10.1021/jp973084f

23. Huang,J. and MacKerell,A.D. (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.*, **34**, 2135–2145.
https://doi.org/10.1002/jcc.23354

24. Cheatham,T.E., Cieplak,P. and Kollman,P.A. (1999) A Modified Version of the Cornell *et al.* Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
https://doi.org/10.1080/07391102.1999.10508297

25. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.*, **92**, 3817–3829.
https://doi.org/10.1529/biophysj.106.097782

http://www.ncbi.nlm.nih.gov/pubmed/17351000

26. Zgarbová,M., Luque,F.J., Šponer,J., Cheatham,T.E., Otyepka,M. and Jurečka,P. (2013) Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.*, **9**, 2339–2354.
https://doi.org/10.1021/ct400154j

27. Zgarbová,M., Otyepka,M., Šponer,J., Mládek,A., Banáš,P., Cheatham,T.E. and Jurečka,P. (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*, **7**, 2886–2902.
https://doi.org/10.1021/ct200162x

28. Mackerell,A.D. (2004) Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.*, **25**, 1584–1604.
https://doi.org/10.1002/jcc.20082

29. Yildirim,I., Stern,H.A., Kennedy,S.D., Tubbs,J.D. and Turner,D.H. (2010) Reparameterization of RNA chi Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine. *J. Chem. Theory Comput.*, **6**, 1520–1531.
https://doi.org/10.1021/ct900604a
http://www.ncbi.nlm.nih.gov/pubmed/20463845

30. Gil-Ley,A., Bottaro,S. and Bussi,G. (2016) RNA Conformational Ensembles: Narrowing the GAP between Experiments and Simulations with Metadynamics. *Biophys. J.*, **110**, 522a-523a.
https://doi.org/10.1016/j.bpj.2015.11.2796

31. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678. https://doi.org/10.1093/nar/gks884
http://www.ncbi.nlm.nih.gov/pubmed/23012264

32. Zgarbová,M., Otyepka,M., Šponer,J., Lankaš,F. and Jurečka,P. (2014) Base Pair Fraying in Molecular Dynamics Simulations of DNA and RNA. *J. Chem. Theory Comput.*, **10**, 3177–3189.
https://doi.org/10.1021/ct500120v

33. Zgarbová,M., Šponer,J., Otyepka,M., Cheatham,T.E., Galindo-Murillo,R. and Jurečka,P. (2015) 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**, 5723–5736. https://doi.org/10.1021/acs.jctc.5b00716

34. Savelyev,A. and MacKerell,A.D. (2014) All-atom polarizable force field for DNA based on the classical drude oscillator model. *J. Comput. Chem.*, **35**, 1219–1239. https://doi.org/10.1002/jcc.23611

35. Dans,P.D., Ivani,I., Hospital,A., Portella,G., González,C. and Orozco,M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, gkw1355. https://doi.org/10.1093/nar/gkw1355
http://www.ncbi.nlm.nih.gov/pubmed/28088759

36. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 11163–8. https://doi.org/10.1073/pnas.95.19.11163
http://www.ncbi.nlm.nih.gov/pubmed/9736707

37. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.*, **85**, 2872–2883. https://doi.org/10.1016/S0006-3495(03)74710-9 http://www.ncbi.nlm.nih.gov/pubmed/14581192

38. Lankaš,F., Gonzalez,O., Heffler,L.M., Stoll,G., Moakher,M. and Maddocks,J.H. (2009) On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, **11**, 10565. https://doi.org/10.1039/b919565n http://www.ncbi.nlm.nih.gov/pubmed/20145802

39. Petkevičiūtė,D., Pasi,M., Gonzalez,O. and Maddocks,J.H. (2014) cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.*, **42**, e153. https://doi.org/10.1093/nar/gku825 http://www.ncbi.nlm.nih.gov/pubmed/25228467

40. Dršata,T., Zgarbová,M., Špačková,N., Jurečka,P., Šponer,J. and Lankaš,F. (2014) Mechanical Model of DNA Allostery. *J. Phys. Chem. Lett.*, **5**, 3831–3835. https://doi.org/10.1021/jz501826q

41. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2010) 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**, 299–313. https://doi.org/10.1093/nar/gkp834 http://www.ncbi.nlm.nih.gov/pubmed/19850719

42. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H., *et al.* (2005) Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps. *Biophys. J.*, **89**, 3721–3740.
https://doi.org/10.1529/biophysj.105.067397
http://www.ncbi.nlm.nih.gov/pubmed/16169978

43. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
https://doi.org/10.1093/nar/gku855
http://www.ncbi.nlm.nih.gov/pubmed/25260586

44. Imeddourene,A. Ben, Xu,X., Zargarian,L., Oguey,C., Foloppe,N., Mauffret,O. and Hartmann,B. (2016) The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res.*, **44**, 3432–47.
https://doi.org/10.1093/nar/gkw084
http://www.ncbi.nlm.nih.gov/pubmed/26883628

45. Ben Imeddourene,A., Elbahnsi,A., Guéroult,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLOS Comput. Biol.*, **11**, e1004631.
https://doi.org/10.1371/journal.pcbi.1004631

46. Tian,Y., Kayatta,M., Shultis,K., Gonzalez,A., Mueller,L.J. and Hatcher,M.E. (2009) ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J. Phys. Chem. B*, **113**, 2596–2603.
https://doi.org/10.1021/jp711203m

47. Zgarbová,M., Jurečka,P., Lankaš,F., Cheatham,T.E., Šponer,J. and Otyepka,M. (2017) Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Inf. Model.*, **57**, 275–287.
https://doi.org/10.1021/acs.jcim.6b00621

48. Maehigashi,T., Hsiao,C., Woods,K.K., Moulaei,T., Hud,N. V and Williams,L.D. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, **40**, 3714–22.
https://doi.org/10.1093/nar/gkr1168
http://www.ncbi.nlm.nih.gov/pubmed/22180536

49. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
https://doi.org/10.1093/nar/gku809
http://www.ncbi.nlm.nih.gov/pubmed/25223784

50. Dans,P.D., Walther,J. and Gómez,H. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45.
https://doi.org/10.1016/J.SBI.2015.11.011

51. Gómez,H., Walther,J., Darré,L., Ivani,I., Dans,P.D. and Orozco,M. (2017) Chapter 7. Molecular Modelling of Nucleic Acids. In.pp. 165–197.
https://doi.org/10.1039/9781788010139-00165

52. Zhang,Q., Beard,D.A. and Schlick,T. (2003) Constructing irregular surfaces to enclose macromolecular complexes for mesoscale modeling using the discrete surface charge

optimization (DISCO) algorithm. *J. Comput. Chem.*, **24**, 2063–2074.
https://doi.org/10.1002/jcc.10337
http://www.ncbi.nlm.nih.gov/pubmed/14531059

53. Bascom,G.D., Myers,C.G. and Schlick,T. (2019) Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub. *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 4955–4962.
https://doi.org/10.1073/pnas.1816424116
http://www.ncbi.nlm.nih.gov/pubmed/30718394

54. Stehr,R., Schöpflin,R., Ettig,R., Kepper,N., Rippe,K. and Wedemann,G. (2010) Exploring the Conformational Space of Chromatin Fibers and Their Stability by Numerical Dynamic Phase Diagrams. *Biophys. J.*, **98**, 1028–1037.
https://doi.org/10.1016/J.BPJ.2009.11.040

55. Kimura,H., Shimooka,Y., Nishikawa,J., Miura,O., Sugiyama,S., Yamada,S. and Ohyama,T. (2013) The genome folding mechanism in yeast. *J. Biochem.*, **154**, 137–147.
https://doi.org/10.1093/jb/mvt033

56. Schlick,T. and Perisić,O. (2009) Mesoscale simulations of two nucleosome-repeat length oligonucleosomes. *Phys. Chem. Chem. Phys.*, **11**, 10729–37.
https://doi.org/10.1039/b918629h
http://www.ncbi.nlm.nih.gov/pubmed/20145817

57. Kulaeva,O.I., Zheng,G., Polikanov,Y.S., Colasanti,A. V., Clauvelin,N., Mukhopadhyay,S., Sengupta,A.M., Studitsky,V.M. and Olson,W.K. (2012) Internucleosomal Interactions Mediated by Histone Tails Allow Distant Communication in Chromatin. *J. Biol. Chem.*, **287**, 20248–20257.
https://doi.org/10.1074/jbc.M111.333104

58. Debye P,H.E. (1923) The theory of electrolytes. I. Lowering of freezing point and related phenomena. *Phys. Zeitschrift*, **24**, 185–206.

59. Hansen,J.C., Ausio,J., Stanik,V.H. and van Holde,K.E. (1989) Homogeneous reconstituted oligonucleosomes, evidence for salt-dependent folding in the absence of histone H1. *Biochemistry*, **28**, 9129–36.
http://www.ncbi.nlm.nih.gov/pubmed/2605246

60. Grigoryev,S.A., Arya,G., Correll,S., Woodcock,C.L. and Schlick,T. (2009) Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci.*, **106**, 13317–13322.
https://doi.org/10.1073/pnas.0903280106

61. Schalch,T., Duda,S., Sargent,D.F. and Richmond,T.J. (2005) X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, **436**, 138–141.
https://doi.org/10.1038/nature03686
http://www.ncbi.nlm.nih.gov/pubmed/16001076

62. Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, **21**, 1087–1092.
https://doi.org/10.1063/1.1699114

63. Alberto Pérez,†,‡, José Ramón Blas,†, Manuel Rueda,†,§, Jose María López-Bes,‡, Xavier de la Cruz,†,‖ and and Modesto Orozco*,†,§,⊥ (2005) Exploring the Essential Dynamics of B-DNA. 10.1021/CT050051S.
https://doi.org/10.1021/CT050051S

64. Hubbard SJ,T.J. (1993) NACCESS.

65. Blanchet,C., Pasi,M., Zakrzewska,K. and Lavery,R. (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.*, **39**, W68-73.
https://doi.org/10.1093/nar/gkr316
http://www.ncbi.nlm.nih.gov/pubmed/21558323

66. Lu,X.-J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
https://doi.org/10.1093/nar/gkg680
http://www.ncbi.nlm.nih.gov/pubmed/12930962

67. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–49.
https://doi.org/10.1093/nar/gku504
http://www.ncbi.nlm.nih.gov/pubmed/24906882

68. Battistini,F., Hunter,C.A., Gardiner,E.J. and Packer,M.J. (2010) Structural Mechanics of DNA Wrapping in the Nucleosome. *J. Mol. Biol.*, **396**, 264–279.
https://doi.org/10.1016/j.jmb.2009.11.040

69. Mitchell,J.S., Glowacki,J., Grandchamp,A.E., Manning,R.S. and Maddocks,J.H. (2017) Sequence-Dependent Persistence Lengths of DNA. *J. Chem. Theory Comput.*, **13**, 1539–1555.
https://doi.org/10.1021/acs.jctc.6b00904

70. Moukhtar,J., Vaillant,C., Audit,B. and Arneodo,A. (2011) Revisiting polymer statistical physics to account for the presence of long-range-correlated structural disorder in 2D DNA chains. *Eur. Phys. J. E*, **34**, 119.

https://doi.org/10.1140/epje/i2011-11119-3

71. Trifonov,E.N., Tan,R.K.Z. and Harvey,S.C. (1988) Static persistence length of DNA. *Struct. Expr. Proc. Fifth Conversat. Discip. Biomol. Stereodyn. held State Univ. New York Albany, June 2-6, 1987 / Ed. by M.H. Sarma R.H. Sarma*.

72. Noy,A. and Golestanian,R. (2012) Length Scale Dependence of DNA Mechanical Properties. *Phys. Rev. Lett.*, **109**, 228101. https://doi.org/10.1103/PhysRevLett.109.228101

CHAPTER III - RESULTS

1. Sequence-dependent properties of B-DNA and structural polymorphisms

Previous studies showed that DNA properties depend on the underlying sequence, elucidating polymorphism that deviate from canonical Arnott's B-DNA description. A typical example is the d(CpG) step, which shows a clear bimodality in the rotational inter base pair parameter twist. Such a polymorphism is strongly coupled to changes in the backbone conformational states (BI/BII), a phenomenon occurring at sub-µs time scale (1, 2), which seems to be coupled to the presence of ions in the grooves.

Several groups, including our own reported deviations from normality of many internal degrees of freedom (1–4) which recently started to get confirmed by experimental evidence (3, 5–7), giving rise to the assumption that different internal degrees of freedom lead to different sequence-dependent conformational sub-states.

However, even though some general observations can be made from experimental data such as crystal or NMR DNA structures, the experimentally resolved set of structures is fragmented and inappropriate for exploring sequence-dependent properties of DNA at a large scale. With increase in computational power, the number and length of MD simulations could be increased so that DNA in many different sequence environments can be extensively studied. In a former study of all the 136 distinct tetranucleotide base sequences, carried out by the ABC (Ascona B-DNA Consortium), nearest neighbor effects of the central base pair step were derived from parmbsc0 simulations, finding non-Gaussian and multi-peaked helical parameter distributions for certain base pair steps and correlations were established between the inter base pair parameters of the central junction of a tetranucleotide and the ζ backbone torsion (8). As the results were obtained using the parmbsc0 force field there are still open questions concerning the sequence-dependent equilibrium distribution of helical parameters and backbone sub-state populations. For this reason, we performed a new set of simulations with the state-of-the-art parmbsc1 force field,

covering the same tetranucleotide space (called miniABC sequence set). In this first work presented in this chapter, we used this set of multi-µs MD simulations to decipher the sequence-dependent polymorphic space at unprecedented detail. We characterized the choreography of backbone and base movements modulating the non-Gaussian or anharmonic effects and the polymorphisms in helical geometries which are particularly present in certain tetranucleotide sequence contexts. These findings allow us to reformulate Calladine-Dickerson rules at the tetranucleotide level.

In a second study, we analyzed in more details the results of the sequence effects of all unique tetranucleotides to pinpoint several cases of tetramers with unusual polymorphic behavior, such as low stability/high flexibility, multimodality in the helical parameter space and high sensitivity to sequence context. Those tetramers comprise only 3-5% of all unique tetranucleotides, indicating that multimodality might have a moderate impact in overall duplex properties, but can be very important to explain local flexibility of certain DNA motifs. We decided to investigate higher-than-nearest-neighbor effects of the d(CpTpApG) tetranucleotide (from now on CTAG), a sequence showing unusual flexibility behavior in the trajectories deposited in our BigNAsim database (8, 9). A potential issue of studying hexa- and octameric effects of the CTAG tetranucleotide could have been the limited sampling, however systematic analysis shows that higher order effects were not an artifact of non-convergence of the simulation. After eliminating this uncertainty, we found non-negligible next-to-nearest neighbor effects of different sequence contexts. Based on the concerted and correlated movements of bases and backbone torsions for the described multi-modal degrees of freedom, and driven by the mechanical limitations imposed by DNA's crankshaft motions, we were able to found a possible explanation on how structural information can travel almost half a helical turn away from the central d(TpA) step. This remote structural 'connection' allows the d(TpA) step to 'feel' its sequence environment beyond the next-to-nearest neighbors, and eventually adopts a different substate if needed which could be important in the cell nucleus where CTAG has been preserved with a low rate of mutation suggesting a possible mechanical role for CTAG at the genomic level.

1.1 Nearest-neighbor effects of DNA dynamics (Publication 1)

We took advantage of the quality of parmbsc1 to analyze the complete tetranucleotide space of B-DNA via MD simulations. The sequence library was designed to minimize the number of short oligomers needed to cover all 136 unique 4-mers. The in-depth analysis shows different equilibrium distributions of intra base pair parameters that are close to harmonic while inter base pair parameter distributions can experience multimodality, most commonly slide for d(GpG), twist for d(CpG) and d(ApG) and shift for YR. In general, shift bimodality is coupled to the appearance of high shift values of above 1Å. Twist bimodality experiences a more complex behavior, as in some cases the second peak of the distribution occurs at higher than canonical values (> 40°), while in other cases it is at low twist values (< 30°).

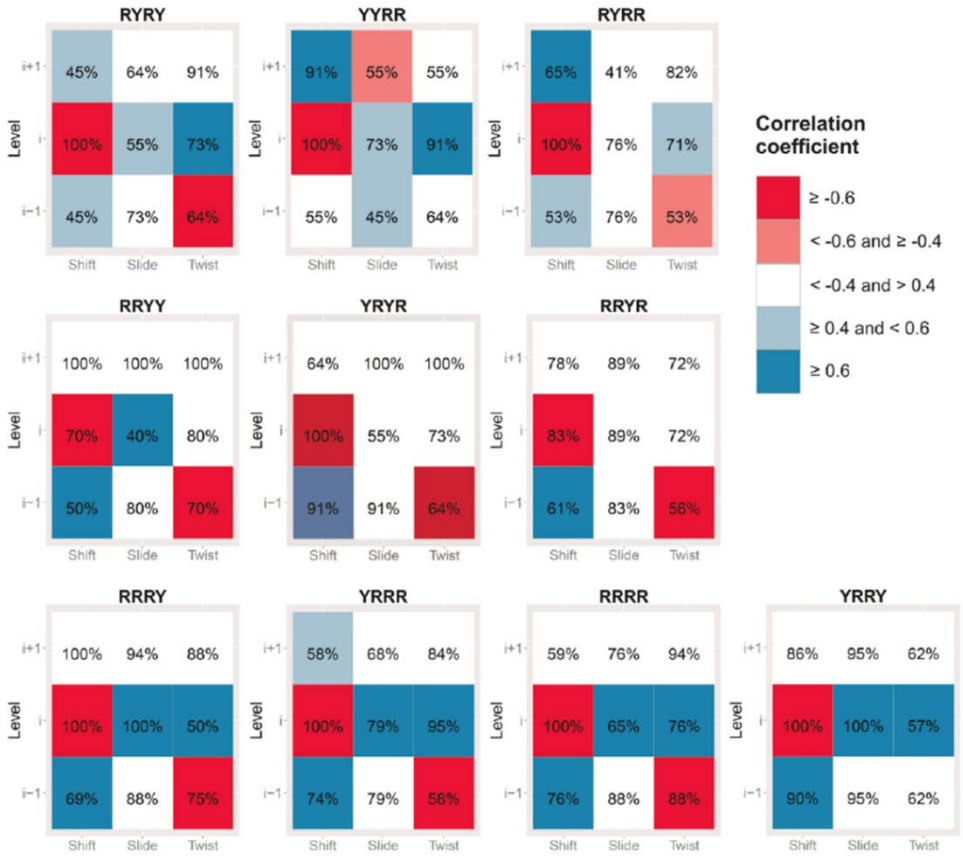


Figure 22. Correlation coefficients between shift, slide, or twist at the positions i-1 (5'-side), i, and i-1 (3'-side), and the backbone sub-state at the junction of base-pair step i in the Watson strand. Results obtained from miniABCBSC1-K dataset. The numbers inside each cell represent the % of specific tetranucleotides within a given family that give rise to the correlation.

In an analysis of the backbone conformations we find that BI -> BII transitions are strongly dependent on the underlying sequence (see Figure 4 in the following publication). For example, RR backbones exhibit quite high BII percentages, especially in the presence of Y at the 5' end of the corresponding tetranucleotide while YY backbones are typically biased towards the BI state.

Connecting the backbone polymorphism with the base pair conformations, with ε/ζ (BI/BII) and inter base pair parameters being the major players, we find that tetranucleotides showing simultaneous sampling of BI and BII conformations are those with bimodality in some inter base pair parameter at the same step (see Supplementary Table S4 and S5 in the following publication). The BI/BII state also correlates with inter base pair helical coordinates in the same and neighboring junctions. For example, the increase in the percentage of BII at the central junction of a given tetranucleotide correlates with larger shift values for all sequences (Supplementary Figures S17 in the following publication) and is also coupled to lower twist and slide values. The BI/BII ratio at a junction i also correlates with shift, twist and slide values at step i+1 and i-1 (see Figure 22), highlighting the subtle mechanical coupling between backbone and base pair step conformations within DNA.

In summary, with this complete study we can formulate some general rules concerning the equilibrium conformation distribution of B-DNA, which represent a significant step beyond Calladine-Dickerson earlier propositions:

- The first and second moments (average and covariance) of the equilibrium distributions of helical coordinates for DNA can only be understood in terms of nonlocal sequence-dependence contexts

- A harmonic model of DNA dynamics will not be able to accurately predict third and higher moments of the equilibrium distribution because significant anharmonic movements arise frequently

- Backbone torsional changes are coordinated in pairs (α/γ, P/χ and ε/ζ). Coordinated changes in the ε/ζ pair lead to the BI/BII polymorphism with coupled impacts on helical parameters. Both ε/ζ and P/χ couplings exhibit sequence dependence.

- The BI/BII conformational change is coupled to the cationic atmosphere surrounding DNA, and to the formation of non-canonical CH---O hydrogen bonds.

- Helical parameters at a given step are not independent, but show a complex backbone-mediated pattern of dependencies.

- Calladine's principles, and Dickerson's algorithms for twist/roll/δ/propeller, can now be transformed into quantitative predictions for all the structural features (helical conformations and backbone substates) of canonical DNA sequences. These extended rules have been implemented on a web server that predicts the average conformation of any B-DNA sequence, in terms of the average helical parameters, base and backbone polymorphisms, and P/χ conformations (see https://mmb.irbbarcelona.org/miniABC/).

Publication:

Pablo D. Dans, Alexandra Balaceanu, Marco Pasi, Alessandro S. Patelli, Daiva Petkevičiūtė, Jürgen Walther, Adam Hospital, Richard Lavery, John H. Maddocks, and Modesto Orozco; The Physical Properties of B-DNA beyond Calladine's rules, Nucleic Acids Research (accepted)

THE PHYSICAL PROPERTIES OF B-DNA BEYOND CALLADINE-DICKERSON RULES

Pablo D. Dans[a,b,1], Alexandra Balaceanu[a,b,2], Marco Pasi[c,d,2], Alessandro S. Patelli[e,2], Daiva Petkevičiūtė[e,f,2], Jürgen Walther[a,b,2], Adam Hospital[a,b], Richard Lavery[d], John H. Maddocks[e,1], and Modesto Orozco[a,b,g,1]

[a]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.
[b]Joint BSC-IRB Research Program in Computational Biology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.
[c]LBPA, École normale supérieure Paris-Saclay, 61 Av. du Pdt Wilson, Cachan 94235, France.
[d]Bases Moléculaires et Structurales des Systèmes Infectieux, Univ. Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, Lyon 69367, France.
[e]Institute of Mathematics, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland.
[f]Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Studentų g. 50, 51368 Kaunas, Lithuania.
[g]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain.

[1]To whom correspondence should be addressed:
Dr. Pablo D. Dans, Tel: +34 934039073, Email: pablo.dans@irbbarcelona.org; or
Prof. John H. Maddocks, Tel: +41 216932762, Email: john.maddocks@epfl.ch; or
Prof. Modesto Orozco, Tel: +34 934037155, Email: modesto.orozco@irbbarcelona.org.

[2]These co-authors equally contributed to this work and were alphabetically sorted.

ABSTRACT

We present a multi-laboratory effort to describe the physical properties of duplex B-DNA under physiological conditions. By processing a large amount of data from atomistic molecular dynamics simulations, we determine the sequence-dependent structural properties of DNA as expressed in the equilibrium distribution of its stochastic dynamics. Our analysis includes a study of first and second moments (or mean and covariance) of the equilibrium distribution, which can be accurately captured by a Gaussian, or harmonic, model, but with nonlocal sequence-dependence. We then further characterize the sequence-dependent choreography of backbone and base movements modulating the non-Gaussian or anharmonic effects manifested in the higher moments of the dynamics of the duplex when sampling the equilibrium distribution. Contrary to prior assumptions, such anharmonic deformations are not rare in DNA and can play a significant role in determining DNA conformation within complexes. Polymorphisms in helical geometries are particularly prevalent for certain tetranucleotide sequence contexts, and are always coupled to a complex network of coordinated changes in the backbone, with BI/BII equilibria being a major determinant. The analysis of our simulations, which contain instances of all 136 distinct tetranucleotide sequences, allow us to reformulate Calladine-Dickerson rules, used for decades to interpret the average geometry of DNA according to presumed local sequence-dependence and harmonic fluctuations, in a more precise manner, leading to an extended set of rules with quantitative predictive power that encompass nonlocal sequence-dependence and anharmonic fluctuations.

SIGNIFICANCE STATEMENT

The article represents the latest effort of the ABC consortium (https://bisi.ibcp.fr/ABC) on the characterization of the sequence-dependent physical properties of B-DNA under physiological conditions. Taking advantage of our recently developed force field (PARMBSC1), and the coordinated effort of the ABC laboratories, we were able to derive general rules concerning the equilibrium conformation of B-DNA, which represent a significant step beyond Calladine-Dickerson earlier propositions. We are now able to predict the appearance of subtle sequence-dependent sub-states at the base and backbone level that arise as a function of tetranucleotide sequence context. The extended Calladine-Dickerson rules presented herein can be transformed into quantitative predictions of the structural features of any canonical B-DNA sequence.

INTRODUCTION

DNA is a flexible and structurally polymorphic polymer whose overall equilibrium geometry strongly depends on its sequence, the solvent environment, and the presence of ligands(*1*, *2*). Conformational changes in DNA are mediated by a complex choreography of backbone rearrangements such as the BI/BII transition(*3*, *4*), the low-twist/high-twist equilibrium(*5*, *6*), or concerted α/γ rotations(*7*–*9*). Such backbone rearrangements lead to local and global changes in the helix geometry(*9*, *10*) impacting on the ability of the DNA to recognize ligands(*11*), and consequently on its functionality.

Binding-induced conformational changes in DNA are required for function, and are expected to follow the sequence-dependent intrinsic deformation modes of DNA, *i.e.* are implicitly coded in the spontaneous deformability of isolated DNA. This suggests that evolution has refined DNA sequence not only to maximize ligand-DNA interactions, but also to reduce the energetic cost of moving from a canonical to a bioactive conformation(*11*, *12*). This leads the notion of "indirect readout", which suggests that the ability of the DNA to adopt the "bioactive" conformation plays a major role in determining the target sequences of a given DNA ligand. Understanding the sequence-dependent physical properties of DNA then becomes crucial to rationalizing how ligands and, most notably, proteins, recognize and modulate DNA activity, *i.e.* the structural basis of gene regulation.

Understanding the sequence-dependent physical properties of DNA has been traditionally hampered by the lack of experimental data. Using simple steric considerations and geometric constraints Calladine developed in 1982 a set of principles to describe the mechanics of DNA(*13*), which have been used for decades to gain some qualitative insight into the sequence-dependence of

expected, or average, local helical geometry. In their original version, those principles suggested that clashes between bases are avoided by a combination of concerted change in twist, roll, and slide, as the base pair propeller increases to improve stacking(*13*). One year later, Dickerson formulated a simple numerical algorithm allowing for a quantification of Calladine's principles, coining the procedure as the "Calladine's Rules"(*14*, *15*). The algorithm could be used to predict correctly the local variation (at the base pair level) in twist, roll, propeller twist and the torsion angle δ for a few B-DNA sequences which were previously determined experimentally(*14*). Unfortunately, the extension and predictive power of these rules, even in the most recent versions, is limited(*1*, *16*). Attempts to gain more quantitative information were based on the analysis of the variability in local helical parameters in structural databases(*17*, *18*), but to date*, isolated B-DNA structures in the Nucleic Acid Databank (NDB) allowed us to obtain flexibility data for only 5 of the 136 distinct tetranucleotides (only AATT, CGCG, CGAA, GCGA and ATTC are represented more than 500 times). Even when the database is extended by including protein-DNA complexes, the sampling is not dense enough to describe sequence-dependent DNA flexibility at the tetranucleotide level (24 out of the 136 tetranucleotides are still represented less than 500 times). In this context, atomistic molecular dynamics (MD) simulations are the only alternative to obtain robust and transferable parameters(*10*, *19*, *20*).

The first requirement for deriving physical descriptors of DNA from MD simulations is the availability of extended simulations for a library of sequence fragments containing all distinct tetranucleotides. This requires a significant computational effort which has encouraged joint projects such as the Ascona B-DNA Consortium (ABC, https://bisi.ibcp.fr/ABC), which have been instrumental, not only in describing physical properties of DNA, but also in refining simulation protocols(*10*, *21–23*). The second major requirement is the availability of accurate force fields, such as the recently developed PARMBSC1(*24*), which has been shown to represent DNA with a quality indistinguishable from experimental measurements(*25*). Thanks to the coordinated effort of several ABC groups, a series of microsecond-scale simulations on a library of DNA duplexes covering all of the 136 distinct tetranucleotides have been performed, and with a number of different simulation conditions *e.g.* using PARMBSC0(*26*) or PARMBSC1, different counter ions, *etc*. Consequently there is a minimum of six total simulations of each independent tetranucleotide. The analysis of this large ensemble of data allows us to not only decipher the rules defining the sequence-dependent equilibrium

*Data from the NDB (http://ndbserver.rutgers.edu/) on the 19th March 2018. We found 727 PDBs with the search string: "Polymer Type: DNA Only + Structural Features: B DNA + Experimental Method: All"; and 3434 PDBs searching for: "Polymer Type: Protein DNA Complexes + Protein Function: All + Structural Features: B DNA + Experimental Method: All". After removing non-canonical and terminal bases, 10,134 tetranucleotides remained in the B-DNA ensemble, and 155,316 tetranucleotides in the Prot-DNA set. Watson and Crick strands were both taken into account, and no filters were applied to reduce the known high redundancy of the database.

3

geometry of B-DNA, but also those determining coordinated backbone conformational changes, and the correlations between various helical deformations. A new, and comprehensive extension of Calladine-Dickerson rules emerges from the analysis of these simulations, including the first predictions of anharmonicity based on sequence context.

METHODS

The choice of sequences. The new ABC sequence library was designed to optimize the number of relatively short oligomers needed to include one copy of each of the distinct 136 tetranucleotides. Applying an adapted version of the Orenstein and Shamir algorithm(*27–29*), we generated 13 oligomers, each containing 18 base pairs (including GC terminals in each end), covering the complete tetranucleotide space (see Table S1 for a list of the designed sequences), and 117 (of the 2080 possible) distinct hexanucleotide sequences. The smaller number of oligomers with respect to previous training libraries(*6*, *10*) made it more practical to obtain multi-microsecond trajectories under several simulation conditions (for example, using both the PARMBSC1(*24*) and PARMBSC0(*30*) force fields, labeled miniABC$_{BSC1}$ and miniABC$_{BSC0}$ respectively), and by changing the ionic environment (from KCl to NaCl, labeled miniABC$_{BSC1}$-K and miniABC$_{BSC1}$-Na respectively). Comparison of results obtained with this library of sequences (miniABC) with respect to the standard ABC-set (μABC(*10*)) allowed us to check for the robustness of our conclusions as a function of the duplexes from which the tetranucleotide parameters were derived.

System preparation and MD simulations. All oligonucleotides were constructed with the *leap* program of AMBERTOOLS 15(*31*) and simulated using the *pmemd.cuda* code(*32*) from AMBER14(*31*), following the standard ABC protocol(*10*). Additional details are provided in Suppl. Material. Trajectories are accessible at the BIGNASim server: https://mmb.irbbarcelona.org/BIGNASim/.(*33*)

Analysis. Trajectories were processed with the *cpptraj*(*34*) module of the AMBERTOOLS 15 package(*31*), and the NAFlex server(*35*) for standard analysis. DNA helical parameters and backbone torsion angles were measured and analyzed with the CURVES+ and CANAL programs(*36*), following the standard ABC conventions(*10*). Duplexes were named following the Watson strand. The letters R, Y and X stand for a purine, a pyrimidine, or any base respectively; base pairs flanking a dinucleotide step were denoted using two dots to represent the central step (*e.g.* R..Y), and one dot when trinucleotides are considered (*e.g.* R.Y), while X:X and XX represent an intra-basepair and inter-basepair (step) respectively. Bayesian Information Criterion (or BIC)(*37*, *38*) was used to quantify the normal or binormal (*i.e.* a mixture of two normals) nature of the distributions of the helical

parameters (see Suppl. Methods). An extension of Helguerro's theorem(*39*, *40*) was used to distinguish those binormal distributions where the two Gaussians are very close (unimodal distributions) from those where they are significantly separated (bimodal distributions). Correlation between backbone and helical parameters was analyzed by clustering the backbone conformations into discrete states using standard rules as described in Suppl. Methods. The similarity between first and second moments (*i.e.* average and covariance) of the helical parameter distributions for different simulation libraries was evaluated using the Kullback-Leibler (KL) divergence, as detailed in the Suppl. Material. More specifically sequence-dependent Gaussian coarse grain cgDNA(*41–43*) model parameters were computed from each of the four MD training libraries used in this work (*i.e.* μABC$_{BSC0}$-K, miniABC$_{BSC0}$-K, miniABC$_{BSC1}$-K, miniABC$_{BSC1}$-Na) in order to be able to generate associated predictions of first and second moments of the helical parameters for fragments of arbitrary sequence. In particular this allowed us to compare PARMBSC0 simulations of the μABC library with the PARMBSC0 simulations of the miniABC library, even though the two libraries have different sequence fragments. See the Supporting Methods for more details.

RESULTS AND DISCUSSION

Sources of uncertainty: the sequence library and the type of salt. Before going into detail with a conformational analysis, we first considered the robustness of our results to changes in the choice of sequence library, because large differences would challenge the general validity of our conclusions. Fortunately, only one of the 1,632 distributions analyzed (namely of 6 intra- plus 6 inter-basepair helical parameters for each of the 136 distinct tetranucleotides), showed significant differences (according to BIC-Helguerro analysis) depending on the choice of library (the previous μABC library, or the current miniABC library; see Suppl. Figure S1). Furthermore, no differences were found depending on the salt (see Table S2 and raw data in https://mmb.irbbarcelona.org/miniABC/), which suggests that our results are robust to the choice between K and Na for the counter-ion. To gain additional confidence in the robustness of our results, we used the explicit form of Kullback-Leibler divergence available for Gaussian (*i.e.* multi-variate normal) distributions to quantify three pairwise differences in cgDNA model predictions (see Methods, and Suppl. Methods) of the mean and covariance for each of the 13 miniABC library sequences for the four different parameter sets extracted from the μABC$_{BSC0}$-K, miniABC$_{BSC0}$-K, miniABC$_{BSC1}$-K, and miniABC$_{BSC1}$-Na simulations. As can be seen in Figure 1, no significant difference arises from the change in sequence library, nor from the difference between K and Na counter ions. However, the results are quite sensitive to the change in force field from PARMBSC0 to PARMBSC1. This is to be expected since the latest PARMBSC1 force field leads to a considerably more realistic representation of

twist/roll and BI/BII distributions (see the analysis and discussion published elsewhere(*9*, *25*)), and to straighter average configurations of duplexes than those obtained from prior force fields. This can be confirmed by considering the differences between static and dynamic persistence lengths (as introduced elsewhere(*44*)) over a large ensemble of sequences (see Suppl. Figure S2).

Strong anharmonic distortions do arise. One of the most important extreme deformations of DNA is the disruption of base pairing, which can be analyzed in detail by aggregating data over all instances of G:C and A:T base pairs . This allowed us to accumulate ensembles on the millisecond time scale. Terminal base pairs (G:C pairs in all the cases) showed open states (water molecules in between H-bonding Watson-Crick groups) in 1-2% of the total simulation time, with short average open life times (around 3 ns, see Table S3) in agreement with time-resolved Stokes shifts spectroscopy(*45*), but most probably too short to lead to isotope exchange signals in NMR experiments(*46*). The opening of central base pairs is less likely to occur (between 0.01% in G:C and 0.05% in A:T of the simulation time), but when it happens, the open state can survive considerably longer (up to 50 ns). Whether or not this time is sufficient to allow proton interchange with the solvent is unclear. Another example of a strong anharmonic deformation arising in our simulations is the temporary formation of a sharp kink (Suppl. Figure S3) associated with anomalous rise and roll(*47*) at an AA inter-basepair within a TAAA tetranucleotide belonging to a relatively long tract of A:T base pairs (seq. 9, see Table S1). Very interestingly, this deformation has been characterized before as one of the origins of bubbling and kinking in natural DNA(*48*, *49*), but to our knowledge, has not been previously observed in atomistic simulations.

Equilibrium distributions of intra-basepair deformations are close to harmonic. A BIC analysis was carried out for the distributions of all six of the intra-basepair helical parameters at the central base pair in all 64 possible trinucleotide contexts. These distributions were all observed to be rather close to Gaussian, *cf.* Figure S4, with the exception of exceptional rare events, as discussed in the last paragraph. Certainly no multi-peaked distribution was ever observed. Nevertheless the average value, or first moment, of each of the six intra-basepair parameters is strongly sequence-dependent to at least the trinucleotide sequence context, see Figure 2. Some qualitative rules on the sequence-dependent variation in the means can be observed. Shear values in G:C intra-basepairs, when G is followed by Y are below average, while the opposite happens for A:T base pairs. Buckle in G:C shows large variations depending on the nature of the 3'-base of G, with an R leading to large positive buckles, and a Y leading to large negative buckles. Propeller also shows clear sequence rules, with A:T intra-basepairs having a sizeable negative value when there is an R 5' to the A, while propeller is close to zero for G:C base pairs within YGR trinucleotides.

Equilibrium distributions of inter-basepair deformations are frequently strongly anharmonic. Bi-normality (*i.e.* deviation from Gaussianity) in the equilibrium distributions of the inter-basepair helical coordinates is common, but clear bimodality (*i.e* the appearance of distinct multiple peaks) is observed in only 3% (miniABC$_{BSC1}$-K+) to 5% (miniABC$_{BSC1}$-Na+) of the inter-basepair helical distributions (Figure 3 and Suppl. Figure S5). Bimodality appears systematically only for slide (several tetranucleotides containing a central GG step), shift (typically in a few tetranucleotides containing a YR central step) and twist (mainly in tetranucleotides containing central CG or AG steps). These conclusions are completely compatible with our prior analysis of PARMBSC0 simulations (see the μABC work(*10*), particularly Figure 8). There are few cases where bimodality affects simultaneously two or more helical parameters, for example, AGGA and GGGA are bimodal in shift and slide (in agreement with experimental data(*50*)) and ACGG, GCGA and GCGG are bimodal in shift and twist in agreement with results derived from the data mining of PDB structures(*5*). The central step of the GTAA tetranucleotide is the only case displaying bimodality in three helical parameters (shift, slide and twist) simultaneously. In general, shift bimodality is coupled to the appearance of high-shift values (above 1 Å). The reverse situation was found for slide, where bimodality displaces the distribution to lower values. Finally twist bimodality displays more complex behavior, as in some cases the second peak of the distribution occurs at lower than canonical values (< 30°), while in others it is at high twist values (> 40°). See Figure 3 and Suppl. Figures S6-S8 for a detailed analysis.

While inter-basepair, or junction, helical coordinates are frequently far from having a normal distribution, the first and second moments of their equilibrium distributions are still well defined, and can be approximated by evaluating the appropriate averages along our MD simulation time series, and over all instances of dinucleotide (or NN, nearest neighbour) or tetranucleotide (NNN, next nearest neighbour) contexts. Only a few general NN rules can be observed for the first moments (or averages): i) As part of Calladine's principles, we also observed that roll (YR) > roll (RY), while the inverse situation is true for twist: twist (RY) > twist (YR); ii) YR inter-basepair steps typically have higher than normal slide and roll; iii) RY steps typically have lower than normal slide and roll; and iv) YY and RR steps have lower than normal tilt values. Any further rules concerning the average values of helical inter-basepair coordinates need to be formulated as the averages for the central junction or step in a specific tetranucleotide sequence context due to strong nonlocal sequence dependence, at least in part due to tetranucleotide dependent anharmonic effects (Figure 3 and discussion below).

Backbone polymorphism. Flexibility of DNA backbones is linked to rotations around seven torsion angles (α, β ,γ, δ, ε, ζ and χ, with δ in the present analysis

being replaced by the sugar phase angle P), which in some cases move in a concerted way (for example α/γ and ε/ζ), leading to conformational sub-states. The best studied of the coupled transitions is the so-called BI/BII transition, which occurs due to the concerted rotation of the ε/ζ torsions. BI→BII transitions are believed to be functionally relevant. They occur in some high-resolution crystal structures(*51*, *52*) and are also detected in ^{31}P NMR spectra(*53*, *54*). Results in Suppl. Figure S9 show that the BII state is much more frequent than expected from simulations performed using previous force fields, matching NMR estimates for equivalent sequences(*55*). Very interestingly (see Figure 4, and Suppl. Tables S4 and S5), the BI/BII equilibrium is strongly dependent on the surrounding base sequence. For example, RR backbones exhibit quite high BII percentages, especially in the presence of Y at the 5' end of the corresponding tetranucleotide, while the YY backbones are typically biased towards the BI state, generating a strong asymmetry at RR:YY steps. While the general trends of BI/BII equilibria are robust with respect to changes in salt, a detailed analysis indicates the existence of subtle differences(*5*), which are especially visible for RR and YR inter-basepairs: in general, Na^+ increases the total percentage of the BII state (Figure 4), but reduces its sequence-dependence, in perfect agreement with experimental data(*56*). As previously reported(*4*, *5*), we found a very strong correlation between BI→BII transitions and the formation of unconventional hydrogen bonds of the type CH---O, which are instrumental in mechanically coupling the movements detected in the backbone with those seen in the bases (see Figure 4, Suppl. Table S6).

In contrast to BI/BII dynamics, the α/γ conformational landscape is dominated by the canonical conformation, which, on average, represents around 90% of the collected ensembles. Non-canonical conformers are more likely to appear in Na^+ simulations than with K^+ (Suppl. Tables S7 and S8). Transitions to non-canonical α/γ conformations are frequent, but the alternative states tend to have a short life time (on average we measured ~500 transitions per μs per nucleotide, with an average residence time ~5 ps). These brief transitions have little impact on the global conformational ensemble(*9*). No clear sequence-related rules can be determined for α/γ transitions, but, as expected, C and G nucleotides show longer-lived and more frequent α/γ transitions than A or T(*8*, *9*, *57*). Phase (P) angle analysis (Suppl. Figure S10) show South (C2'-endo, ~150°) conformations are dominant as expected, but East conformers are common, and sampling North states is not rare, especially for pyrimidines(*9*). As also expected, glycosidic torsions (χ) are always in the *anti* region (-180 to -90°), with purines sampling more frequently than pyrimidines the *high-anti* conformations (-90 to -30°; see Suppl. Figure S11). Finally, all nucleotides exhibit the same wide distribution for the β angle, spanning from 120° to 240°, with a strongly marked peak at the canonical value (180°) and a marginal population at ~70° (gauche+, see Suppl. Fig. S12), in good agreement with results from the data mining of X-ray structures(*58*).

The choreography of correlated motions in the DNA. The movements of the DNA duplex often involves concerted changes in conformational degrees of freedom, generating a complex choreography. As an example, puckering (measured by the phase angle P) and glycosidic torsions (measured by the χ angle) are tightly coupled, and the population of East and North puckering leads to a marked displacement of χ to lower values (Suppl. Figure S13). Furthermore, χ and P torsions are coupled to the ε/ζ changes in a sequence-dependent manner (Figure S14). Thus, in purines the population of the BII state is coupled to a displacement of puckering to the East (P) and (χ) high-anti regions, while in pyrimidines the population of BII conformers leads only to a slight displacement to the high-anti region, without significant puckering changes.

When the conformational analysis is carried out at the intra-basepair level, a pattern of sequence-dependent correlated movements emerges. All distinct trinucleotides show moderate-to-high correlations in shear-opening, shear-stretch, and stagger-buckle. The pattern of correlation is less clear for the remaining intra-basepair parameters, although several trinucleotides show stretch-opening correlations (Suppl. Fig. S15). A more complex sequence-dependent picture of correlated movements can be obtained by analyzing the inter-basepair step helical parameters (Suppl. Fig. S16). For example, mild to strong correlations are found in shift-tilt, slide-twist, rise-tilt, shift-slide, and shift-twist movements for RR steps. For RY, weaker correlations can be found (depending on the tetranucleotide sequence-environment) in shift-tilt, slide-rise and roll-twist. Finally, YR inter-basepairs may exhibit moderate to strong correlations for shift-tilt, slide-twist, rise-twist and roll-twist (Suppl. Fig. S16). Interestingly, for all the tetranucleotides, shift-slide and roll-twist always show negative correlations, while shift-tilt and slide-twist always show positive correlations. As expected, correlations also emerge when combining inter and intra helical parameters in the same analysis. Thus, a significant number of tetranucleotides show moderate to strong correlations of opening with shift, buckle with rise, and stagger with tilt (data not shown). It is also worth noting that the network of correlations extends to neighbouring steps. As an example, twist in the central YR step of XYRR tetranucleotides is highly correlated with slide in the adjacent RR step(*5*, *10*), which again stresses the limitations of simple nearest neighbours interpretations of DNA conformational mechanics, and points the way to coarse grain models such as cgDNA cites, that encompass longer range coupling, with associated longer range sequence-dependence of the observed means and many non-vanishing covariances.

Lastly, backbone and base pair conformations are connected in a complex way, with ε/ζ (BI/BII) being the major determinant in the polymorphism. Very often, tetranucleotides showing simultaneous sampling of BI and BII conformations are those with bimodality in some helical parameter at the same step (70% of the

bimodal inter-basepair helical parameters occur in steps with bimodal BI/BII distributions, see Figure 3 and Suppl. Table S4 and S5). The BI/BII state also correlates with inter-basepair helical coordinates in neighbouring junctions, explaining part of the geometrical constraints postulated by Calladine. For example, the increase in the percentage of BII at the central junction of a given tetranucleotide correlates with larger shift values for all sequences (Suppl. Figures S17), and is also coupled to lower twist and slide values. The BI/BII ratio at a junction *i* also correlates with shift, twist and slide values at step *i+1* and *i-1* (Suppl. Figures S18 and S19), highlighting the subtle mechanical coupling between backbone and base pair step conformations within DNA(*58*).

All the observations made above can be unified in a global flexibility scheme for B-DNA (Figure 5), showing that all base pair junctions contain potentially polymorphic elements (BI/BII, shift, slide, or twist) that can lead to bimodal behavior depending on the specific tetranucleotide environment. The analysis we have carried out leads to a scheme with strong predictive power at the tetranucleotide level. As a single example, we can now say with confidence that when the choice of X and Y within an XYRY tetranucleotide leads to bimodality, this will be expressed in shift and twist, coupled with a low-to-moderate percentage of BII in the Watson strand. In contrast, when XRRX tetranucleotides are considered, bimodality will show up in either shift, slide or twist, coupled with a moderate-to-high percentage of BII in the Watson strand of the central junction.

The experimental validation of this new and extended set of rules to predict B-DNA conformation based on the sequence is however very difficult to achieve. To date, only the sequence known as DDD (Drew-Dickerson Dodecamer) was determined experimentally enough times, using significantly different techniques, protocols, and laboratory conditions, to have a view, yet limited, of the structural fluctuations of a given isolated B-DNA(*59*). Moreover, it's the only sequence for which two independent ^{31}P-NMR experiments were performed, and from which accurate BI% were obtained(*54*, *60*). Consequently, we took 93 structures from the PDB(*59*), plus the two ^{31}P-NMR experiments to compare the results of our predictive rules on the DDD sequence based on the two PARMBSC1 MD dataset computed in this work (Suppl. Fig. S20). Helical conformations predicted in terms of the inter-basepair parameters were in excellent agreement with the experimental ensemble reproducing sequence-dependent features of this prototypical B-DNA sequence. Moreover, we correctly predicted base polymorphisms in shift and twist(*5*, *10*, *18*), and most importantly the backbone substates, in particular BI/BII (Suppl. Fig. S20).

CONCLUSIONS

The analysis of numerous molecular dynamics trajectories obtained with an accurate, last generation, force field has allowed us to derive some general rules concerning the equilibrium conformation distribution of B-DNA, which represent a significant step beyond Calladine-Dickerson earlier propositions. Specifically, we are now able to predict when significantly anharmonic distributions will arise as a function of tetranucleotide sequence context:

- The first and second moments (average and covariance) of the equilibrium distributions of helical coordinates for DNA can only be understood in terms of nonlocal sequence-dependence contexts, to at least the trinucleotide level for intra-basepair coordinates, and the tetranucleotide level for inter-basepair coordinates.
- A harmonic model of DNA dynamics will not be able to accurately predict third and higher moments of the equilibrium distribution because significant anharmonic movements arise frequently. In fact, the distribution of many inter-basepair coordinates is significantly binormal and, in a non-negligible number of cases, actually bimodal (*i.e.* multi-peaked). Such bimodality, and the relative population of corresponding local minima of the free energy, is dependent on the tetranucleotide context. Slide for GG, twist for CG and AG, and shift for YR are the most common steps and helical coordinates exhibiting bimodality, with the tetranucleotides most commonly enhancing bimodality being AGGA, GGGA, ACGG, GCGA, GCGG, and GTAA.
- Backbone torsional changes are coordinated in pairs (α/γ, P/χ and ε/ζ). Movements in α/γ lead to the generation of short-lived non-canonical states, which can however be populated in the presence of ligands, as it was previously observed for protein-DNA complexes(*7*). Changes in sugar puckering to the East region leads to lower χ values, while coordinated changes in the ε/ζ pair lead to the BI/BII polymorphism with coupled impacts on helical parameters. Both ε/ζ and P/χ couplings exhibit sequence dependence.
- The BI/BII conformational change is coupled to the cationic atmosphere surrounding DNA, and to the formation of non-canonical CH---O hydrogen bonds. BI/BII transitions are especially prevalent for YRRX sequences and often are associated to bimodality in helical coordinate distributions at the inter-basepair level. They are a major source of polymorphism in B-DNA. In general, the population of the BII state is coupled to large shift, and low slide and twist at the same junction, but distant and more complex correlations exist between BI/BII conformational states and the helical conformation of neighbouring steps.
- Helical parameters at a given step are not independent, but show a complex backbone-mediated pattern of dependencies. For example, shift-tilt and roll-twist always show negative correlations, and the opposite applies to

shift-tilt and slide-twist coupling. On the contrary, correlations between slide-twist, shift-slide and shift-twist vary as a function of base sequence. Moreover, helical coordinate correlations may extend to neighbouring base pairs as a function of the local sequence.

- Calladine's principles, and Dickerson's algorithms for twist/roll/δ/propeller, can now be transformed into quantitative predictions for all the structural features (helical conformations and backbone substates) of canonical DNA sequences. These extended rules have been implemented on a web server that predicts the average conformation of any B-DNA sequence, in terms of the average helical parameters, base and backbone polymorphisms, and P/χ conformations (see https://mmb.irbbarcelona.org/miniABC/).
- Furthermore, using the predictive cgDNA coarse-grained model (and its dinucleotide dependent parameter sets fit to MD simulations), the nonlocal sequence-dependent first (average) and second (covariance) helical coordinate moments can be computed interactively for an arbitrary sequence on the cgDNAweb(*61*) server http://cgdnaweb.epfl.ch/, including interactive visualisation of the expected or ground state conformation.
- Additionally, the local and global flexibility of arbitrary canonical B-DNA sequences can be obtained by using the rigid inter-basepair MCDNA coarse grain model, to provide the user a limited number of alternative equilibrium B-DNA conformations according to the tetranucleotide states of the underlying sequence. (https://mmb.irbbarcelona.org/MCDNAlite/). Using the extended Calladine-Dickerson rules derived herein, the backbone, sugar, and base conformational substates are predicted and rebuilt at atomic resolution, based on the spontaneous values of inter-basepair helical parameters.

ACKNOWLEDGMENTS

P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher. A.B. and J.W. are La Caixa PhD fellows (UB and IRB Barcelona, Spain). M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avançats) researcher. The authors thank the Ascona B-DNA Consortium for the trajectories in the standard set (μABC). The authors are also grateful with Genís Bayarri for setting up the web page: https://mmb.irbbarcelona.org/miniABC/. Calculations were performed in the Laboratory for Computation and Visualization of Mathematics and Mechanics at the EPFL, Lausanne, and at IRB Barcelona.

FUNDING

This work has been supported by: 1) the Spanish Ministry of Science (BFU2014-61670-EXP, BFU2014-52864-R), the Catalan SGR, the Instituto Nacional de Bioinformática, the European Research Council (ERC SimDNA), the European Union's Horizon 2020 research and innovation program under grant agreement N°676556, the European Union's Horizon 2020 research, BioExcel and MuG EU-projects and the Biomolecular & Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (to M.O.). 2) The MINECO Severo Ochoa Award of Excellence, Government of Spain (awarded to IRB Barcelona). 3) CNRS and the ANR project CHROME (ANR-12-BSV5-0017-01) (to R. L.). 4) Swiss National Science Foundation [200020_163324] (to J.H.M.).

AUTHOR CONTRIBUTIONS

The miniABC sequence library was designed by M.P. and R.L. Simulations were performed by A.S.P. with the assistance of D.P., A.H., and P.D.D. All co-authors were involved in producing results and further discussions. P.D.D. integrated all the results, and was the scientific coordinator for the project. P.D.D., J.H.M., and M.O. discussed the analysis and wrote the manuscript with contributions from all the co-authors. The original idea of the project came from R.L., J.H.M., and M.O.

REFERENCES

1. S. Neidle, *Principles of nucleic acid structure* (Elsevier, 2008).
2. W. Fuller, T. Forsyth, A. Mahendrasingam, Water-DNA interactions as studied by X-ray and neutron fibre diffraction. *Philos. Trans. R. Soc. B Biol. Sci.* **359**, 1237–1248 (2004).

3. B. Hartmann, D. Piazzola, R. Lavery, BI-BII transitions in B-DNA. *Nucleic Acids Res.* **21**, 561–8 (1993).
4. A. Balaceanu *et al.*, The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.* **8**, 21–28 (2017).
5. P. D. Dans *et al.*, Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.* **42**, 11304–11320 (2014).
6. M. Zgarbová *et al.*, Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Inf. Model.* **57**, 275–287 (2017).
7. P. Várnai, D. Djuranovic, R. Lavery, B. Hartmann, Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res.* **30**, 5398–406 (2002).
8. A. Pérez, F. J. Luque, M. Orozco, Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.* **129**, 14739–14745 (2007).
9. P. D. Dans *et al.*, Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res.* **44**, 4052–4066 (2016).
10. M. Pasi *et al.*, μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **42**, 12272–12283 (2014).
11. R. Rohs *et al.*, Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.* **79**, 233–69 (2010).
12. D. D. Boehr, R. Nussinov, P. E. Wright, The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
13. C. R. Calladine, Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.* **161**, 343–52 (1982).
14. R. E. Dickerson, A. Klug, Base sequence and helix structure variation in B and A DNA. *J. Mol. Biol.* **166**, 419–441 (1983).
15. A. V Fratini, M. L. Kopka, H. R. Drew, R. E. Dickerson, Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *J. Biol. Chem.* **257**, 14686–707 (1982).
16. T. E. Cheatham, B. R. Brooks, P. A. Kollman, III, *Curr. Protoc. nucleic acid Chem.*, in press, doi:10.1002/0471142700.nc0705s06.
17. W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, V. B. Zhurkin, DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 11163–8 (1998).
18. P. D. Dans, A. Pérez, I. Faustino, R. Lavery, M. Orozco, Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.* **40**, 10668–10678 (2012).
19. A. Pérez, F. J. Luque, M. Orozco, Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.* **45**, 196–205 (2012).
20. P. D. Dans, J. Walther, H. Gómez, M. Orozco, Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.* **37**, 29–45 (2016).
21. D. L. Beveridge *et al.*, Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. I. Research Design and Results on d(CpG) Steps. *Biophys. J.* **87**, 3799–3813 (2004).
22. S. B. Dixit *et al.*, Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps. *Biophys. J.* **89**, 3721–3740 (2005).
23. R. Lavery *et al.*, 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**, 299–313 (2010).
24. I. Ivani *et al.*, Parmbsc1: A refined force field for DNA simulations. *Nat. Methods.* **13**, 55–58 (2015).
25. P. D. Dans *et al.*, How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.* **45**, 4217–4230 (2017).
26. A. Pérez *et al.*, Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92**, 3817–29 (2007).
27. J. Mintseris, M. B. Eisen, Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC Bioinformatics.* **7**, 429 (2006).
28. P. Medvedev, K. Georgiou, G. Myers, M. Brudno, in *Algorithms in Bioinformatics* (Springer Berlin Heidelberg, Berlin, Heidelberg; http://link.springer.com/10.1007/978-3-540-74126-8_27), pp. 289–301.
29. Y. Orenstein, R. Shamir, Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers. *Bioinformatics.* **29**, i71-9 (2013).
30. A. Pérez *et al.*, Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92**, 3817–29 (2007).
31. D. Case *et al.*, AMBER14 (2014), (available at ambermd.org).
32. R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, R. C. Walker, Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
33. A. Hospital *et al.*, BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.* **44**, D272–D278 (2016).
34. D. R. Roe, T. E. Cheatham, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
35. A. Hospital *et al.*, NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.* **41**, W47–W55 (2013).
36. R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, K. Zakrzewska, Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **37**, 5917–5929 (2009).
37. G. Schwarz, Estimating the Dimension of a Model. *Ann. Stat.* **6**, 461–464 (1978).
38. R. E. Kass, A. E. Raftery, Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
39. M. F. Schilling, A. E. Watkins, W. Watkins, Is Human Height Bimodal? *Am. Stat.* **56**, 223–229 (2002).
40. F. de Helguero, Sui Massimi Delle Curve Dimorfiche. *Biometrika*. **3**, 84 (1904).
41. D. Petkevičiūtė, M. Pasi, O. Gonzalez, J. H. Maddocks, cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.* **42**, e153 (2014).
42. O. Gonzalez, M. Pasi, D. Petkevičiūtė, J. Glowacki, J. H. Maddocks, Absolute versus Relative Entropy Parameter Estimation in a Coarse-Grain Model of DNA. *Multiscale Model. Simul.* **15**, 1073–1107 (2017).
43. O. Gonzalez, D. Petkevičiūtė, J. H. Maddocks, A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.* **138**, 055102 (2013).
44. J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, J. H. Maddocks,

Sequence-Dependent Persistence Lengths of DNA. *J. Chem. Theory Comput.* **13**, 1539–1555 (2017).
45. † Daniele Andreatta *et al.*, Ultrafast Dynamics in DNA: "Fraying" at the End of the Helix (2006), doi:10.1021/JA0582105.
46. U. D. P. and, J. . Alexander D. MacKerell, NMR Imino Proton Exchange Experiments on Duplex DNA Primarily Monitor the Opening of Purine Bases (2005), doi:10.1021/JA056445A.
47. M. Pasi, R. Lavery, Structure and dynamics of DNA loops on nucleosomes studied with atomistic, microsecond-scale molecular dynamics. *Nucleic Acids Res.* **44**, 5450–5456 (2016).
48. G. Altan-Bonnet, A. Libchaber, O. Krichevsky, Bubble Dynamics in Double-Stranded DNA. *Phys. Rev. Lett.* **90**, 138101 (2003).
49. A. Zeida, M. R. MacHado, P. D. Dans, S. Pantano, Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **86**, 1–7 (2012).
50. T. Maehigashi *et al.*, B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.* **40**, 3714–3722 (2012).
51. D. Djuranovic, B. Hartmann, Conformational Characteristics and Correlations in Crystal Structures of Nucleic Acid Oligonucleotides: Evidence for Sub-states. *J. Biomol. Struct. Dyn.* **20**, 771–788 (2003).
52. A. Madhumalar, M. Bansal, Sequence Preference for BI/BII Conformations in DNA: MD and Crystal Structure Data Analysis. *J. Biomol. Struct. Dyn.* **23**, 13–27 (2005).
53. B. Heddi, N. Foloppe, N. Bouchemal, E. Hantz, B. Hartmann, Quantification of DNA BI/BII Backbone States in Solution. Implications for DNA Overall Structure and Recognition. *J. Am. Chem. Soc.* **128**, 9170–9177 (2006).
54. Y. Tian *et al.*, ^{31}P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J. Phys. Chem. B.* **113**, 2596–2603 (2009).
55. A. Ben Imeddourene *et al.*, Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLOS Comput. Biol.* **11**, e1004631 (2015).
56. B. Heddi, N. Foloppe, C. Oguey, B. Hartmann, Importance of Accurate DNA Structures in Solution: The Jun–Fos Model. *J. Mol. Biol.* **382**, 956–970 (2008).
57. T. Dršata *et al.*, Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.* **9**, 707–721 (2013).
58. D. Svozil, J. Kalina, M. Omelka, B. Schneider, DNA conformations and their sequence preferences. *Nucleic Acids Res.* **36**, 3690–3706 (2008).
59. P. D. Dans *et al.*, Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.* **44**, 4052–4066 (2016).
60. C. D. Schwieters, G. M. Clore, A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. *Biochemistry.* **46**, 1152–66 (2007).
61. L. De Bruin, J. H. Maddocks, cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.* (2018), doi:10.1093/nar/gky351.

FIGURES

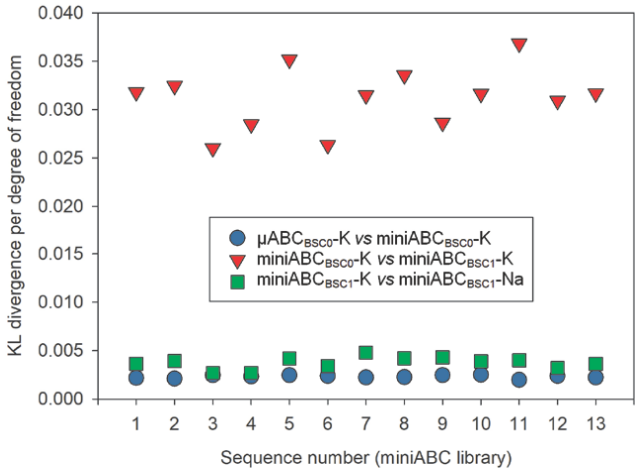


Figure 1. Symmetric Kullback-Leibler divergence per degree of freedom between Gaussian distributions, which is a combined measure of differences in values of first and second moments, for each of the thirteen oligomers in the miniABC training library, but for cgDNA model parameter sets fit to different MD simulation protocols (see Methods and Suppl. Methods).

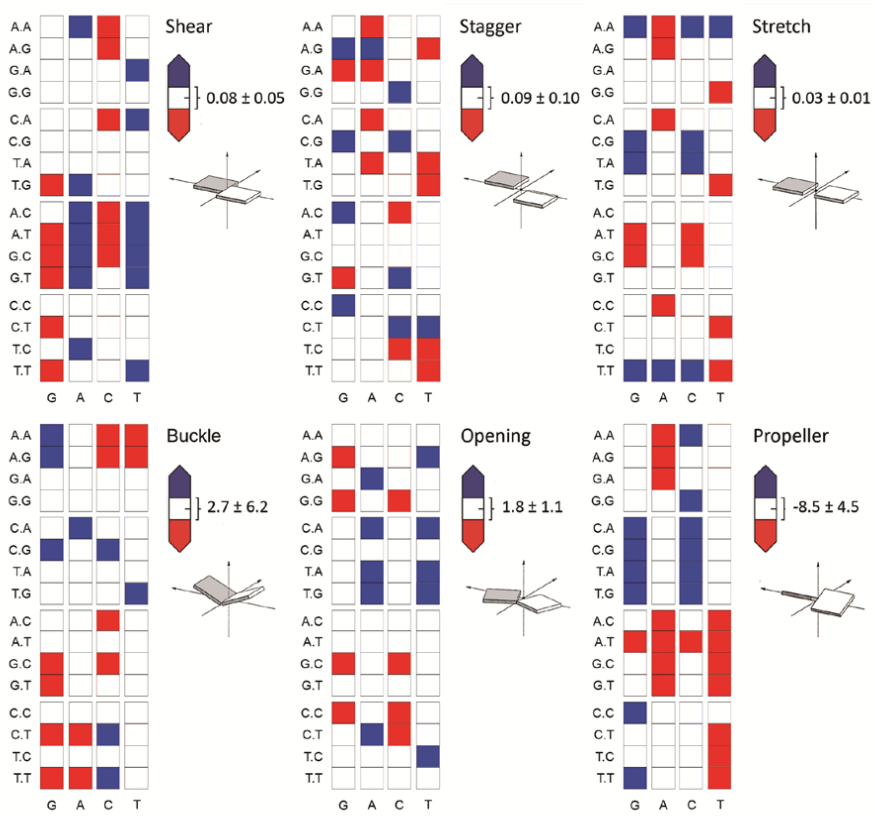


Figure 2. Average values of intra-basepair helical coordinates of the central base-pair (x-axis) in all possible 64 trinucleotide sequence contexts (y-axis). Results obtained from the miniABC$_{BSC1}$-K simulations. The global averages (white) are over all sequence contexts and standard deviations reflect the variation among trinucleotide contexts. The blue squares mean that a specific base-pair has an average value above the global average plus one standard deviation, while red squares mean an average value below the global average plus one standard deviation.

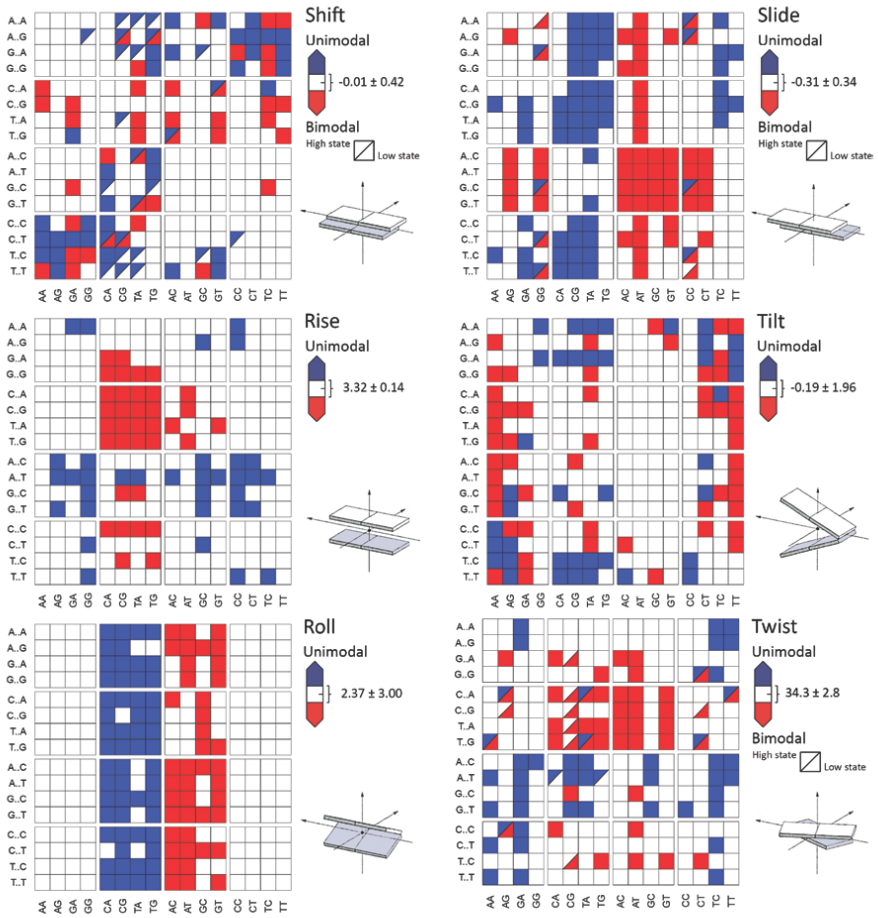


Figure 3. Average values of inter-basepair, or step, helical coordinates for the central junction (x-axis) in all possible 256 tetranucleotide contexts (y-axis). Results obtained from the miniABC$_{BSC1}$-K simulations. Tetranucleotides classified as bimodal (half-square) are polymorphic (*i.e.* they sample two clear conformational substates). The global averages (white), exhibited on the legend at the right of each squared-plot, were computed from the weighted-averages obtained through BIC (see Methods and Suppl. Methods), while the standard deviations reflect the variation along the tetranucleotide sequences that share the same central base pair step. The blue squares mean that a specific step has an average value above the global average plus one standard deviation, while red squares mean an average value below the global average plus one standard deviation.

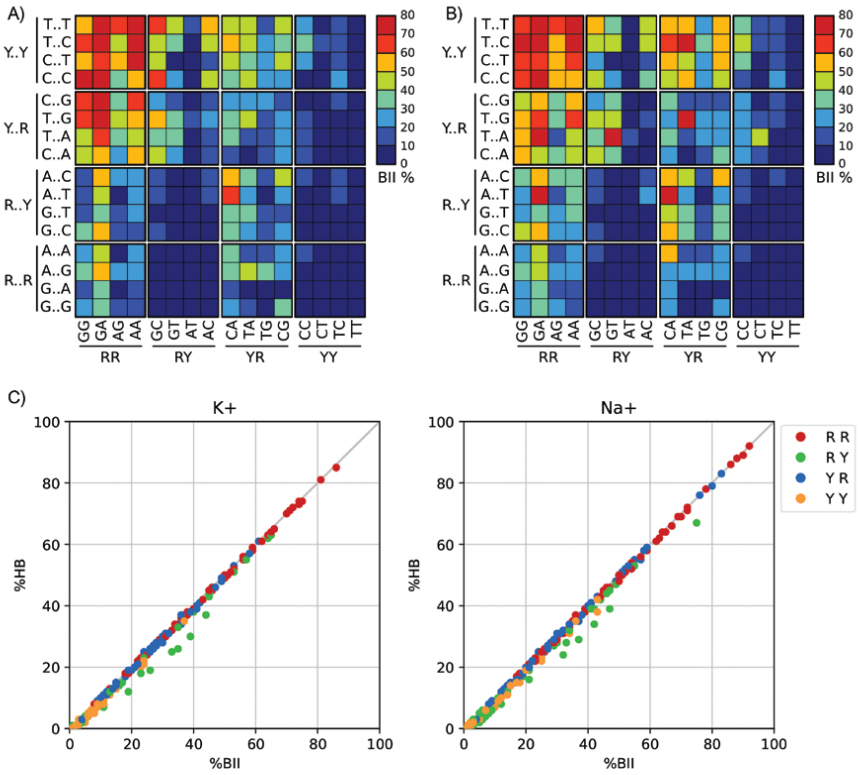


Figure 4. Sequence dependence of BII backbone conformations comparing K^+ and Na^+. A) **mini**ABC_{BSC1}-K BII percentages. B) **mini**ABC_{BSC1}-Na BII percentages. C) Correlation between the percentage of BII (%BII, horizontal axis) and of occurrence of formation of the C–H···O H-bonds (%HB, vertical axis) at the central base step of each of the 256 possible tetranucleotide sequences, colour-coded according to base type of the central base step.

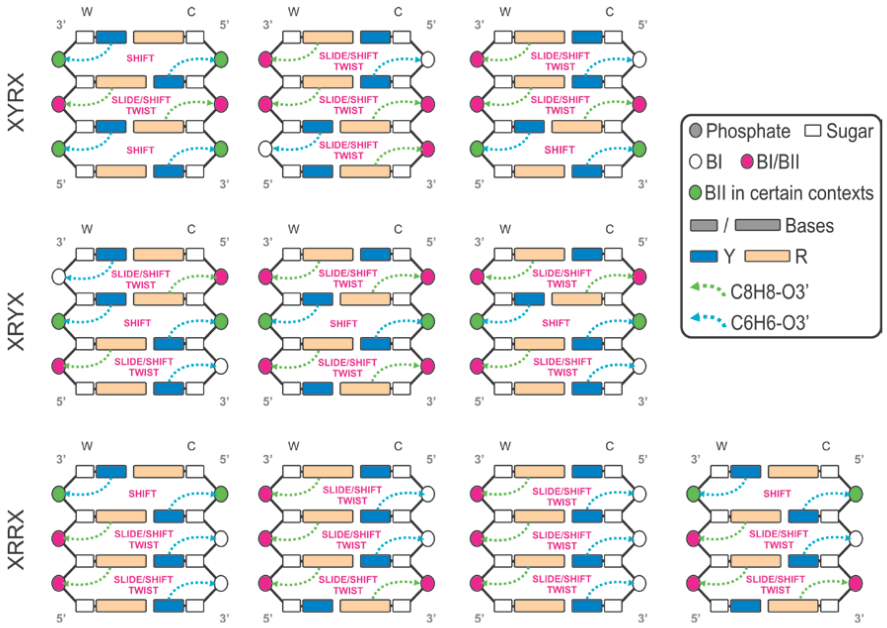


Figure 5. Schema of the polymorphic, or multi-well, landscape exhibited by B-DNA at the tetranucleotide level expressed in the purine (R)/pyrimidine (Y) alphabet, for which only 10 distinct combinations exist, but which still distinguish all possible behaviours. The only helical coordinates that exhibited multi-modality are shift, slide and twist, and each junction in the figure is marked with which coordinates can be multi-modal in it. There is a very high correlation between the occurrence of multi-modality and the formation of a noncanonical hydrogen bond in either the same or a neighbouring junction, along with its associated BI/BII backbone transition (see text).

SUPPLEMENTARY MATERIAL

THE PHYSICAL PROPERTIES OF B-DNA BEYOND CALLADINE'S RULES

Pablo D. Dans[a,b,1], Alexandra Balaceanu[a,b,2], Marco Pasi[c,d,2], Alessandro S. Patelli[e,2], Daiva Petkevičiūtė[e,f,2], Jürgen Walther[a,b,2], Adam Hospital[a,b], Richard Lavery[d], John H. Maddocks[e,1], and Modesto Orozco[a,b,g,1]

[a]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.
[b]Joint BSC-IRB Research Program in Computational Biology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.
[c]LBPA, École normale supérieure Paris-Saclay, 61 Av. du Pdt Wilson, Cachan 94235, France.
[d]Bases Moléculaires et Structurales des Systèmes Infectieux, Univ. Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, Lyon 69367, France.
[e]Institute of Mathematics, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland.
[f]Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Studentų g. 50, 51368 Kaunas, Lithuania.
[g]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain.

[1]To whom correspondence should be addressed:
Dr. Pablo D. Dans, Tel: +34 934039073, Email: pablo.dans@irbbarcelona.org; or
Prof. John H. Maddocks, Tel: +41 216932762, Email: john.maddocks@epfl.ch; or
Prof. Modesto Orozco, Tel: +34 934037155, Email: modesto.orozco@irbbarcelona.org.

[2]These co-authors equally contributed to this work and were alphabetically sorted.

SUPPLEMENTARY METHODS

Simulation details. Canonical duplexes were generated using Arnott B-DNA fiber parameters(*1*), and solvated by a truncated octahedral box of SPC/E(*2*) water molecules with a minimum distance of 10 Å between DNA and the closest face of the box. Systems were neutralized with K^+ or Na^+ ions adding additional 150 mM of K^+Cl^- (or Na^+Cl^-). PARMBSC0(*3*) and PARMBSC1(*4*) force fields were used to describe DNA, while Dang's parameters were used for ions(*5*). Systems were optimized and equilibrated as described elsewhere(*6*), and simulated for 1 μs in the NPT ensemble, using Particle-Mesh Ewald corrections(*7*) and periodic boundary conditions. SHAKE was used to constrain bonds involving hydrogen(*8*), allowing 2 fs integration step.

Typically, analyses presented here correspond to the second part of the trajectory (last 500 ns).

Bayesian Information Criterion (BIC), Bayes Factors, and the Helguerro's theorem. We used the BIC methodology to determine the optimal number of Gaussian functions needed to fit a given distribution. This is done by finding the set of parameters that minimizes the BIC values (the model with the lower BIC is chosen) according to(*9*):

$$-2lnp(x|k) \approx BIC = -2\ln(L) + kln(n)$$

Where *x* are the observed data, *k* is the number of free parameters to be estimated, and *p(x|k)* is the probability of the observed data given the number of parameters, or, in other words, the likelihood of the parameters given the dataset. *L* is the maximized value of the likelihood function for the estimated model, and *n* is the number of data points in *x* (the number of observations). In this work we limit the BIC to considering a maximum of two Gaussians, leading to the classification of each distribution as uninormal (fitted with one Gaussian) or binormal (fitted with a combination of two Gaussians).

The Bayes Factors that can be extracted from the BIC analysis were used to determine the strength of the evidence in favour of the model chosen by BIC(*10*, *11*). This leaded to a third classification labelled as "insufficient evidence", when either of the two models determined with BIC (uninormal or binormal) couldn't be statistically supported.

Finally, when there was sufficient evidence to favour a binormal fitting, we used an extension of the Helguerro's theorem(*12*, *13*) to define the modality of the distribution and distinguish the cases where the two peaks of the fitted Gaussians are close together from those where they are significantly separated. This is the most important distinction in terms of understanding DNA dynamics. In the first case, for practical purposes, the use of a single Gaussian distribution may often be justified to represent the data (the overall distribution may be interpreted as binormal-unimodal), while it cannot be used to estimate higher moments in the second multi-peaked case (binormal-bimodal distributions). For a given parameter, we defined an inter-basepair, or intra-basepair as polymorphic from the structural point of view, when a given distribution was classified using these three approaches as binormal-bimodal.

Correlations between substates. For each tetranucleotide we calculated the correlation between the backbone state at the central step (inter-basepair i) and the helical parameters shift, slide and twist at two consecutive levels around the central

2

dinucleotide (i-1, i, and i+1). The substates of the torsion angles of the backbone were categorized following the standard definition: *gauche positive* (g+) = 60 ± 40 degrees; *trans* (t) = 180 ± 40 degrees; and *gauche negative* (g–) = 300 ± 40 degrees. For the correlations with BI/BII, we assigned to the backbone one of two possible discrete values, either BI or BII, according to the sub-state of the ζ torsion (g- or t respectively) at the central bps junction. All frames where the ζ torsion didn't fall inside the ranges defined by g- and t were not considered in the analysis. This leads to a strong reduction of the noise that comes from specific tetranucleotides, when trying to find patterns by grouping them (*e.g.* the "noise" arising from the individual behavior of the GAGA, GGGG, and AAGA tetranucleotides when considering the RRRR family). The point-biserial(*14*) correlation coefficient, mathematically equivalent to the Pearson correlation(*15*), was used as a measure of the correlation between these discrete substates of the backbone and the continuous values of the inter-basepair helical parameters. The obtained correlation values were divided in five categories: i) ≥ -0.6, strong negative correlation; ii) < -0.6 and ≥ -0.4, mild negative correlation; iii) >-0.4 and< 0.4, no correlation; iv) ≥ 0.4 and < 0.6, mild positive correlation; and finally v) ≥ 0.6, strong positive correlation. We then group each of these categorized correlation matrices according to the 10 non-redundant tetranucleotide combinations of Y/R bases, and for each entry selected the dominant mode to describe the subset (i.e. the most common situation shared by the individual tetranucleotides within a family). In the same way, correlations between sum and differences of helical parameters have been computed, as previously done in Calladine's works(*16*, *17*).

Kullback-Leibler (KL) divergence between configuration distributions. For each MD simulation we fit a Gaussian or multi-variate normal distribution on the helical coordinates by estimating a mean shape vector $\hat{w}$ and a stiffness, or inverse covariance matrix K, from the MD time series. (This Gaussian is in dimension 12N-6 for a fragment with N base pairs, so dimension 210 for the case N=18 considered here.) The KL divergence(*18*) is a convenient way to quantify the difference between two probability distributions. When both distributions are Gaussian with mean vectors $\hat{w}_1$, $\hat{w}_2$ and inverse covariance matrices $K_1$and K_2, then the divergence can be explicitly evaluated as:

$$D_{12} = \frac{1}{2}\left[K_1^{-1} : K_2 - \ln\left(\frac{\det K_2}{\det K_1}\right) - I : I\right] + \frac{1}{2}(\hat{w}_1 - \hat{w}_2) \cdot K_2(\hat{w}_1 - \hat{w}_2),$$

Where a colon denotes the standard Euclidean inner product for square matrices and I denotes the identity matrix of the same dimension as K_1 and K_2. The second term of this expression is interesting to look at separately: it quantifies the difference in expected

shapes, weighted by one of the inverse covariance, and is equal to the square of the Mahalanobis distance:

$$M_{12} = \frac{1}{2}(\hat{w}_1 - \hat{w}_2) \cdot K_2(\hat{w}_1 - \hat{w}_2),$$

Both KL divergence and Mahalanobis distance are non-symmetric, but here we chose to report the symmetrized values: $D = \frac{1}{2}(D_{12} + D_{21})$ and $M = \frac{1}{2}(M_{12} + M_{21})$. To give a meaning to values of the KL divergence, the KL values were scaled by 12N-6 (being N the number of base-pairs in each oligomer), obtaining in this way a divergence per degree of freedom.

cgDNA calculation of DNA Persistence Length. The cgDNAmc code(*19*) allows efficient generation of ensembles of configurations over ensembles of sequences, so that the possible range of values of various expectations can be examined as the sequence of the DNA duplex varies. One standard set of expectations to compute is tangent-tangent correlations along the duplex in order to determine the associated decay rate or persistence length ℓ_p along a given fragment. The persistence length ℓ_p is often taken as an overall proxy for the stiffness of the duplex, with longer persistence length indicating greater stiffness. However it is known (see *e.g.* the discussion in ref 19) that the value of ℓ_p depends on both the stiffness of the duplex and on its intrinsic curvature, with bent sequences having lower persistence lengths. For this reason ℓ_p is sometimes called apparent persistence length. A sequence-dependent dynamic persistence length ℓ_d was introduced(*19*), which largely eliminates dependence on intrinsic curvature. Thus ℓ_d is a better proxy for an overall stiffness, while the difference (ℓ_d - ℓ_p) is an overall measure of how intrinsically bent the duplex is. Fig S2A provides spectra (or histograms) of possible values of both ℓ_p and ℓ_d for 10K sequences according to a cgDNA model parameter set fit to MD simulations of the miniABC library using the PARMBSC0 MD potentials. The range of variation in ℓ_d is small compared to that of ℓ_p, and it can be verified that all exceptionally low values of ℓ_p correspond to highly bent sequences. The same data for the same 10K sequences, but for a cgDNA model parameter set fit to MD simulations of the miniABC library using the PARMBSC1 MD potentials is shown in Fig S2B. The fact that the spectra of dynamic persistence lengths ℓ_d shifts to the right indicates that the PARMBSC1 potentials lead to duplexes that are slightly stiffer than for PARMBSC0, while the fact that the spectra of apparent persistence lengths has a smaller tail on the left indicates that PARMBSC1 leads to duplexes that have smaller intrinsic bends than for PARMBSC0. Figure S2 also provide the values of apparent and dynamic persistence lengths for the six independent dinucleotide tandem repeats poly(XZ). As such sequences are very straight, their apparent and dynamic persistence lengths are all very close. And for both the PARMBSC0 and PARMBSC1 parameter sets

the sequence poly(AA) is the high outlier among all sequences, with poly(AT) being by far the low outlier for ℓ_d among all sequences.

Statistics, graphics and molecular plots. The statistical analysis, including the Bayesian Information Criterion (BIC), Bayes Factor analysis, Helguerro's theorem, Kullback-Lieber divergence, and correlations, as well as associated graphics, were obtained with R 3.0.1 statistical package(*20*), the MatLab R2016b package, numpy(*21*) and matplotlib(*22*). The molecular plots were generated using VMD 1.9(*23*).

5

SUPPLEMENTARY TABLES

Table S1. DNA sequences in the miniABC library.

Seq. number	Watson strand (5'-3' direction)
1	GCAACGTGCTATGGAAGC
2	GCAATAAGTACCAGGAGC
3	GCAGAAACAGCTCTGCGC
4	GCAGGCGCAAGACTGAGC
5	GCATTGGGGACACTACGC
6	GCGAACTCAAAGGTTGGC
7	GCGACCGAATGTAATTGC
8	GCGGAGGGCCGGGTGGGC
9	GCGTTAGATTAAAATTGC
10	GCTACGCGGATCGAGAGC
11	GCTGATATACGATGCAGC
12	GCTGGCATGAAGCGACGC
13	GCTTGTGACGGCTAGGGC

Table S2. Sequence-averaged conformational parameters obtained from the different miniABC simulations.[a]

	miniABC$_{BSC0}$-K		miniABC$_{BSC1}$-K		miniABC$_{BSC1}$-Na	
Parameter	Average	SD	Average	SD	Average	SD
Shear (Å)	0.02	0.30	0.02	0.30	0.02	0.30
Stretch (Å)	0.03	0.12	0.03	0.12	0.03	0.11
Stagger (Å)	0.06	0.40	0.10	0.38	0.10	0.38
Buckle (°)	0.8	10.8	1.5	9.9	1.6	9.7
Propeller (°)	-12.0	8.2	-9.0	8.1	-9.3	8.2
Opening (°)	2.2	4.5	1.8	4.3	1.8	4.2
Xdisp (Å)	-1.77	1.52	-0.88	1.36	-0.64	1.43
Ydisp (Å)	0.03	1.27	0.00	1.13	-0.01	1.17
Inclination (°)	8.2	7.1	4.0	6.6	2.8	7.0
Tip (°)	0.2	6.7	0.3	6.3	0.3	6.4
Shift (Å)	-0.03	0.69	-0.03	0.80	-0.04	0.83
Slide (Å)	-0.51	0.62	-0.29	0.55	-0.22	0.55
Rise (Å)	3.32	0.32	3.32	0.30	3.32	0.29
Tilt (°)	-0.3	4.3	-0.3	4.4	-0.3	4.5
Roll (°)	4.5	5.8	2.4	5.7	1.7	5.8
Twist (°)	32.1	5.6	34.4	5.5	34.7	5.3
α (°)	-71.1	13.9	-72.1	15.4	-72.3	15.4
β (°)	170.3	13.8	167.8	21.0	166.9	21.2
γ (°)	56.3	12.3	55.0	18.9	55.0	19.1
δ (°)	119.4	21.3	135.3	15.5	136.2	14.7
ε (°)	-167.4	25.4	-160.4	25.8	-158.6	27.1
ζ (°)	-94.1	33.5	-111.4	41.6	-113.8	43.8
χ (°)	-120.5	20.2	-112.1	17.0	-111.2	16.9
Phase (°)	128.3	37.6	151.4	26.5	152.3	25.0
Amplitude (°)	38.4	7.0	41.6	6.6	41.8	6.6

[a] Capping base pairs were removed from the analysis. For the dihedral angles only the Watson strand was considered.

Table S3. DNA breathing and fraying. Base opening statistics based on the analysis of the WC hydrogen bonds.

	Loss of one Hbond [a]		Loss of two Hbonds		Loss of three Hbonds		Solvent exchange[b]	
	Occ.[c] (%)	<t½>[d] (ns)	Occ. (%)	<t½> (ns)	Occ. (%)	<t½> (ns)	Occ. (%)	<t½> (ns)
				K+Cl-				
C:G bp terminal	3.73	0.099	2.55	0.754	1.73	1.332	2.14	3.436
C:G bp terminal(-1)[e]	0.33	0.327	0.01	15.53	<0.01	---	<0.01	---
C:G bp central	0.45	0.251	0.03	10.47	0.01	315.2	0.01	149.5
A:T bp central	1.67	0.089	0.06	7.700	---	---	0.03	41.54
				Na+Cl-				
C:G bp terminal	2.81	0.095	1.57	0.761	0.87	2.209	1.20	3.552
C:G bp terminal(-1)	0.38	0.288	0.01	14.39	<0.01	---	<0.01	.---
C:G bp central	0.52	0.222	0.03	8.651	<0.01	---	<0.01	---
A:T bp central	1.59	0.094	0.04	8.963	---	---	0.01	62.49

[a] We consider a hydrogen bond broken when the distance between the heavy atoms involved in the Watson-Crick interactions was greater than 3.5 Å. [b] Solvent exchange refers to base openings where at least one donor-acceptor distance of WC hbonds is larger than 6 Å. These large separations allow water molecules to interact directly with the base, and eventually exchange protons with imino groups of the bases. [c] Occ. stands for occurrence in %. [d] Average open base lifetime. [e] Refers to the C:G base-pair prior to last (residue numbers 2:35 and 17:20), see Table S1.

Table S4. BII percentages for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-K.

	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	52	74	64	74	65	44	11	57	49	49	19	47	24	9	14	1
T..C	66	86	40	81	45	39	6	40	58	41	22	39	37	11	15	2
C..T	56	62	56	70	45	6	2	13	42	30	24	45	23	2	10	3
C..C	72	86	37	53	64	23	5	40	36	41	22	24	9	5	24	1
C..G	62	71	36	64	23	19	4	13	24	26	27	18	8	2	11	1
T..G	65	75	47	50	53	33	11	24	30	49	15	26	14	8	7	2
T..A	45	66	31	43	35	35	6	13	32	14	14	28	15	6	7	1
C..A	40	59	25	50	49	26	5	11	18	9	12	20	14	5	5	0
A..C	19	51	24	29	16	5	1	15	53	30	13	46	13	1	11	1
A..T	12	46	8	23	15	5	1	17	61	28	9	24	8	1	10	1
G..T	13	38	13	23	8	2	0	5	31	36	12	15	3	1	9	1
G..C	34	56	11	19	13	4	1	3	39	28	14	21	9	1	6	1
A..A	23	46	8	23	9	2	1	6	36	12	18	22	10	1	8	0
A..G	33	59	21	26	9	2	1	5	30	41	30	27	8	1	7	1
G..A	14	37	8	13	6	2	0	4	10	9	4	7	4	0	3	0
G..G	22	38	15	18	6	4	1	1	27	11	7	31	6	1	3	1

9

Table S5. BII percentages for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-Na.

	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	67	78	67	92	55	33	7	41	58	53	22	51	36	18	15	2
T..C	72	88	52	90	46	42	9	42	67	80	34	59	43	14	17	3
C..T	62	64	59	69	29	7	1	10	52	43	28	51	34	7	14	3
C..C	65	86	54	54	49	21	6	30	49	57	27	41	14	8	25	2
C..G	45	51	34	59	21	37	5	7	38	18	19	19	20	5	16	2
T..G	54	65	34	63	47	47	8	17	29	76	21	22	21	11	12	3
T..A	50	70	15	45	34	75	12	3	37	17	21	27	20	43	7	2
C..A	50	49	35	47	44	32	6	9	20	5	14	36	25	10	7	1
A..C	31	54	39	39	10	4	1	19	55	40	14	52	19	0	12	1
A..T	23	72	14	30	14	4	1	21	83	23	5	24	10	1	6	1
G..T	25	36	23	36	6	3	0	5	41	30	15	30	5	1	7	1
G..C	44	57	26	23	14	4	1	4	50	31	12	32	9	1	4	1
A..A	26	49	18	26	11	4	1	7	50	13	16	29	11	1	8	0
A..G	32	46	21	28	8	2	1	6	22	29	25	21	9	1	8	0
G..A	21	34	17	16	5	4	1	5	20	2	4	8	5	0	2	0
G..G	29	30	21	15	6	7	1	2	27	9	8	27	8	2	4	1

10

Table S6. Pearson correlation coefficients between BII% and the formation of the C-H···O H-bond.

	BII% vs C8-H8···O3'		BII% vs C6-H6···O3'		
Set	RR	YR	RY	YY	Total
miniABC$_{BSC1}$-K	1.000	0.999	0.994	0.996	0.998
miniABC$_{BSC1}$-Na	1.000	0.999	0.995	0.997	0.998

11

Table S7. Percentages of α/γ torsions in the canonical sub-state (characterized by α in g- and γ in g+) for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-K.

Flanks	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	97	97	90	98	98	99	99	96	90	93	98	97	96	98	94	96
T..C	97	95	98	95	91	87	97	97	97	98	99	96	94	98	99	98
C..T	97	98	98	97	99	89	93	98	94	97	99	94	99	96	97	99
C..C	95	94	97	97	96	98	98	99	90	97	97	89	95	96	90	92
C..G	98	93	97	97	94	97	98	98	96	96	97	97	99	98	95	96
T..G	97	90	95	97	97	98	97	98	89	95	99	87	98	95	95	100
T..A	97	97	96	98	98	97	98	97	98	99	99	99	99	96	97	94
C..A	97	92	98	91	98	97	98	98	96	99	95	95	91	99	99	96
A..C	97	86	98	96	99	97	97	97	97	99	92	90	89	96	99	97
A..T	96	92	95	94	96	95	98	99	98	99	99	99	97	89	97	94
G..T	97	93	97	95	99	99	90	89	97	99	83	94	99	96	94	97
G..C	92	96	93	91	97	83	99	83	95	97	97	91	89	99	98	98
A..A	97	97	64	93	96	97	98	98	95	100	99	97	99	98	98	100
A..G	95	97	97	96	97	98	98	98	98	98	96	96	97	92	99	100
G..A	91	96	94	98	100	99	99	91	99	99	100	99	96	99	87	96
G..G	69	92	89	98	99	93	96	99	91	90	90	75	99	94	95	95

Step

12

Table S8. Percentages of α/γ torsions in the canonical sub-state (characterized by α in g- and γ in g+) for all the 256 tetranucleotides obtained from miniABC$_{BSC1}$-Na.

Flanks \ Step	GG	GA	AG	AA	GC	GT	AT	AC	CA	TA	TG	CG	CC	CT	TC	TT
T..T	97	97	98	99	97	95	95	71	92	84	98	96	99	99	99	98
T..C	86	97	90	98	97	70	97	98	95	96	99	95	99	87	93	99
C..T	96	97	96	94	98	96	97	98	98	92	87	79	98	99	99	96
C..C	96	91	95	86	95	98	85	99	95	93	82	97	94	96	99	95
C..G	94	95	95	97	95	97	95	99	95	99	92	99	99	99	98	98
T..G	94	93	88	96	87	93	79	97	92	99	92	90	97	92	90	99
T..A	95	97	98	96	95	98	98	92	95	97	97	96	97	91	87	100
C..A	94	95	97	96	96	99	87	98	97	96	91	93	95	92	99	99
A..C	82	97	97	97	97	96	98	97	99	97	89	99	99	100	96	100
A..T	98	99	82	98	97	98	98	99	95	99	98	100	97	97	95	98
G..T	87	95	97	98	93	96	98	98	95	92	97	92	96	98	97	95
G..C	97	94	92	98	96	98	97	96	97	97	93	98	97	99	94	93
A..A	93	97	98	98	98	88	97	98	95	96	99	99	99	67	99	96
A..G	94	94	98	98	92	98	94	98	97	94	94	96	98	100	88	97
G..A	91	96	98	97	96	98	96	96	72	98	92	93	98	77	99	98
G..G	89	91	93	93	95	94	93	97	98	98	93	97	99	94	99	98

13

SUPPLEMENTARY FIGURES

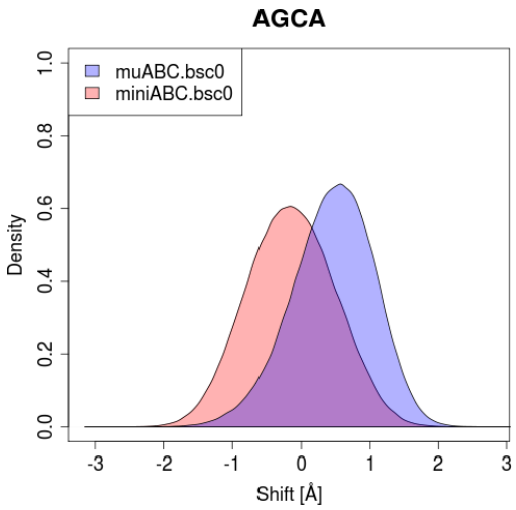


Figure S1. Shift distribution of the AGCA tetranucleotide obtained from μABC$_{BSC0}$-K and miniABC$_{BSC0}$-K. Both are bell-shaped Gaussian distributions, with a similar standard deviation, but different mean. All 1,631 pairs of other analogous marginal distributions were more similar one to the other.

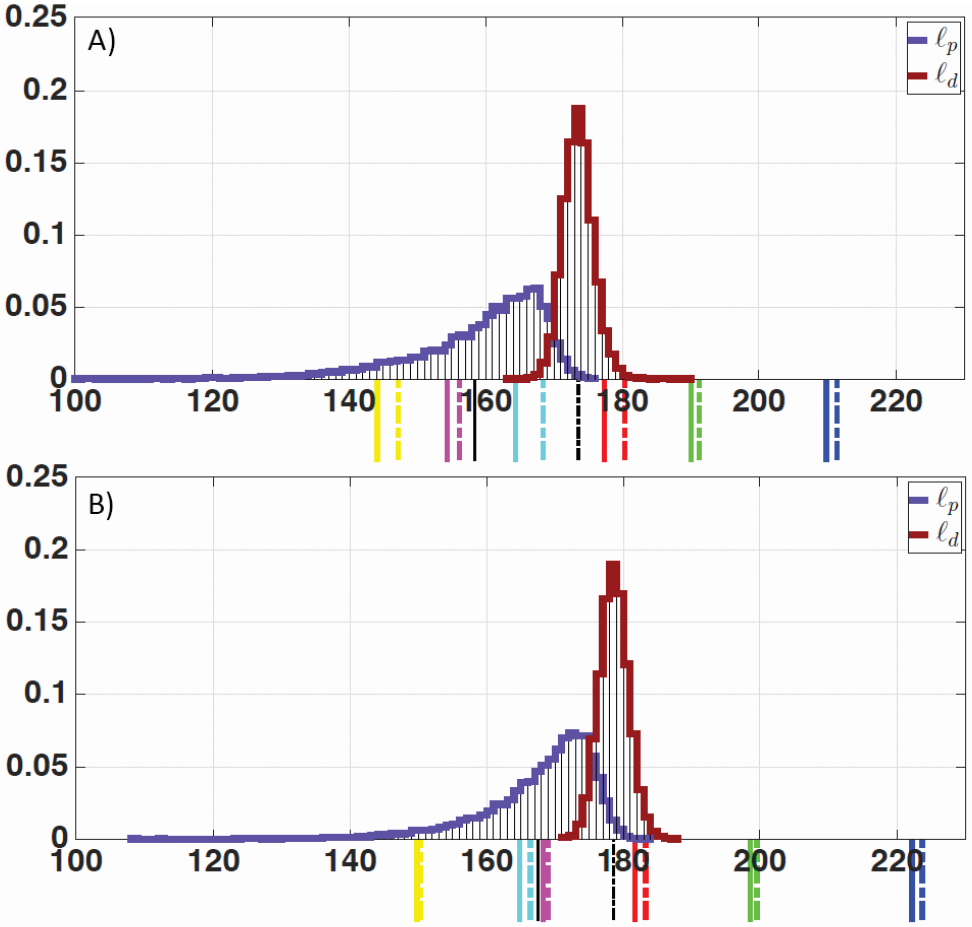


Figure S2. Spectra of ℓ_p (dark blue) and ℓ_d (dark red) persistence lengths computed over an ensemble of 10K sequences for A) PARMBSC0, and B) PARMBSC1 parameter sets, with mean for ℓ_p (black solid line) and mean for ℓ_d (black dashed line). The ℓ_p (coloured solid line) and ℓ_d (dashed solid line) values for the 6 distinct dinucleotide tandem repeats are also indicated in each case. The x-axis is in units of basepair, while the frequency is reported on the y-axis. Note that using an average rise of 0.33 nm, the peaks reported between 160 to 180 base pairs represent persistence lenths between 52 to 59 nm.

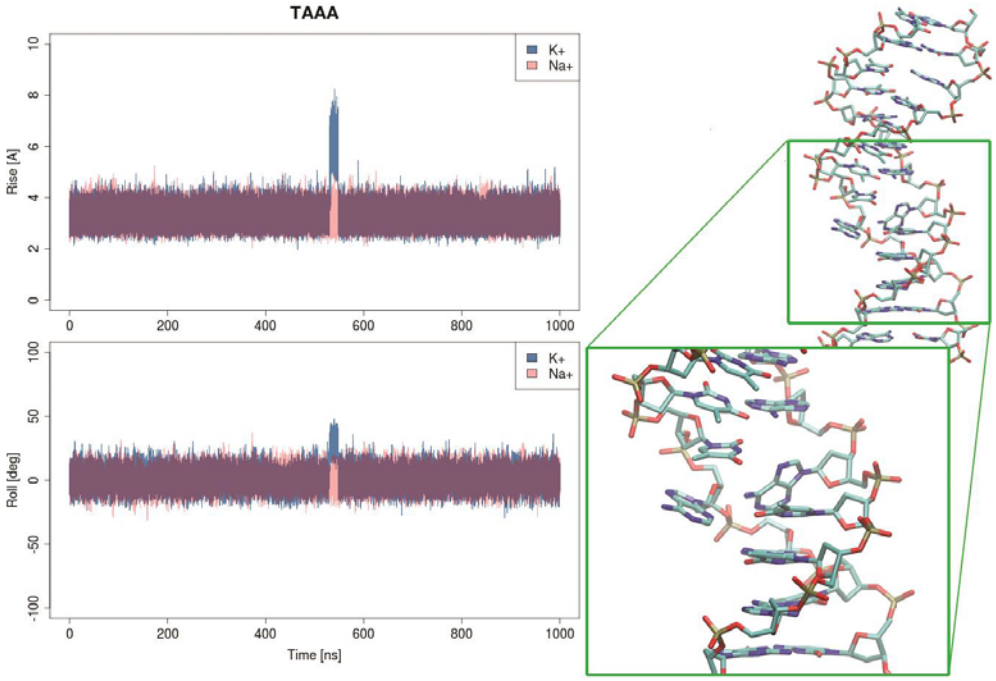


Figure S3. Time evolution of rise and roll for the TAAA tetranucleotide. The trajectory performed in K^+ (blue) shows the formation of a reversible kink near 550 ns, not present using Na^+ (pink). During the formation of the kink, up to two consecutive adenines lose their Watson-Crick H-bonds and are partially un-stacked. Note that this local distortion does not affect the main double helical structure of the oligomer.

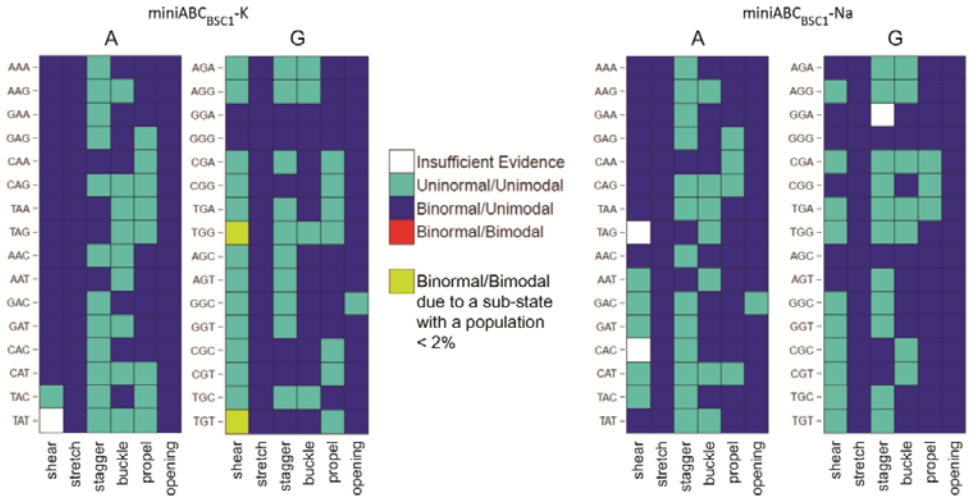


Figure S4. Structural polymorphisms (normality and modality) in intra-basepair helical conformations for all distinct trinucleotides. Results obtained from miniABC$_{BSC1}$-K and miniABC$_{BSC1}$-Na.

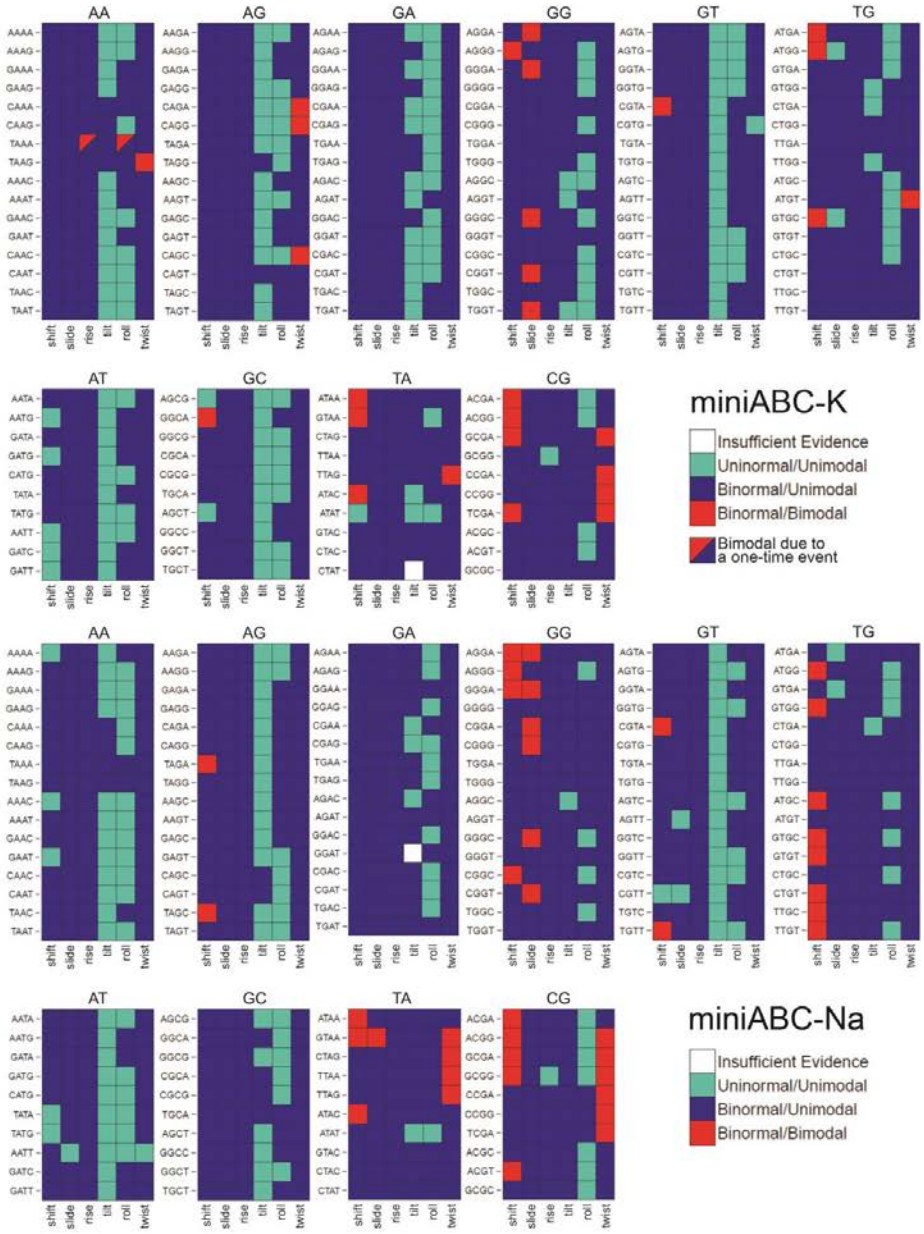


Figure S5. Structural polymorphisms (normality and modality) in inter-basepair helical conformations for all the 136 distinct tetranucleotides. Results obtained from miniABC$_{BSC1}$-K (top) and miniABC$_{BSC1}$-Na (bottom). Tetranucleotides classified as binormal/bimodal (red) are considered as polymorphic (exist in two clear conformational sub-states).

18

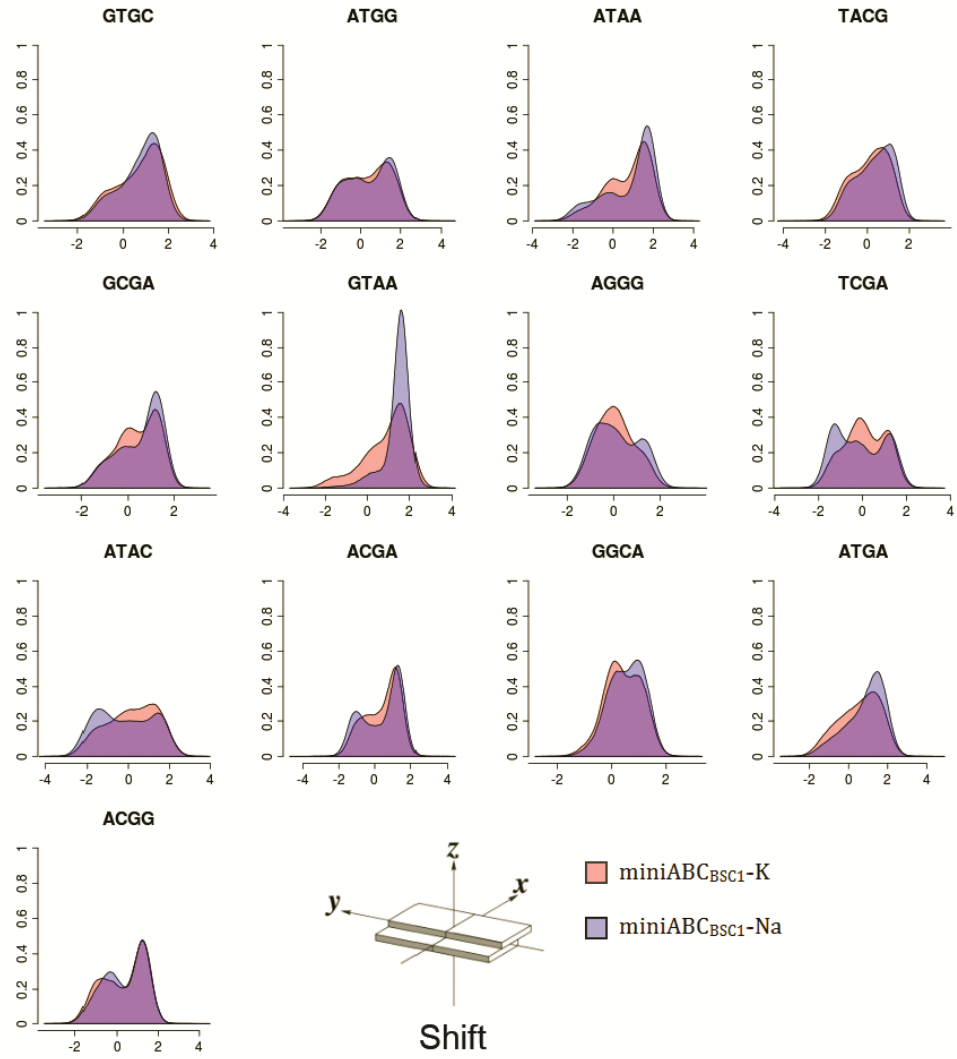


Figure S6. Normalized shift distributions for all the bimodal cases found in the miniABC$_{BSC1}$-K dataset, overlapped with their counterpart computed using Na+. X-axes represent the shift helical parameter in Å.

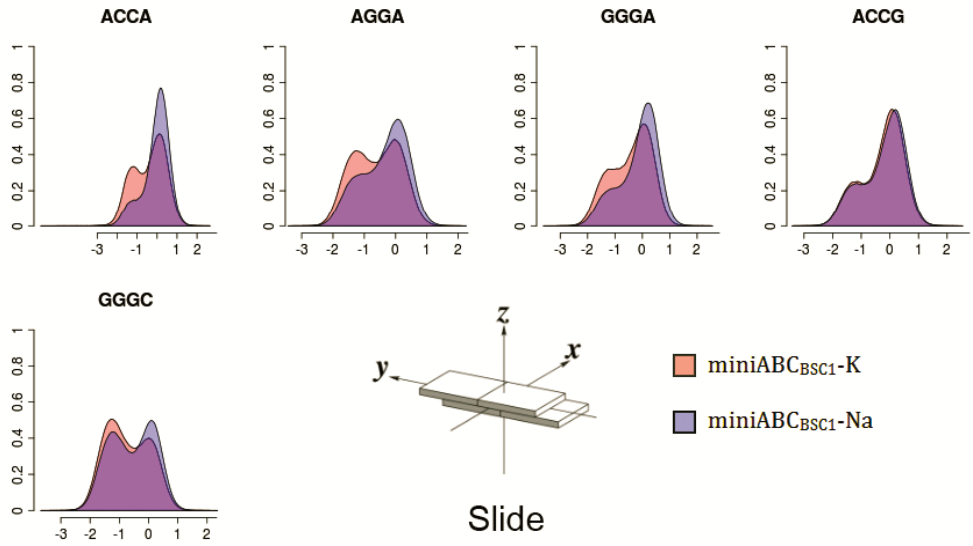


Figure S7. Normalized slide distributions for all the bimodal cases found in the miniABC$_{BSC1}$-K dataset, overlapped with their counterpart computed using Na+. X-axes represent the slide helical parameter in Å.

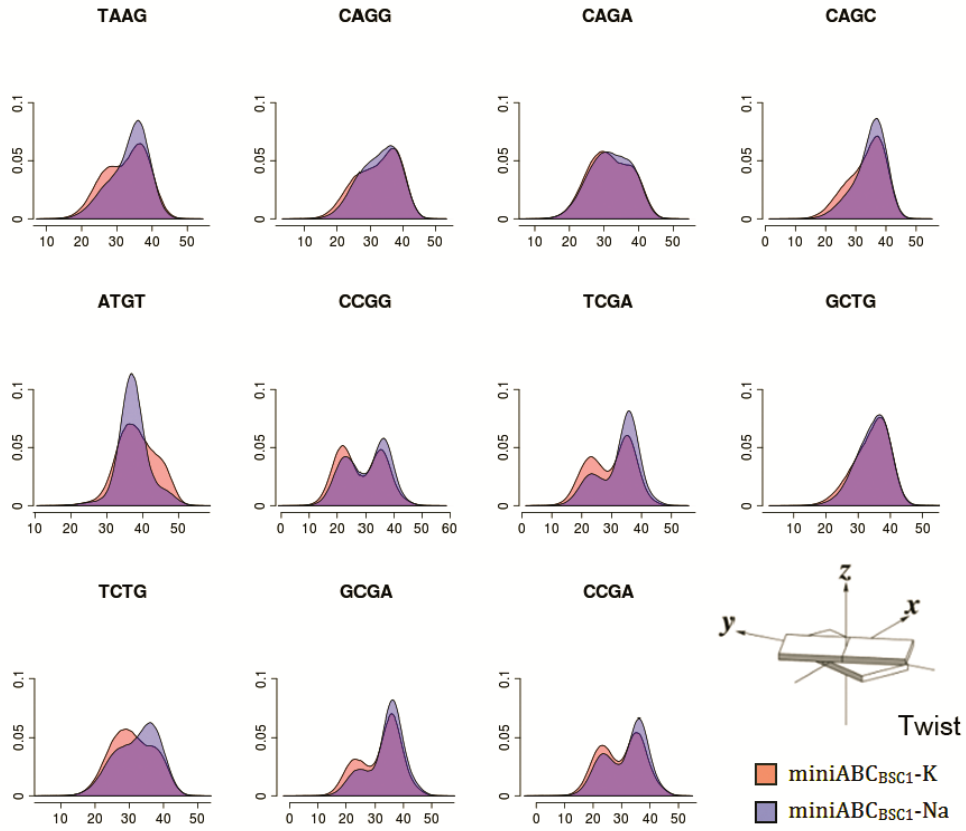


Figure S8. Normalized twist distributions for all the bimodal cases found in the miniABC$_{BSC1}$-K dataset, overlapped with their counterpart computed using Na+. The x-axes represent the twist helical parameter in degrees. Nota that the two peaks observed are in agreement with X-ray structures of DNA and protein-DNA complexes deposited in the Protein Data Bank(*11*).

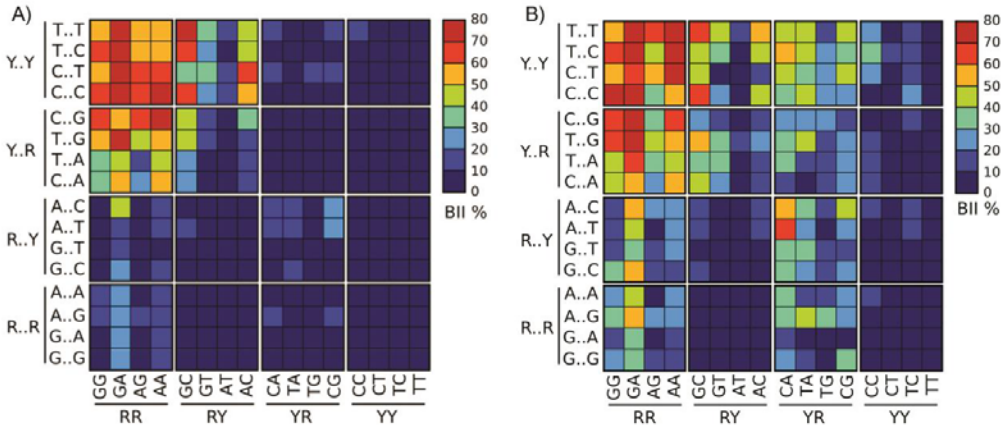


Figure S9. Sequence dependence of BII backbone conformations. The percentage occurrence of BII backbone states for the phosphodiester junction at the central base step of each of the 256 possible tetranucleotide sequences is shown (BII%), using the color code defined on the right (0% is dark blue; 80% is dark red). The sequences are arranged so that each column represents one of 16 dinucleotide steps, and each row corresponds to one of the 16 possible flanking sequences; columns and rows are further grouped on the basis of base type (R = purine and Y = pyrimidine). A) μABC$_{BSC0}$-K BII percentages(*24*); B) miniABC$_{BSC1}$-K BII percentages.

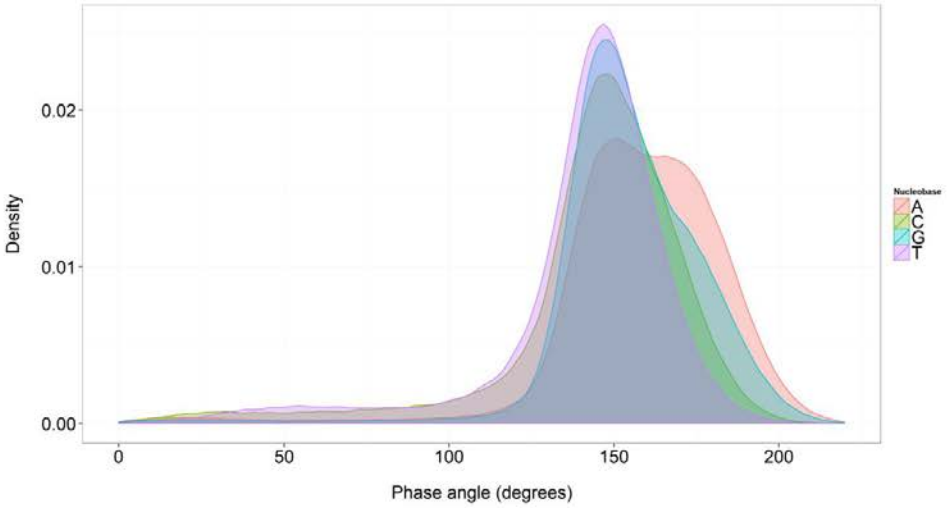


Figure S10. Normalized distribution of the P angle for A, C, G and T bases (in degrees), obtained from miniABC$_{BSC1}$-K dataset.

23

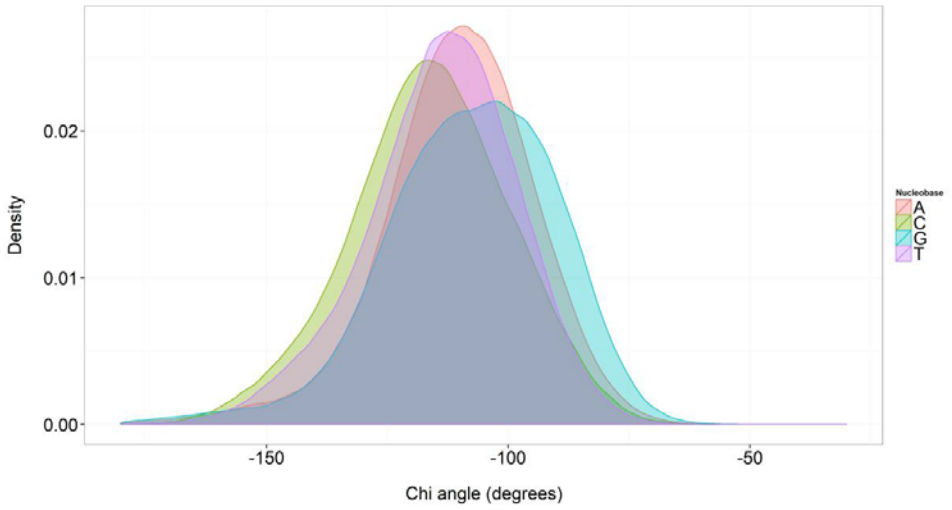


Figure S11. Normalized distribution of the χ angle for A, C, G and T bases (in degrees), obtained from miniABC$_{BSC1}$-K dataset.

24

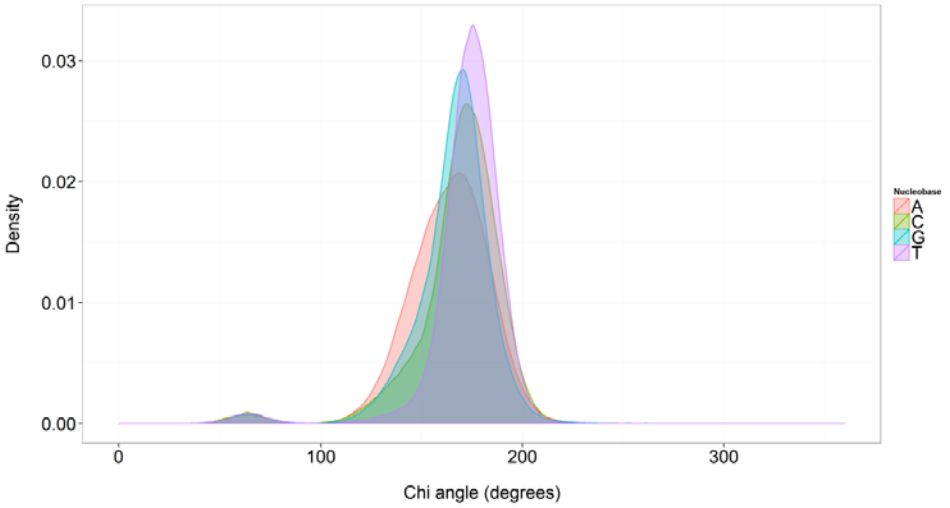


Figure S12. Normalized distribution of the β angle for A, C, G and T bases (in degrees), obtained from miniABC$_{BSC1}$-K dataset.

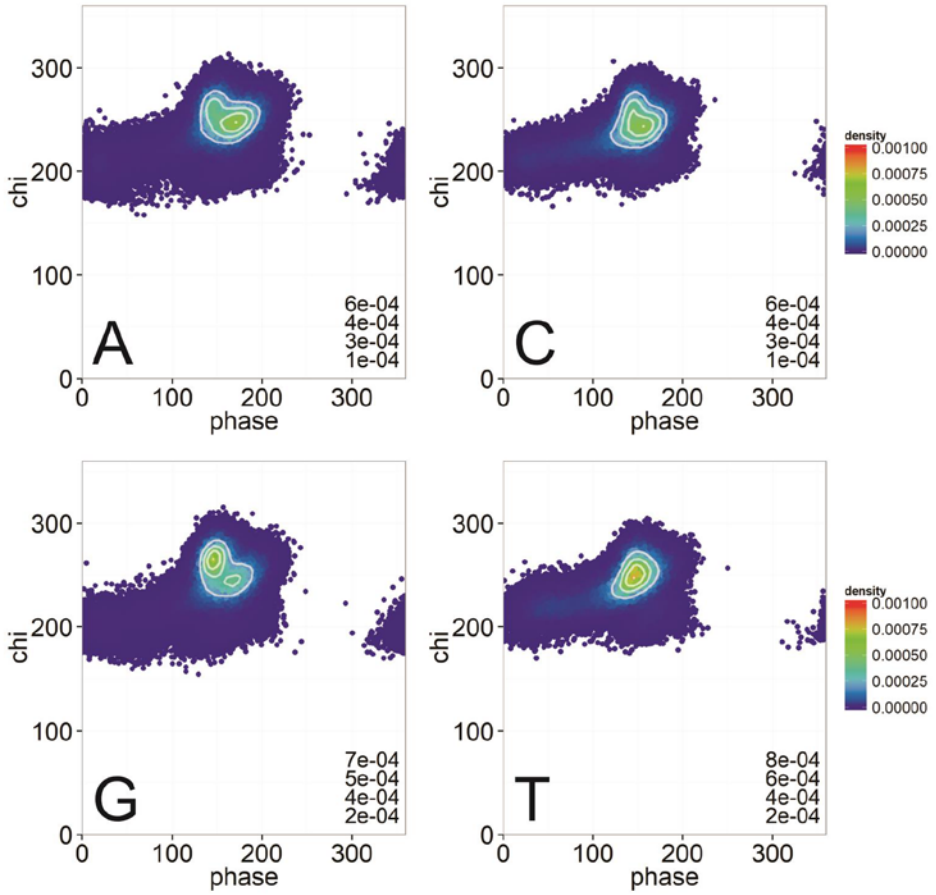


Figure S13. Phase vs χ distribution plot (in degrees) obtained from miniABC$_{BSC1}$-K dataset for A, C, G and T bases.

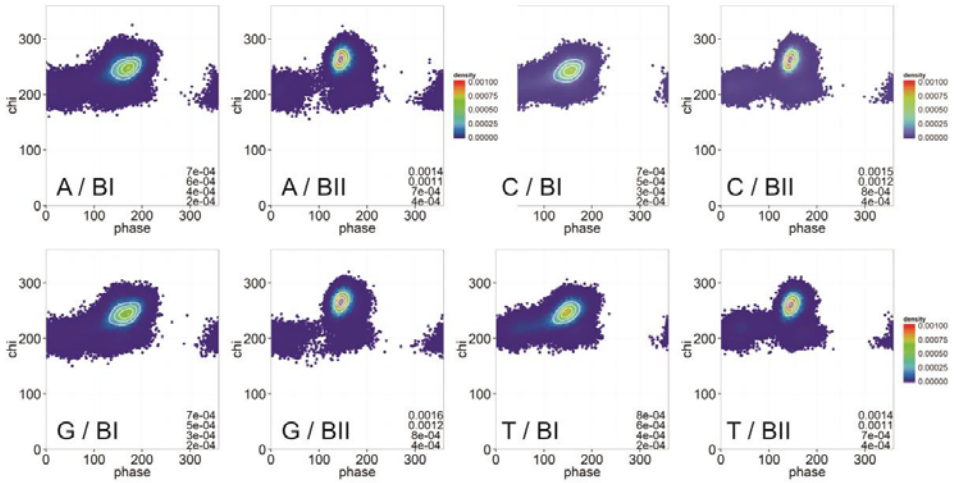


Figure S14. Phase vs χ distribution plot (in degrees) obtained from miniABC$_{BSC1}$-K dataset and filtered according to BI/BII for A, C, G and T bases.

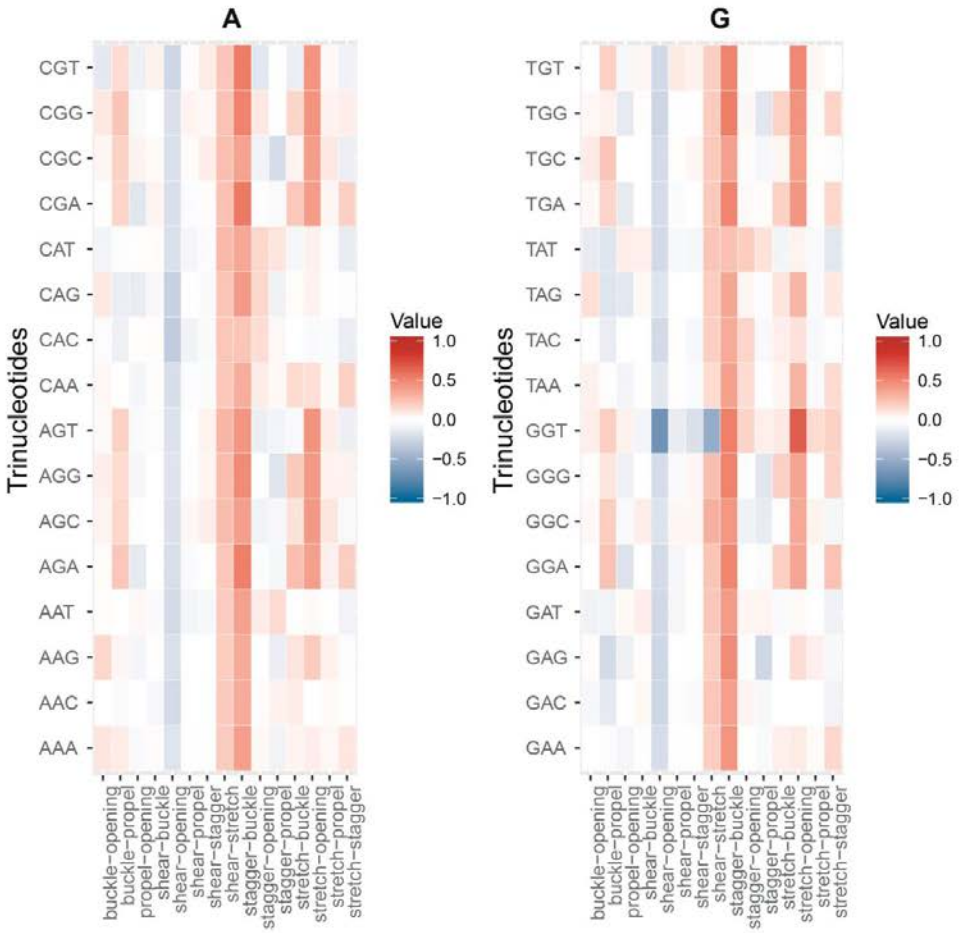


Figure S15. Correlation coefficients between intra-basepair helical parameters (shear, stretch, stagger, propeller, buckle and opening) belonging to the same base pair in the Watson strand. Results obtained from miniABC$_{BSC1}$-K dataset for all bps.

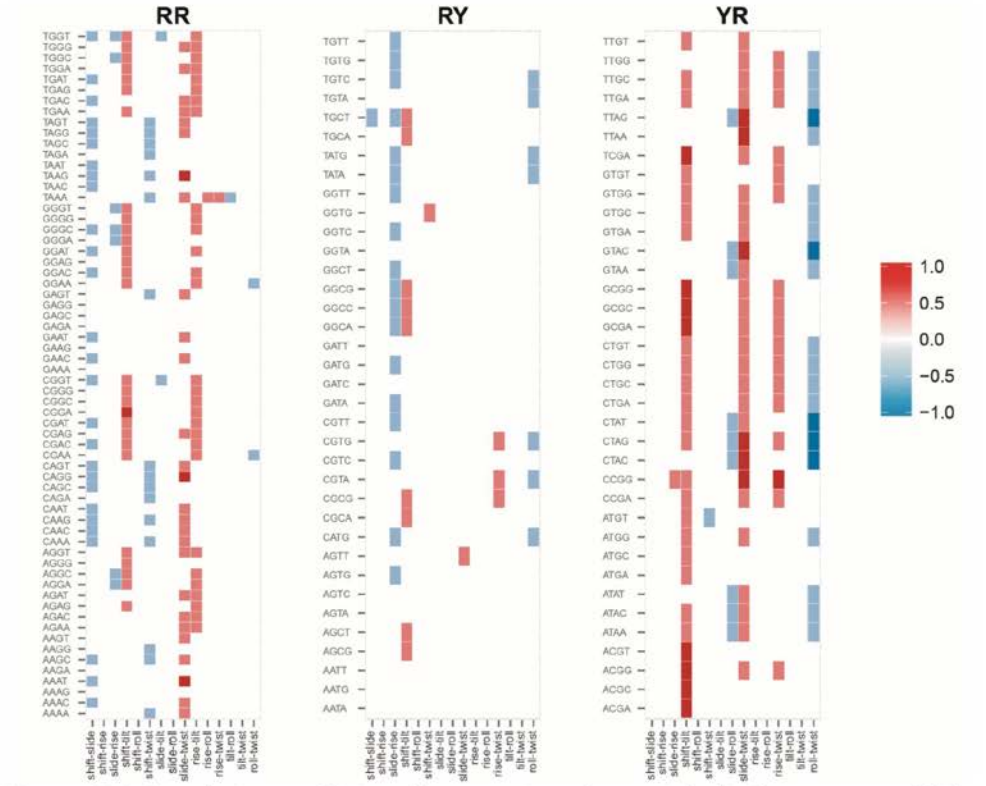


Figure S16. Correlation coefficients between inter-basepair helical parameters (shift, slide, rise, tilt, roll, and twist) belonging to the same step in the Watson strand. Results obtained from miniABC$_{BSC1}$-K dataset for all RR, RY and YR bps.

29

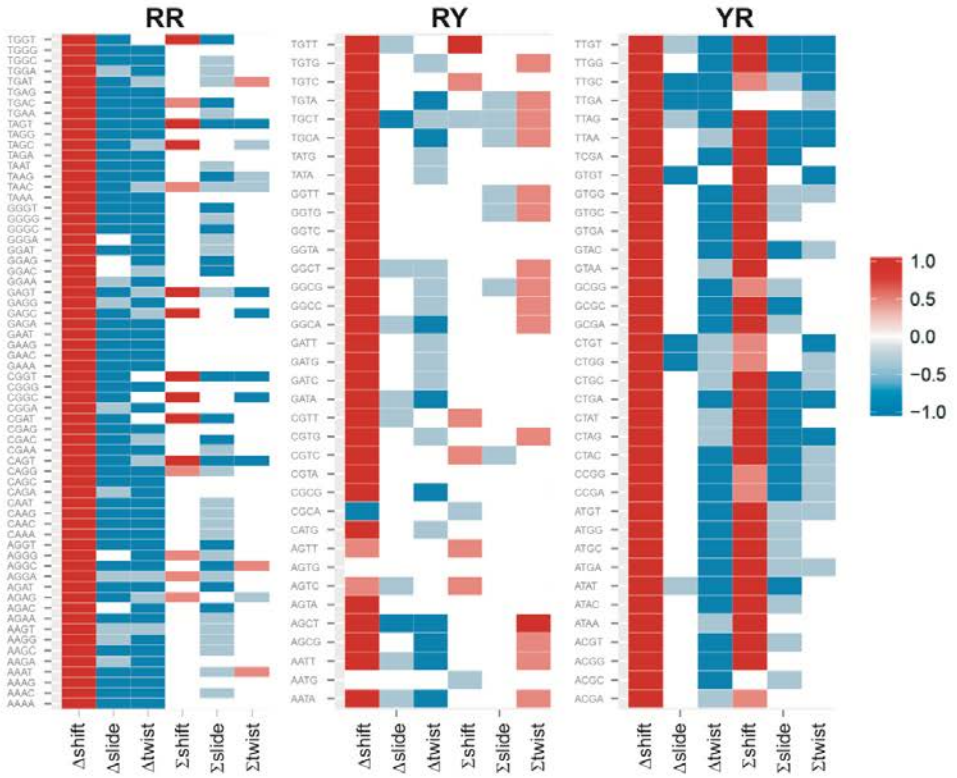


Figure S17. Correlation coefficients between differences (Δ) and sums (Σ) of inter-basepair parameters and the BII state in the central junction. Results obtained from miniABC$_{BSC1}$-K dataset for all steps grouped by RR, RY and YR.

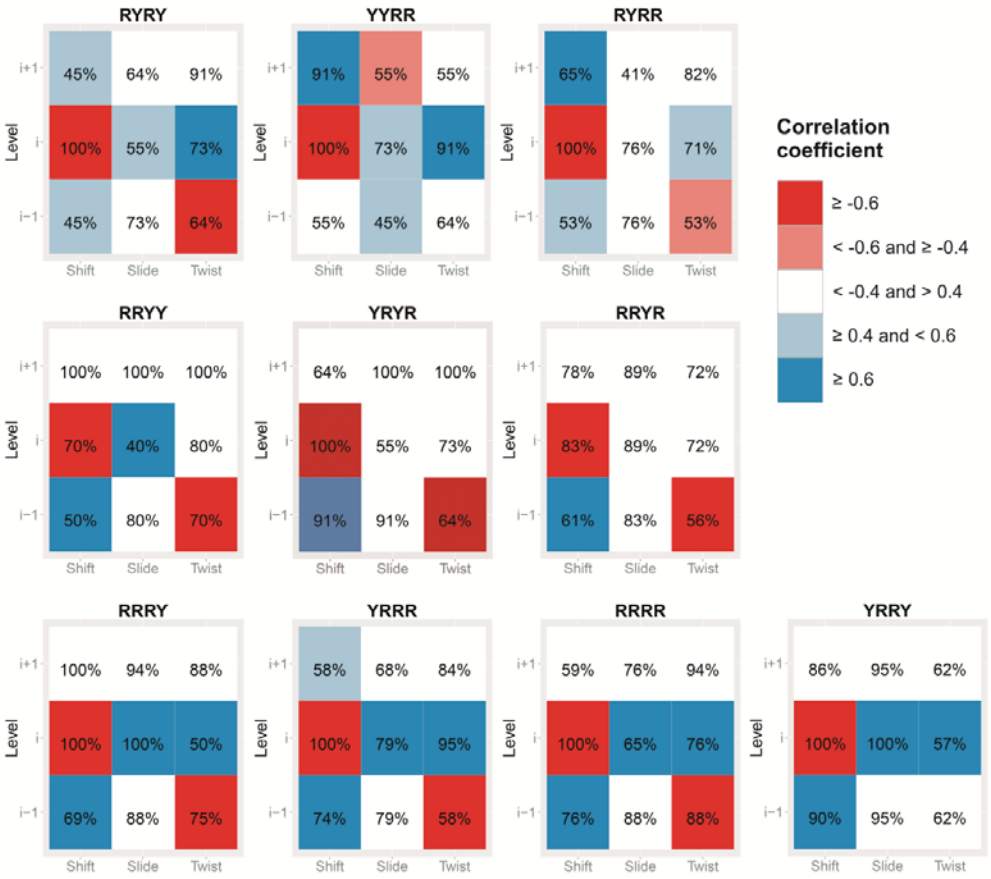


Figure S18. Correlation coefficients between shift, slide, or twist at the positions i-1 (5'-side), i, and i+1 (3'-side), and the backbone substate at the junction of inter-basepair i in the Watson strand. Results obtained from miniABC$_{BSC1}$-K dataset. The numbers inside each cell represent the % of specific tetranucleotides within a given family that give rise to the correlation.

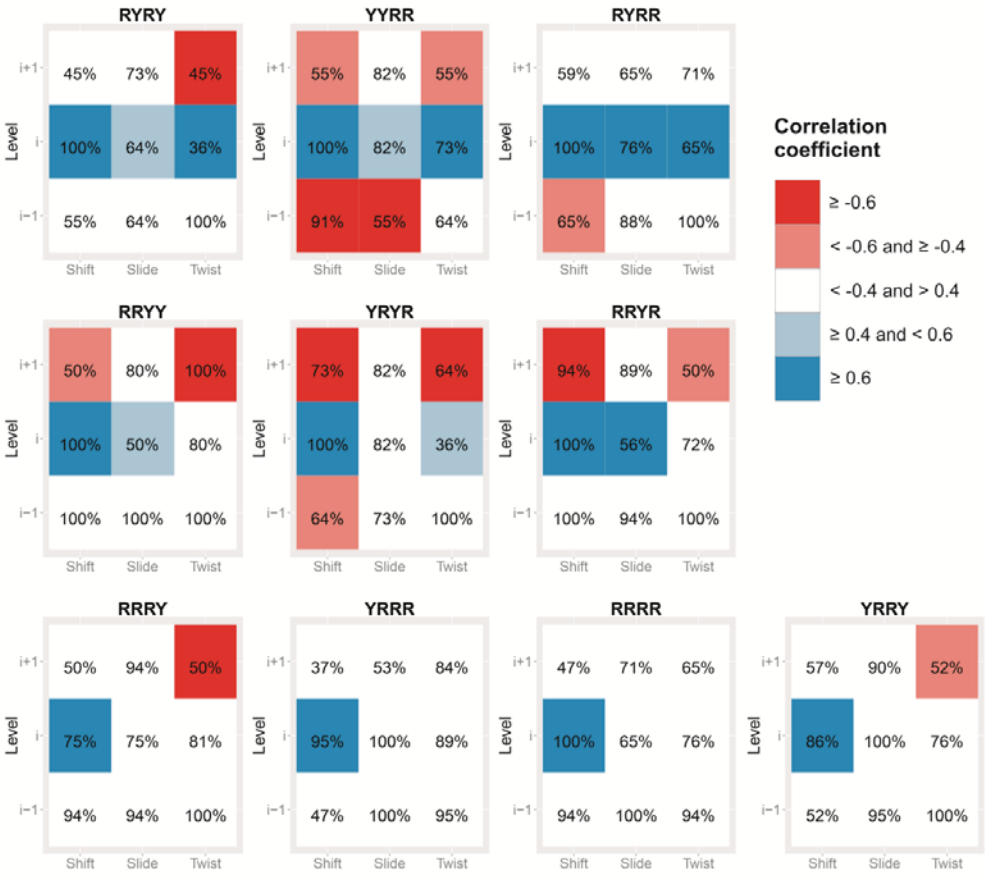


Figure S19. Correlation coefficients between shift, slide, or twist at the positions i-1 (5'-side), i, and i+1 (3'-side), and the backbone substate at the junction of inter-basepair i in the Crick strand. Note that we refer everything to the Watson strands (see Methods), so in this plot, RRRR means YYYY since we are analyzing the correlation with the Crick strand. Results obtained from miniABC$_{BSC1}$-K dataset.

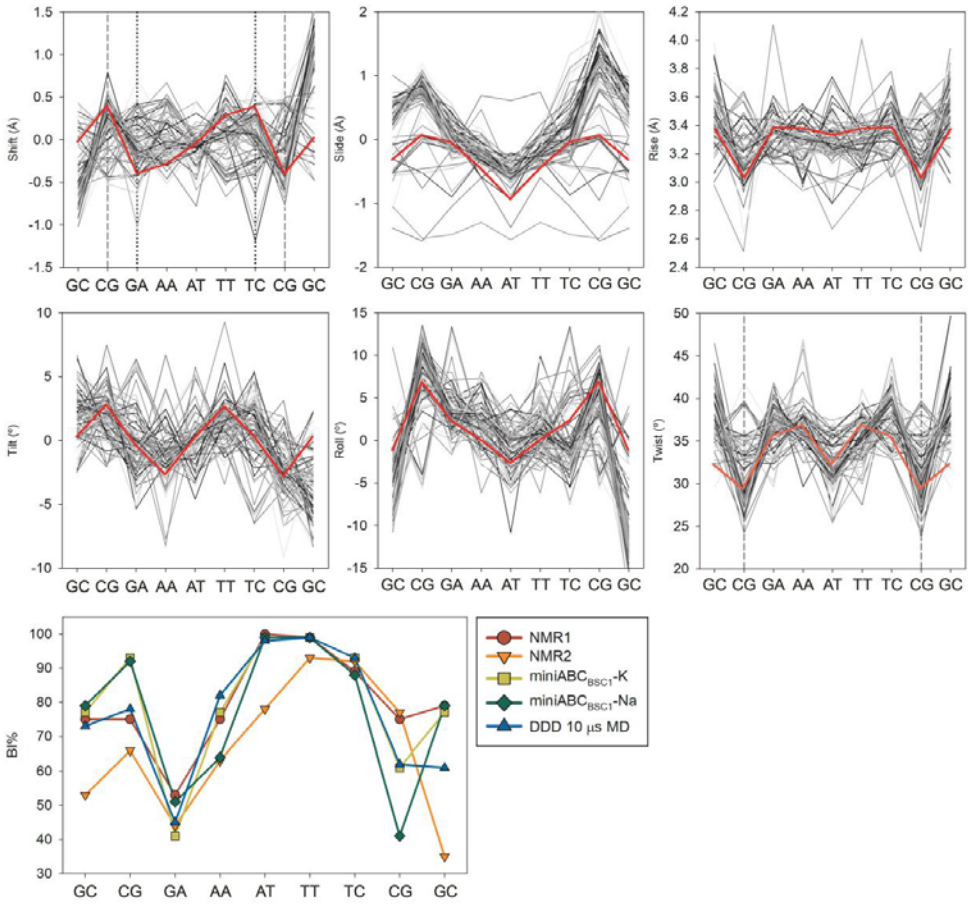


Figure S20. Comparison between experimental structures (X-ray and NMR) determined for the Drew-Dickerson Dodecamer(*25*), and conformations predicted by using the MD datasets produced herein in conjunction with our extended rules. The six intra-basepair parameters were predicted (red lines), and compared with all the experimental structures (grey lines). Vertical dashed lines represent predicted binormal/bimodal steps, while vertical dotted lines represent steps with clear multi-peaked distributions although not bimodal according to Helguerro (see Methods). In the last row, predicted BI% (yellow and green) were compared with ^{31}P-NMR gold-standard measurements, and a very long MD simulation of the same sequence using PARMBSC1 force field(*25*). NMR1 stands for the work by Schwieters *et al.*(*26*), and NMR2 from Tian *et al.*(*27*).

REFERENCES

1. S. Arnott, D. W. L. Hukins, Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.* **47**, 1504–1509 (1972).
2. H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
3. A. Pérez *et al.*, Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92**, 3817–29 (2007).
4. I. Ivani *et al.*, Parmbsc1: A refined force field for DNA simulations. *Nat. Methods.* **13**, 55–58 (2015).
5. L. X. Dang, Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J. Am. Chem. Soc.* **117**, 6954–6960 (1995).
6. M. Pasi *et al.*, μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **42**, 12272–12283 (2014).
7. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
8. J.-P. Ryckaert, G. Ciccotti, H. J. . Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
9. G. Schwarz, Estimating the Dimension of a Model. *Ann. Stat.* **6**, 461–464 (1978).
10. R. E. Kass, A. E. Raftery, Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
11. P. D. Dans, A. Pérez, I. Faustino, R. Lavery, M. Orozco, Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.* **40**, 10668–10678 (2012).
12. F. de Helguero, Sui Massimi Delle Curve Dimorfiche. *Biometrika*. **3**, 84 (1904).
13. M. F. Schilling, A. E. Watkins, W. Watkins, Is Human Height Bimodal? *Am. Stat.* **56**, 223–229 (2002).
14. G. V Glass, K. D. Hopkins, *Statistical methods in education and psychology* (Boston : Allyn & Bacon, 3rd ed., 2008).
15. K. Pearson, Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. London*. **58**, 240–242 (1895).
16. C. R. Calladine, Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.* **161**, 343–52 (1982).
17. C. R. Calladine, *Understanding DNA : the molecule & how it works* (Elsevier Academic Press, 2004).
18. D. V. Lindley, Information Theory and Statistics. *Solomon Kullback* . New York: John Wiley and Sons, Inc.; London: Chapman and Hall, Ltd.; 1959. Pp. xvii, 395. $12.50. *J. Am. Stat. Assoc.* **54**, 825–827 (1959).
19. J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, J. H. Maddocks, Sequence-Dependent Persistence Lengths of DNA. *J. Chem. Theory Comput.* **13**, 1539–1555 (2017).
20. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing (2013).
21. T. E. Oliphant, Python for Scientific Computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).

22. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
23. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–8, 27–8 (1996).
24. A. Balaceanu *et al.*, The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.* **8**, 21–28 (2017).
25. P. D. Dans *et al.*, Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.* **44**, 4052–4066 (2016).
26. C. D. Schwieters, G. M. Clore, A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. *Biochemistry*. **46**, 1152–66 (2007).
27. Y. Tian *et al.*, ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J. Phys. Chem. B*. **113**, 2596–2603 (2009).

1.2 Higher than tetranucleotide effects of d(CpTpApG) (Publication 2)

This work expands the previous study of tetranucleotide effects on the central base pair step to effects of a specific tetranucleotide in different hexa- and octanucleotide environments. The chosen tetranucleotide, CTAG, showed unusual flexibility behavior in the trajectories deposited in our BigNAsim database as well as in the study of the 136 unique tetranucleotides. In total we studied 40 different sequence contexts. In the analysis of these trajectories we find evidence of intrinsic multi-modality of the individual trajectories in three inter base pair parameters (shift, slide and twist). Shift distribution is tri-modal, while twist and slide distributions are bimodal but only 4 specific combinations of substates are possible (see Figure 23). Additionally, different sequence environments experience large differences in the distributions of these helical coordinates of the d(TpA) step (see Figure 2 in the following publication).

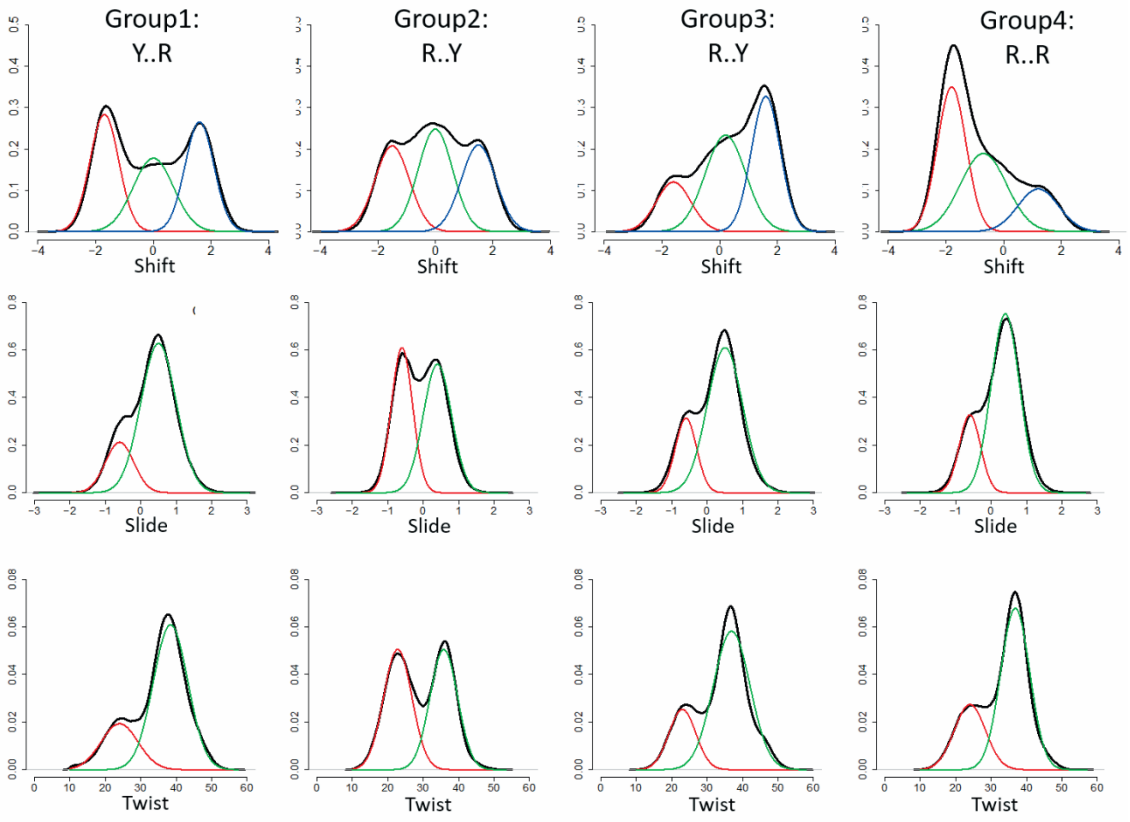


Figure 23. Normalized frequencies for shift, slide and twist (black line), and the BIC decomposition in Gaussians (red, green, and blue lines) of the four groups obtained by PCA and subsequent clustering

The pathway of the information traveling through concerted movements of backbone and bases which are also coupled to the formation of unconventional hydrogen bonds influences the central d(TpA) step up to octamer level. We further examine the remote effects beyond hexanucleotide level (see Figure 24), pointing out which types of sequences are more susceptible to transmitting information and at which steps communication vanishes.

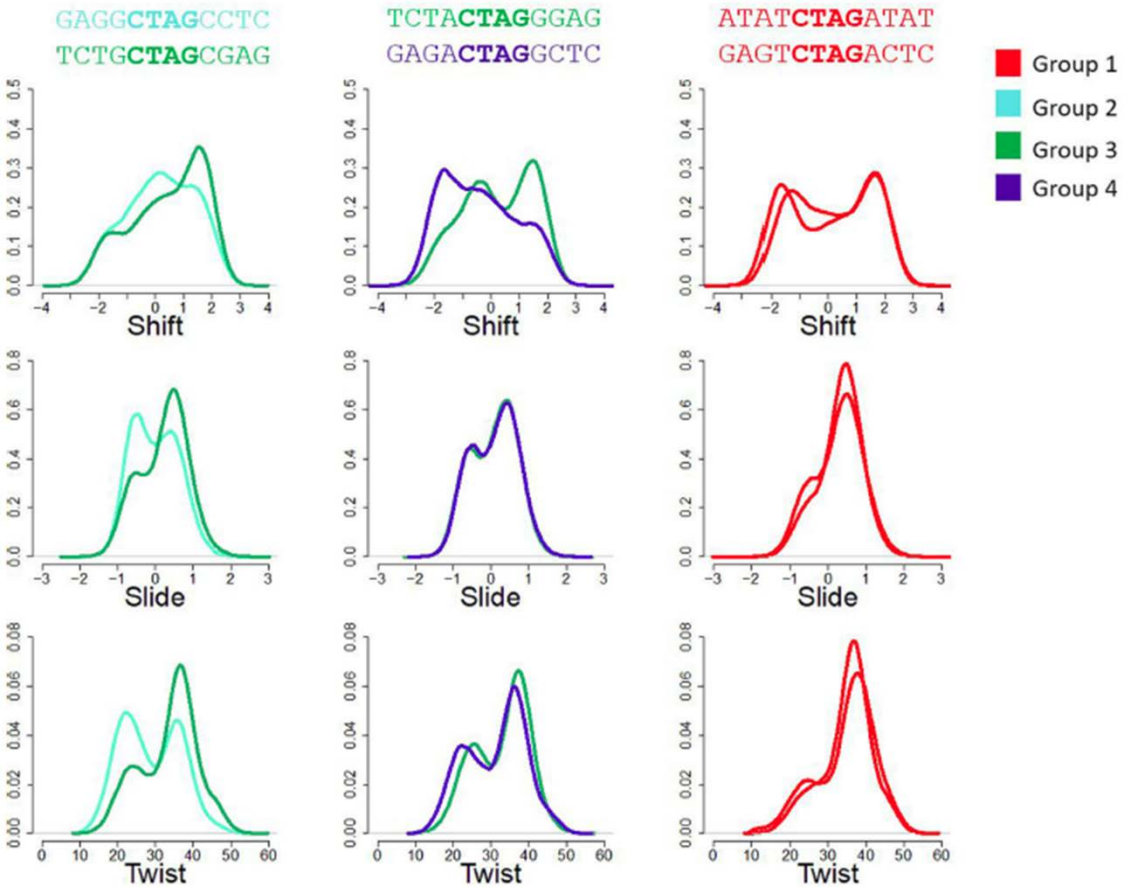


Figure 24. Normalized frequencies of shift, slide and twist at the central TpA step for three pairs of selected sequences showing non-negligible effects beyond next-to-nearest neighbours. The colours used are related to the groups found in the clustering analysis.

PCA and subsequent clustering (for more details see 'Materials and Methods' of following publication) show 4 distinct clusters of hexanucleotide variability, with all four clusters experiencing distinct pattern of flanking purines or pyrimidines (see Figure 5 in the following publication). A comparison of the resolved structures in the PDB database containing CTAG reveals values for the shift, slide, roll and twist helical parameters that cover the multi-modal distributions obtained in our trajectories (see Figure 8 in the following publication), confirming our claims on the bimodal nature of slide and twist, with peaks in the distributions that fit well to

our results which provides an indirect, but strong support of the 4-state model of the dynamics of the central junction in CTAG.

Finally, an analysis of the genome of several different species uncovered that CTAG is one of the lowest populated tetranucleotides appearing mainly on intergenic regions and very rarely in genes (see Figure 9 and Supplementary Figure 10 of the following publication). Due its good conservation, we can conclude that CTAG is important for the functionality of the cell or they are easily accessible to the mismatch repairing machinery. The low mutation rate of CTAG in cancer cell lines suggests that the cell takes advantage of the unusual properties of CTAG as points of high flexibility that might help to fold chromatin.

In summary, this in-depth study uncovered previously unknown features of one of the most structurally polymorphic tetranucleotides found in B-DNA. Analysis of the helical space of all hexanucleotide environments and some octamer sequence contexts alongside with data mining in the PDB data base lead to the assumption that CTAG exists in different conformational substates where inter base pair parameters are tightly coupled to the backbone by concerted and correlated movements which lets information travel up to half a helical turn away from the affected base pair step.

Publication:

Alexandra Balaceanu, Diana Buitrago, Jürgen Walther, Pablo D. Dans and Modesto Orozco; Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond, Nucleic Acids Research, 2019, doi: 10.1093/nar/gkz255

Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz255

Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond

Alexandra Balaceanu[1], Diana Buitrago[1], Jürgen Walther[1], Adam Hospital[1], Pablo D. Dans[1] and Modesto Orozco[1,2,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain and [2]Department of Biochemistry and Biomedicine, University of Barcelona, 08028 Barcelona, Spain

Received February 04, 2019; Revised March 23, 2019; Editorial Decision March 28, 2019; Accepted March 30, 2019

ABSTRACT

We used extensive molecular dynamics simulations to study the structural and dynamic properties of the central d(TpA) step in the highly polymorphic d(CpTpApG) tetranucleotide. Contrary to the assumption of the dinucleotide-model and its nearest neighbours (tetranucleotide-model), the properties of the central d(TpA) step change quite significantly dependent on the next-to-nearest (hexanucleotide) sequence context and in a few cases are modulated by even remote neighbours (beyond next-to-nearest from the central TpA). Our results highlight the existence of previously undescribed dynamical mechanisms for the transmission of structural information into the DNA and demonstrate the existence of certain sequences with special physical properties that can impact on the global DNA structure and dynamics.

INTRODUCTION

Early structural models of DNA derived from fibre diffraction data provide a static and averaged picture of the double helix (1–3), which despite its simplicity was sufficient to represent the general shape of DNA in physiological conditions. However, as more accurate structural techniques appeared, the intrinsic polymorphism of double-stranded DNA become evident (4–7) as significantly different conformations were described depending on the sequence, the environment or the presence of ligands (8–11). Six decades after the development of the first duplex models, we understand that DNA as a flexible and polymorphic molecule is able to sample a wide range of helical geometries, thanks to a complex choreography of backbone rearrangements, which allows the conformational changes required for DNA functionality (11–19).

Attempts to determine the principles relating sequence and structure originated in the eighties when by processing the scarce experimental data available, Calladine *et al.* (20), developed a series of heuristic rules relating sequence with some structural characteristics of DNA (21,22). In the late nineties (23), Olson *et al.* developed a complete set of parameters defining the expected distribution of helical parameters of the 10 unique base pair steps (bps). Parameters were derived from the analysis of the available crystal data on DNA–protein complexes and provided information not only on the equilibrium geometry but also on the expected flexibility of the bps (extracted from the variability of the same bps in different crystals). Twenty years after their generation, Olson-Zhurkin parameters are still used to represent DNA by means of helical mesoscopic descriptors. However, we cannot ignore the strong assumptions involved in their derivation: (i) the ensemble of configurations obtained from the analysis of crystal structures should define a densely populated Gaussian distribution; (ii) a dinucleotide (step) model is enough to represent DNA sequence variability, i.e. the helical geometry can be decomposed at the bps level; (iii) conformational variability found in structures in PDB should exclusively depend on the flexibility of the step and finally (iv) binding of a protein should not introduce anharmonic distortions in the duplex geometry.

The eruption of atomistic molecular dynamics (MD) simulations gave the community an alternative source of parameters to describe DNA structure and flexibility. Compared with results derived from the analysis of experimental structures, the MD-based ones are more robust as they are obtained from processing an extremely large number of snapshots, and provide information on flexibility that is not contaminated by the presence of ligands, crystal lattice or any other environmental artifacts. As a major caveat, MD-derived descriptions of DNA properties are dependent on the length of trajectories as well as on the quality of the force field parameters used to describe DNA interactions. Thus, early attempts to describe DNA from multinanosecond trajectories led to artefactual results due to a previously unknown error of the most used force field at that time (24). A newer force field (25) and higher computa-

*To whom correspondence should be addressed. Tel: +34 93 40 37156; Email: modesto.orozco@irbbarcelona.org

tional capabilities provided descriptions of DNA properties that were more reasonable, but still far from the required accuracy (12,26,27). The availability of the highly accurate PARMBSC1 force field (28,29) and the development of new MD codes taking advantage of a new generation of computers (30–33) provide the community with the possibility to derive reliable representation of the sequence-dependent physical properties of DNA from the analysis of microsecond long trajectories collected under highly controlled simulation conditions.

Results collected by the Ascona B-DNA Consortium (34–37) revealed two major findings that challenged current models of DNA flexibility. First, the dinucleotide-model is insufficient to describe DNA flexibility, as the variability in bps parameters depending on tetranucleotide environment can be more pronounced than the variability found when comparing different bps for a given tetranucleotide context. Second, several distributions of helical parameters considering the nearest neighbours deviate from normality and a part of them are in fact multi-modal, which means that the physical properties of such tetranucleotides cannot be represented by a single set of elastic parameters (equilibrium values and associated stiffness). Analysis of MD data revealed that the changes between substates happen towards a series of coordinated changes along the backbone (17,37,38), where unusual H-bond interactions and subtle changes in the solvent environment play a key role (18,39). The analysis of ABC data and of additional trajectories stored in our BigNASim database (40) suggested that a nearest neighbour-based model was in general sufficient to derive transferable descriptors of DNA structure and flexibility, but a few exceptions to this general rule emerged; the clearest one is the d(CpTpApG) tetranucleotide (in the following CTAG): a very polymorphic stretch of DNA, with 50% G-C content, for which results were significantly different depending on the simulation. The structural peculiarities of TpA steps have been qualitatively pointed out in the past by analysing a small number of experimental structures, especially when immersed in short A-tracks (41,42).

We present here a detailed analysis of CTAG in different sequence contexts. Results demonstrate that next-to-nearest effects modulate the geometrical properties of the central d(TpA) step. Such structural effects are very visible when hexanucleotides are considered, but quite surprisingly extend beyond the next-to-nearest level, indicating the existence of a complex mechanism of information transfer across DNA through the coordinated backbone and base movements.

MATERIALS AND METHODS

The choice of sequences and the simulation details

The systematic study of sequence-dependent effects beyond the tetranucleotide level has been to date impossible, due to the huge number of sequences that need to be considered. For example, the study of all hexanucleotides would require the simulation of 2,080 sequences, while to consider all octanucleotides 32,826 sequence combinations are needed. Fortunately, the analysis of ABC simulations where tetranucleotides appear in different molecular environments suggests that sequences effects beyond the tetranucleotide are rare, and if they exist, are localized in certain ultra-flexible sequences. We focused our interest here in one of the most flexible tetranucleotide: CTAG. Thus, we built a library of 40 different sequences covering the entire hexanucleotide space (XpCpTpApGpX) as well as all possible pyrimidine(Y)/purine(R) combinations at the octanucleotide level in several repeats (see Supplementary Methods). All the sequences were prepared using the leap module of AMBERTOOLS 16 (43) and standard ABC protocol (37). Accordingly, systems were built from Arnott's B-DNA average parameters, neutralizing the DNA with K^+ ions, adding water (at least 10 Å of water separate DNA from the faces of the box) and extra 150 mM KCl. Systems were then optimized, thermalized and equilibrated before production (34,35). Water was represented with the SCP/E model (44), Smith-Dang parameters were used for ions (45–47) and the recent PARMBSC1 force field was considered to represent nucleic acids interactions (28). Trajectories (collected in the NPT ensemble $T = 298$ K, $P = 1$ atm) were extended from 0.5 μs to up to 9 μs. All simulations were performed with the pmemd.cuda code using periodic boundary conditions and Particle Mesh Ewald (31,48). Movements of hydrogen atoms were annihilated using SHAKE (49), which allowed us the use of a 2 fs integration step. All trajectories collected here are accessible through the MuG BigNASim portal (40): https://mmb.irbbarcelona.org/BIGNASim/

Analysis

Standard analysis was done using *cpptraj* module of the AMBERTOOLS 16 package (43), the NAFlex server (50) CURVES+ and CANAL programs (51), following the standard ABC-conventions (37). The CANION module from Curves+ (38) was used to determine distributions of ion populations in curvilinear cylindrical coordinates for each snapshot of the simulations with respect to the instantaneous helical axis. Duplexes were named following the Watson strand (e.g. ATGG stands for (ATGG)·(CCAT)). The letters R, Y and X stand for a purine, a pyrimidine or any base respectively, while X:X and XX represent a base pair and base pair step, respectively. Base pairs flanking the CTAG were denoted using two dots to represent the central tetrad (e.g. R··Y). The normality and modality of the helical distributions were evaluated using Bayesian Information Criteria (52,53) and Helguerro's theorem (54) as described elsewhere (12). Classification of the torsional states of the different rotatable bonds in the DNA backbone was done using standard criteria (55). Correlations between different torsions were determined by circular correlation analysis (see Supplementary Methods for additional details). The meta-trajectory analysis was used to define the global characteristic of the d(TpA) essential deformation space. With this purpose, the 40 individual trajectories were grouped and subjected to principal component analysis (56,57) in the helical space of the central d(TpA) step after Lankaš' normalization of the different rotational and translational degrees of freedom (58). The essential dynamics of the central d(TpA) step is then used to define the set of key movements explaining the global deformation at the d(TpA) step. The distributions of the four informative bps deformations were subjected to detailed analysis (see Supplementary Method

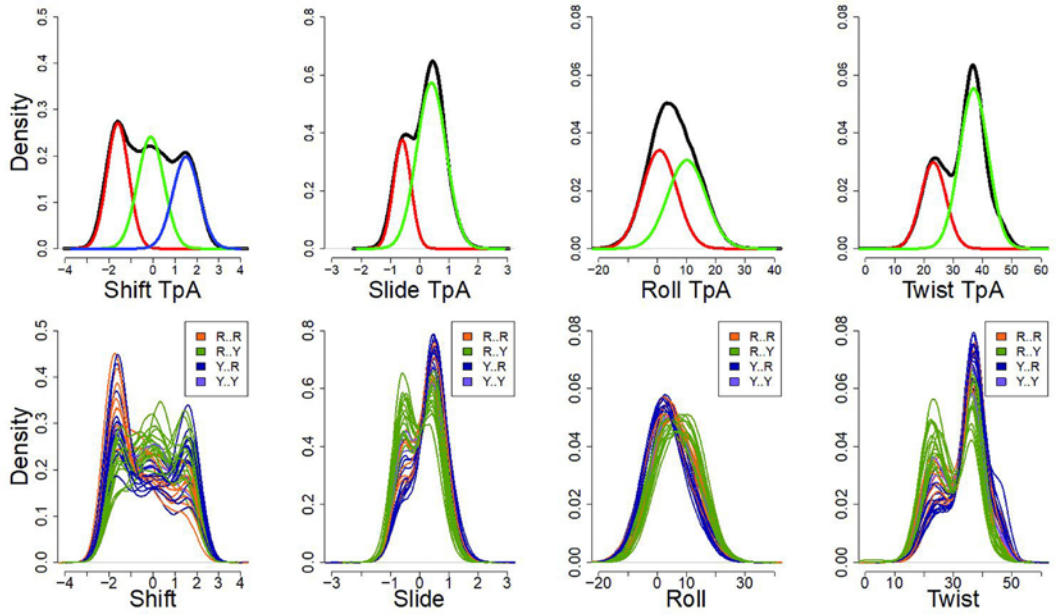


Figure 1. Normalized frequencies of those bps helical parameters found to be bi-normal and tri-normal according to the BIC analysis. First row: Density obtained from the meta-trajectory (black line), and the BIC decomposition in two Gaussians (slide, roll and twist: red and green lines) or in three Gaussians (shift: red, green and blue lines). Second row: Overlapped density of the shift, slide, roll and twist parameters at the central TpA step of the 40 sequences studied (see Supplementary Table S1).

for additional details). Comparison and clustering of the individual trajectories of the central d(TpA) for the 40 sequences studied (all with a common CTAG central tetranucleotide) were done using symmetrized Kullback-Leibler (KL) divergences (58) followed by hierarchical cluster analysis using Ward's clustering criterion (59), where the dissimilarities are squared before cluster updating (60), using as descriptive variable the six distinguished helical variables detected by the PCA of the meta-trajectory (see Supplementary Methods for additional details). The clusters obtained in this manner were subsequently analysed in detail, further highlighting the differences between their individual accessible helical spaces. Ion analysis was performed as described elsewhere (18,38) to unravel the connections between the binding of cations on the DNA and its mechanistic properties. Stacking strengths were followed by geometrical criteria for the central dinucleotide in the meta-trajectory filtered by the three main states in helical space, as described in detail in Supplementary Methods. Structural database analysis was done using all DNA structures containing the CTAG tetranucleotide. Genomic analysis was done to determine the prevalence of the CTAG tetranucleotide in different wild-type genomes and its resilience to mutation. Genomes of *Homo sapiens* (hg19), *Escherichia coli* (NC_000913.3) and *Saccharomyces cerevisiae* (sacCer3) were analysed. Occurrences of this tetranucleotide were then mapped, using Homer software (61), to the annotated regions of each organism obtained from UCSC and compared to the overall frequency of each annotation type. To compute the resilience to mutation, the frequency of mutations for each tetranucleotide along the genome in 30 different cancer types (data from (62)) was determined normalizing by tetranucleotide occurrence along the genome. Single-nucleotide polymorphisms (SNPs) in the human genome were retrieved from Ensembl Variation database (63), and the number of SNPs per tetranucleotide was computed, normalizing by genome-wide tetranucleotide frequency.

RESULTS AND DISCUSSION

The CTAG shows dramatic and complex structural polymorphism

We collected trajectories for 40 oligonucleotides containing the CTAG tetranucleotide in a central position (see 'Materials and Methods' and Supplementary Table S1). All the trajectories were stable along time in the sub-microsecond timescale, sampling structures that fit well in the B-like double helical conformation. As suggested by the analysis of ABC-simulations (37), and of trajectories deposited in BigNASim, (40) CTAG is highly polymorphic as seen from clear bimodal distributions of some helical parameters. To check that the multi-peaked distributions were not artefacts due to limited sampling, we extended trajectories for selected tetranucleotides up to 9 μs (Supplementary Table S1), tracing the changes in the distribution of helical parameters. The good convergence shown in Supplementary Figure S1 supports the robustness of our results and sug-

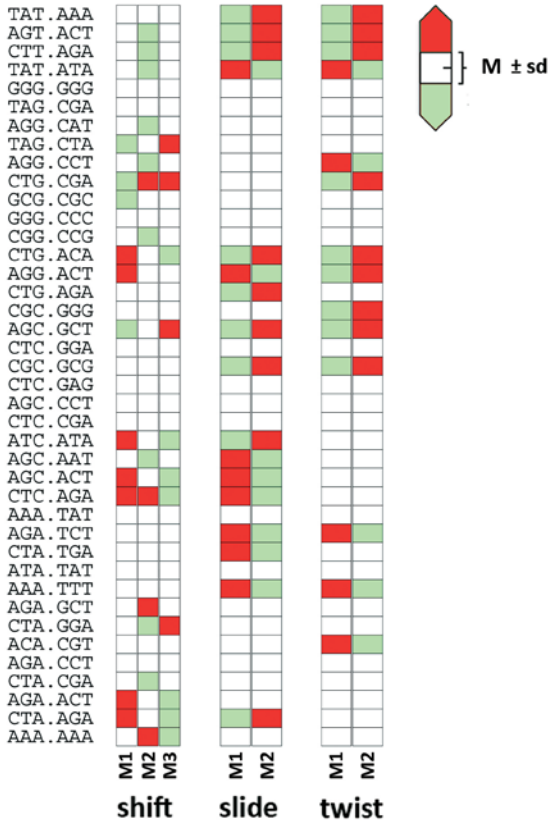


Figure 2. Relative propensities of the multi-modal bps helical coordinates of the central TpA in all 40 sequence contexts. Comparison to the global average propensities over all sequence contexts per component of the multi-modal distributions with standard deviations that reflect the variation of the propensity of each component amongst sequences. The propensity values were computed BIC analysis (see 'Materials and Methods' section and Supplementary Methods).

gests a fast dynamic of interchange of the different states (see 'Discussion' section below).

In order to obtain a global average picture of the conformational space accessible to the CTAG tetranucleotide, we joined the 40 individual trajectories (equal number of snapshots in all cases) to generate a meta-trajectory, which was then subjected to PCA and BIC analysis. Four base-parameters (the symmetric buckle and propeller twist of d(T·A) and d(A·T)) and four bps parameters at the central d(TpA) step (roll, twist, shift and slide) emerged as determinant to explain 60% of the variance in the meta-trajectory; Six of which were used for further analysis. As seen in the BIC analysis summarized in Figure 1, deviations from Gaussianity in the form of multi-peaked distributions are the main responsible for the structural polymorphisms detected at the bps level. Such deviations could in principle emerge from two different sources: (i) intrinsic multi-modality in the individual trajectories and (ii) individual distributions (coming from the 40 sequences studied) are Gaussian, but they are centred at different average values. To analyse which is the real origin of the deviation from normality in meta-trajectories, we repeated the analysis for individual trajectories (Figure 1). Roll distributions were unimodal in all cases, but the position of the peak was displaced towards slightly higher values when the central tetranucleotide is surrounded by R at 5′ and Y at 3′ (i.e. RpCpTpApGpY hexanucleotides), leading to a bi-normal distribution of the meta-trajectory (see Figure 2). The situation is completely different for twist, slide and shift where bi- or even tri-modality (three peaks in the distribution) is clear for individual sequences (see Figure 2 and Supplementary Figure S2), with the different substates being sampled in a fast equilibrium along the time scale of the simulations (see examples in Supplementary Figure S3).

As shift distribution is tri-modal and twist and slide distributions are bi-modal, we could in principle expect 12 states. However, many of the combinations of twist, slide and shift substates are not possible, and in practice, only four states appear when meta-trajectory is projected in the twist-slide-shift 3D space (Figure 3). In fact, one of them (high twist/positive slide/zero shift; HPZ) is populated only in some of the simulations and has globally a reduced impact in the meta-trajectory ensemble, which is dominated by three main states (Figure 4): high twist/positive slide/negative shift (HPN); high twist/positive slide/positive shift (HPP) and low twist/negative slide/zero shift (LNZ). Experimental validation of the suggested polymorphisms is difficult as experimental structures are always averaged (i.e. assuming a normal unimodal distribution). However, plotting the scarce experimental data available for the CTAG tetranucleotide on the 2D population plots (shift-twist, shift-slide and twist-slide) derived from meta-trajectories provides an indirect, but strong support to our results. For example, the shift distribution is very narrow and centred around zero for low slide values, while when slide increases, larger values (either positive or negative) of shift are sampled, in perfect agreement with MD meta-trajectories. Similarly, low twist appears experimentally only in zero shift conformations, while high shift (either negative or positive) is found only in experimental structures with a high twist. Although the major discrepancies between MD and experiments seem to occur for the twist-shift plane, filtering the shift values according to low/high twist reconcile partially the matching between experiments and theory (Supplementary Figure S4). Finally, the twist-slide plot shows only two regions of high probability consistent with the same slide/twist correlation found experimentally (see Figure 3 and 'Discussion' section below).

Next-to-nearest dependence in central d(TpA) conformation

All the sequences studied here correspond to the same tetranucleotide, so a similar distribution of helical parameters at the central d(TpA) step could be expected. However, this is not the case as shown in selected examples in Supplementary Figure S2, where large differences in the distributions of helical coordinates for the d(TpA) step appear. Analysis of the trajectories (Figure 1) reveals that the origin

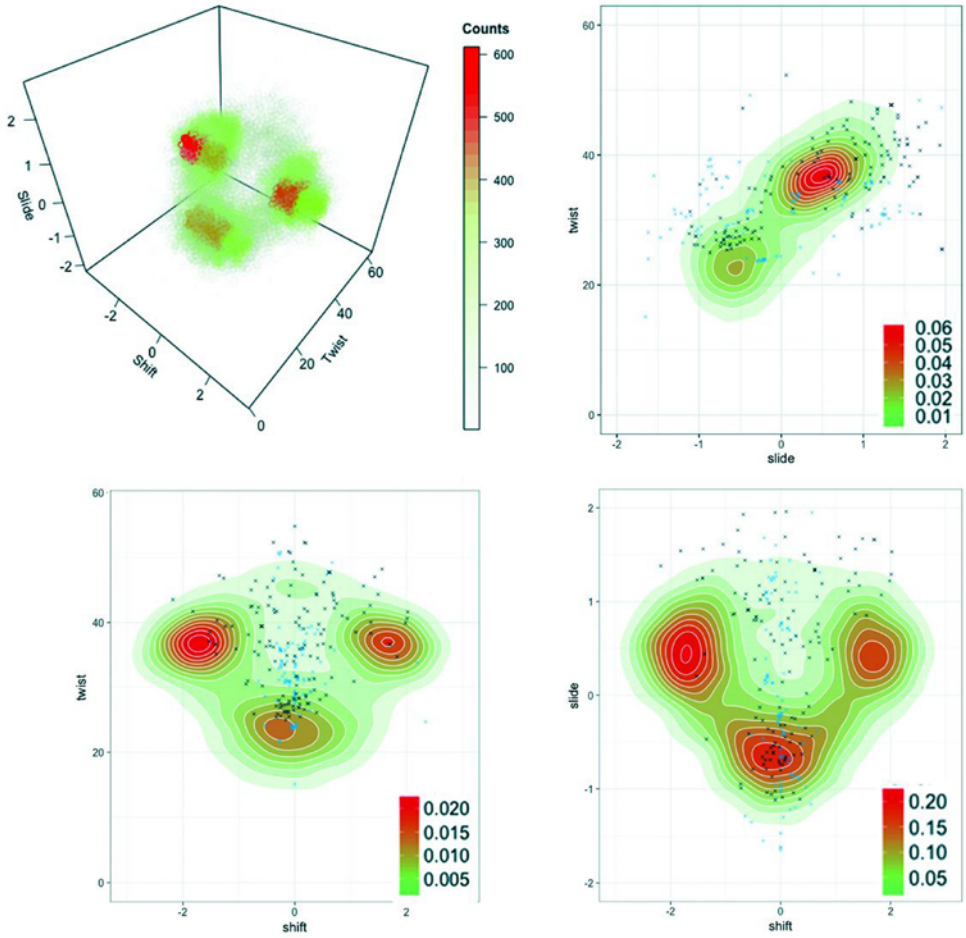


Figure 3. 3D and 2D counts in the shift, slide and twist planes from MD simulations at the central TpA step. In the 2D density plots, experimental structures from the PDB (see Supplementary Methods) were added as black crosses (protein–DNA complexes) or blue crosses (isolated DNA).

of the difference emerges from the different weights of the individual substates defining the global distributions (see a global summary in Figure 2). Moreover, we observe that the varying populations of these substates are a direct consequence of sequence context. To go deeper in the analysis of this hexanucleotide variability, we perform Kullback-Leibler (KL) analysis of the 40 trajectories in the 6D space defined from the PCA analysis as informative of the entire flexibility space of the helix (see above). Clustering analysis can be performed from the KL results to determine the similarity between sequences based on the dynamics of the central d(TpA) step and organized in the relational dendrogram (Figure 5), which clearly shows the presence of at least two major clusters. The first one is populated mainly by sequences where the central tetranucleotide is flanked by Y at 5′ and R at 3′, but also contains two 5′Y·3′Y sequences. The other cluster, the largest one, is subdivided into three different subclusters, two of which are formed almost exclusively of sequences where the central tetranucleotide is surrounded by R at 5′ and Y at 3′; finally, the last cluster corresponds to situations where the CTAG tetrad is surrounded by 5′R··3′R. Examples of prototypical distributions obtained for representative sequences in each cluster are shown in Supplementary Figure S5, which demonstrate that the hexanucleotide content has a non-negligible role in defining the properties of the central d(TpA) step in the CTAG tetranucleotide, a clear exception of the nearest neighbour model. Furthermore, the presence of some hexanucleotides in different clusters suggests that some couplings might be possible even beyond the next-to-nearest neighbour level (see below). The rules that govern the sampling of a given substate of the sequences in each cluster can be understood by analysing sequence-dependent stabilizing

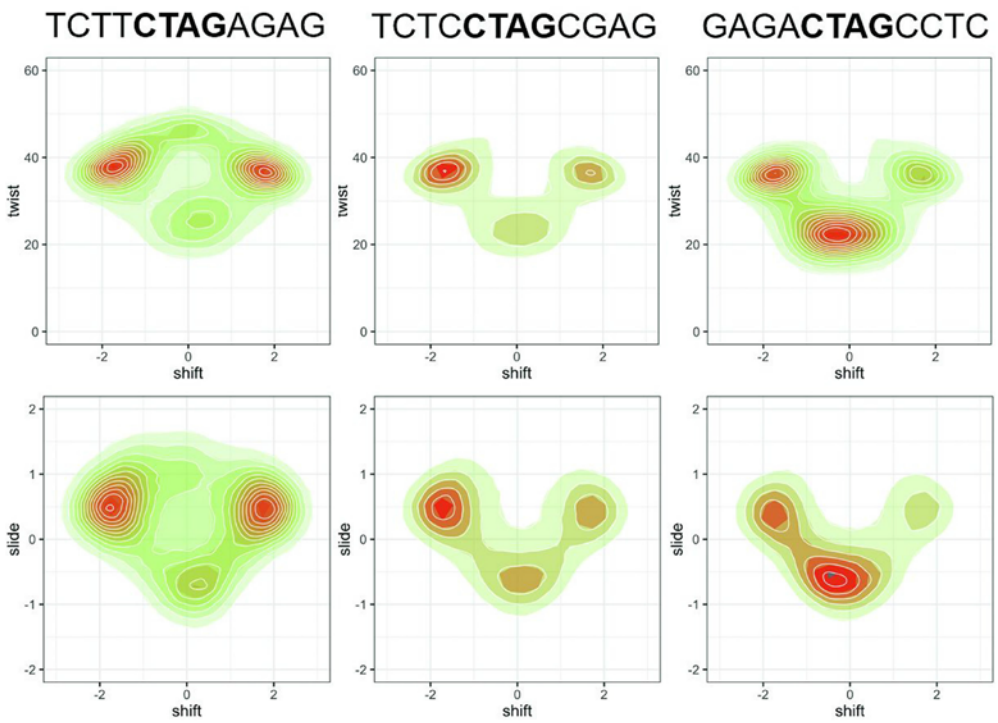


Figure 4. 2D density plots in the shift/twist and shift/slide planes at the central TpA step for three selected sequences.

factors that give rise to the characteristic distributions of helical parameters depicted in Supplementary Figure S5.

The existence of such effects implies that the motion of the central TpA step must be somehow connected to the distant base pairs. Mechanical information should travel from one site to the other to allow the TpA step to 'feel' its environment and respond in a different way according to the nature of the base pairs located almost half helical turn away. We were able to find a possible explanation based on the concerted and correlated movements of the backbone and bases, by first noting that the twist polymorphism at TpA was behaving as the better well-known YpR step: d(CpG) (18,34,37,39). The two possible twist substates (HT/LT) at the TpA step were connected to the backbone BI/BII polymorphism at the next GA junction (note that BI/BII interconversion is mainly governed by the ζ torsion). Furthermore, the BI/BII polymorphism at GpA is possible due to the formation of the intra C8H8-O3′ h-bond and the shift polymorphism in the same junction (Figure 6A and B) (39). Similar results were found if looking to the correlation of twist at the central TpA step with the bps at the 5′-side (CpT). It is then clear that the main backbone polymorphism (BI/BII) is linked to the base polymorphisms, mainly to shift and twist (Supplementary Table S2) up to the next-to-nearest neighbours. The information travels through successive backbone and base polymorphisms, which are limited to some specific substates due to DNA's crankshaft motion (Supplementary Table S2). This dynamically concerted movement of either (alone or in combination) shift/slide/twist step parameters and the ζ torsion could be appreciated from the Pearson correlation coefficients that clearly show a correlation/anti-correlation pattern in successive bps. Since intra-molecular CH-O h-bonds are mainly responsible for the information transfer between the backbone and the base (39) (with perhaps a small contribution from the known sugar puckering flexibility, see Supplementary Table S2), both backbone and base polymorphisms can be followed by looking only to the formation of those C8H8-O3′ hbonds in RpR and YpR steps, or C6H6-O3′ hbonds in RpY and YpY steps. The correlated/anti-correlated formation of these h-bonds away from the central TpA step clearly explains the transfer of mechanical information up to the next-to-nearest neighbours, and also beyond depending on the sequence (see 'Discussion' section below and Figure 6C). As a general rule, at the tetranucleotide level, the BII backbone state is significantly favoured at the 3′ side on either strand (i.e. at GpA step). The correlations of backbone substates with the helical parameters at TpA paint a picture where negative shift is related to having more BI at the GpA of the Watson strand and more BII at GpA on the Crick strand, with positive shift being favoured in the exactly opposite situation. Additionally, the TpA can be found in a low twist state only when both 3′ GpA junctions are in BII, while the simultane-

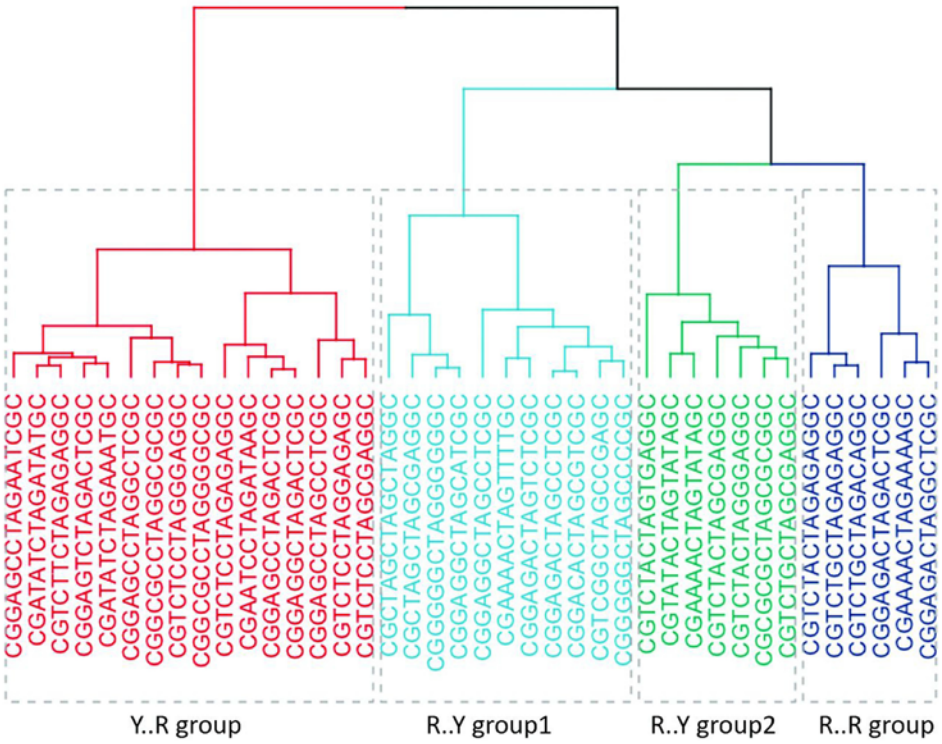


Figure 5. Dendrogram obtained from a hierarchical clustering method using Ward's criterion to classify the sequences. The distances were obtained from the symmetric Kullback-Leibler (KL) divergence in the space of six helical parameters: shift, slide and twist of TpA step, buckle and propeller of dT, and the buckle of dA (see Supplementary Methods).

ous BI state on both strands at GpA will promote high twist at TpA. The next-to-nearest context and sometimes more remote sequence effects can modulate the relative populations of BI/BII on the two strands, which in turn will affect the helical parameters at the central TpA. It's worth noting that the correlations between helical parameters in consecutive steps are mostly anti-correlations, and in general the global twist distribution of a tetra- or hexanucleotide segment can be perfectly described by a single Gaussian function. This means that, from a static and averaged view, the correlations/anti-correlations between substates in consecutive steps are leading to compensatory effects.

In addition to the backbone movements and h-bonds, each substate at the TpA step is modulated and stabilized by other factors, such as interactions with ions and stacking between consecutive bases. For CpG, a relatively simple mechanism was found where the entrance of Na^+/K^+ inside the minor groove triggered and stabilized the low twist state and hence BII (18). For TpA, the mechanism is much more complex, since it involves a combination of shift/slide/twist substates and the movements of K^+ from the major groove of CpT to the major groove of ApG, when going from HPN (high twist/positive slide/negative shift) to HPP (high twist/positive slide/positive shift) (Supplementary Figure S6). A depletion of cations inside both grooves for the whole tetranucleotide was observed when moving to the LNZ substate (low twist/negative slide/zero shift). All the sequences studied share the same redistribution of K^+ when moving between the substates, but the sequence-specific populations of each substate lead to different overall ion distributions when changing the next-to-nearest neighbour's context (Supplementary Figure S7). Finally, we found that at the TpA step, the stacking strength on either strand increased significantly when shift moves toward the minor groove at high twist and positive slide, an interaction that further stabilizes the BII state at the 3′ junction (Supplementary Figure S8).

Structural information travels beyond next-to-nearest neighbours

Sequences studied here cover all the next-to-nearest neighbours' space with some redundancy that allowed us to check for some remote effects beyond this level. As noted above, such effects are clearly visible already in Figure 5, where sequences containing the same hexanucleotide sequence appear in two very different branches of the dendrogram, indicating the tuning of hexanucleotide preferences by more remote effects. Analysis of the different octanucleotidic en-

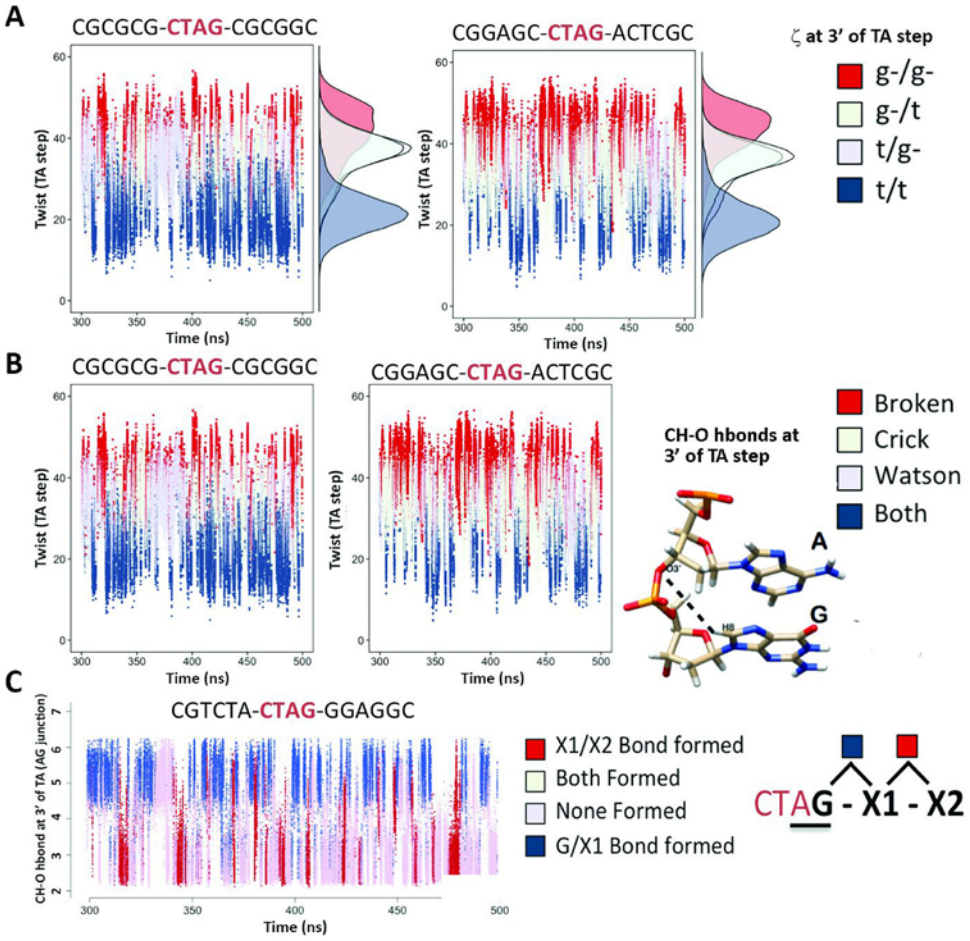


Figure 6. Concerted movements along the backbone and the bases explain the flow of structural information from the central TpA step and beyond next-to-nearest neighbours. (**A**) Correlation between twist and the BI/BII population (reduced to the ζ torsion at the 3′-side of TpA) at the TpA junction. (**B**) Correlation between twist at TpA and the CH-O h-bond formed at the ApG junction (bps +1). (**C**) Correlation between the CH-O h-bond at the ApG junction with the CH-O h-bond at bps+1 (hexanucleotide context) and bps+2 (octanucleotide context). Note that the CH-O h-bonds are always coupled to BII propensities, stabilizing the BII substate.

vironments (R··R/Y··Y), (Y··R) and (R··Y) reveals the existence of a quite differential behaviour (see Figure 7). For example, the conformational substates of the central TpA step in YpCpTpApGpR sequences (Y··R) are fully defined at the next-to-nearest neighbours level, with remote effects being negligible: all (Y··R) hexanucleotides appear in the same cluster in the dendrogram of Figure 5, and they display consistent distributions in all multi-modal helical parameters (shift has two main populations at ±2 Å, with the zero shift state being less favoured). Slide and Twist are, as a consequence, pushed towards higher values. This makes sense, considering that, irrespective of the octanucleotide level base, when ApG is followed by an R base on both strands, the junction at ApG will be pushed out of the BII state. This frustration of high BII propensity of two adjacent bps (a direct consequence of the crankshaft effect) will result in an overall higher BI population at ApG, which corresponds to the high twist, positive slide and negative/positive shift equilibrium at TpA. On the contrary, R··Y hexanucleotides (RpCpTpApGpY sequences) have two very distinct behaviours depending on the next flanking base: Central TpA steps in RpRpCpTpApGpYpY (RR··YY) octanucleotides tend to populate zero shift states and have equal populations of high/low twist as well as of negative/positive slide. On the contrary, TpA in YR··YR octanucleotide contexts have a strong preference for positive shift and rarely visit low twist or negative slide. Inspection of the trajectories suggests that this is probably due to a domino effect of h-bond proclivity so that depending on the base pairs flanking the R··Y hexanucleotide there is ei-

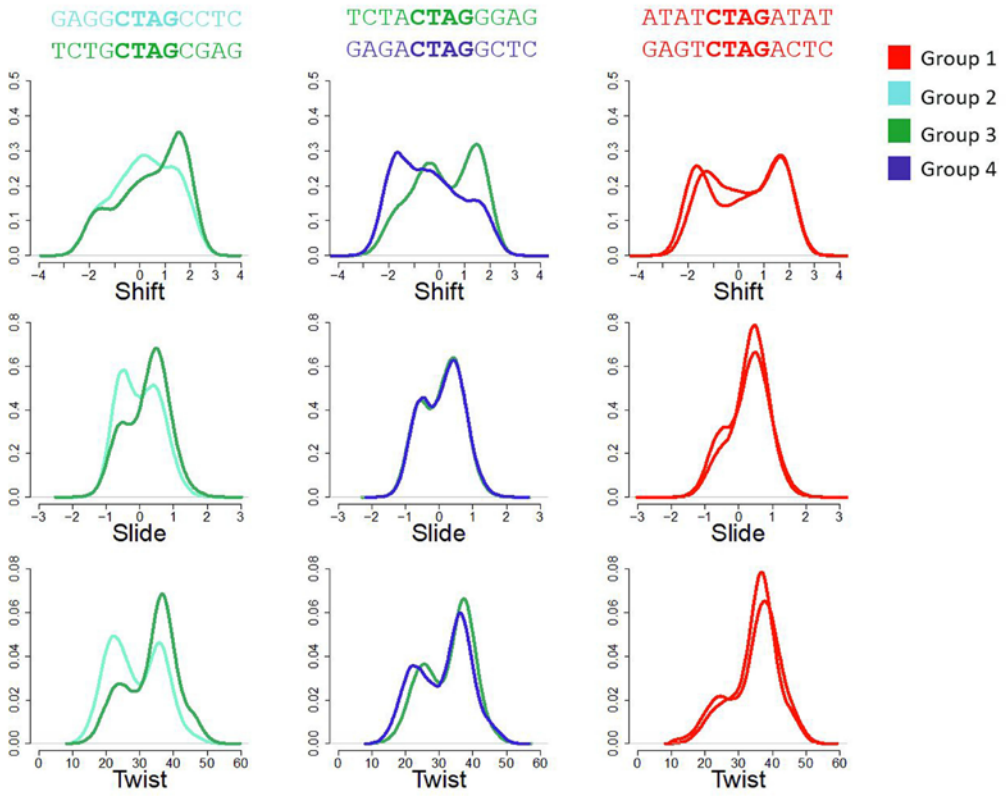


Figure 7. Normalized frequencies of shift, slide and twist at the central TpA step for three pairs of selected sequences showing non-negligible effects beyond next-to-nearest neighbours. The colours used are related to the groups found in the clustering analysis.

ther an equally strong preference towards BII at ApG on the two strands, or the Watson strand BII state is favoured over the Crick, which is necessarily compensated by shifting the bases towards the major groove. Finally, remote sequence effects are present just in a few cases for R··R/Y··Y hexanucleotides and lead to a change in shift from the minor to the major groove, maintaining similar distributions of twist and slide (Figure 7). In summary, our results suggest that CTAG is one of the few tetranucleotides (amongst the unique 136) where next-to-nearest neighbours and beyond effects are observed, while in general, nearest neighbour models can accurately explain by 'concatenation of tetranucleotides' the described remote effects in longer sequences.

Data mining of structural databases and genomic implications

We analysed the structures of DNA obtained experimentally (X-ray and NMR) and stored in the Protein Data Bank that contained the CTAG tetranucleotide sequence in order to validate our results. Only 106 occurrences of CTAG in naked DNA structures were found (some with small ligands or metal ions), and 160 occurrences in structures of protein–DNA complexes. Moreover, only a fraction of the tetranucleotide sequence space is covered (next-to-nearest neighbours), and barely any of the hexanucleotide context (octanucleotides of the type XpXpCpTpApGpXpX, where X = C, T, A, G) is found (Supplementary Table S3). This scarcity of data clearly limits the generality of the conclusions that could be derived from the data mining of the PDB, although a BIC analysis of the experimental structural parameters of TpA steps flanked by 5′C-3′G at least confirms that multi-modality is not a force field artefact (Supplementary Figure S9). PDB structures containing the CTAG tetranucleotide have values for the shift, slide, roll and twist helical parameters that cover the multi-modal distributions obtained in our trajectories, confirming our claims on the bimodal nature of slide and twist, with peaks in the distributions that fit well to our results (see Figure 8 and Supplementary Figure S9). For shift, TpA steps distribution displays peaks 2 Å towards both the minor or major groove in several protein-bound DNA structures, but the data on naked DNA seem to be insufficient to cover these deformations: there is a small peak at +2 Å, but highly underestimated compared to our results. Finally, roll has a

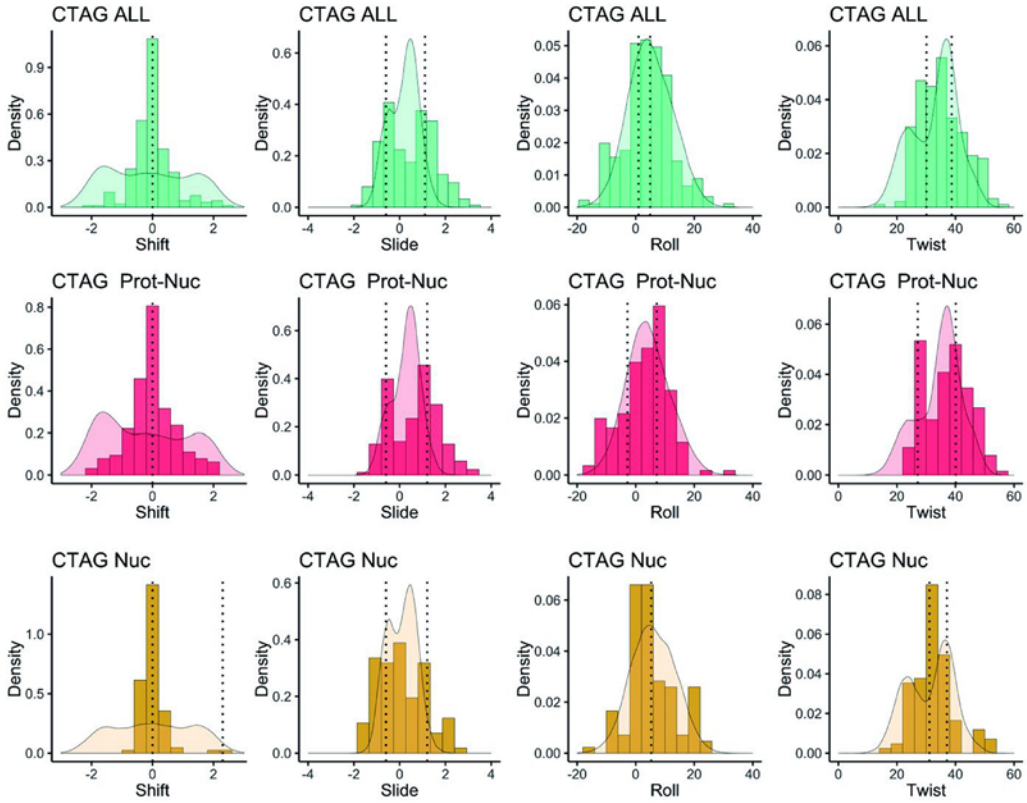


Figure 8. Normalized frequencies of shift, slide, roll and twist from MD meta-trajectory of representative hexanucleotides (G··C for free DNA; A··G, G··A, A··T and T··A for protein-bound DNA and their combination for all DNA structures) compared to those obtained from the data mining of the PDB for all structures containing DNA (first row), for Protein–DNA complexes (second row) and for isolated DNA structures (third row). The mean values of the BIC components of the experimental helical parameters data are shown as vertical dotted lines in each case.

broad distribution, similar to what we obtain from MD simulations, being bi-normal, but unimodal.

All analyses performed in this work suggests that CTAG has really unique physical properties, which should provide the genome with a point of high flexibility and polymorphism. Very remarkably, CTAG is one of the lowest populated tetranucleotides in the analysed species (see Figure 9) appearing mainly on intergenic regions and very rarely on genes (Supplementary Figure S10). We further highlighted this by analysing comparatively, with and without including exons, all the tetranucleotides containing the trinucleotide TpApG (XTAG or TAGX, where X could be A, C, G or T), which is known as the amber stop codon. Our results still confirm the low rate of the CTAG tetranucleotide, even removing the TpApG stop codon (Supplementary Figure S11). Interestingly, this infrequent CTAG tetranucleotide is well conserved, which suggest that (i) despite being far from coding regions they are important for the functionality of the cell, or alternatively, (ii) they are easily accessible to the mismatch repairing machinery, avoiding the stabilization of mutations. The same conclusion can be reached from the analysis of cancer genomic data, which show that again CTAG is very rarely mutated in cancer (Supplementary Figure S12). The unusual physical properties of the CTAG tetranucleotide matches its unusual prevalence and distribution in the genome and its extreme resilience to somatic (cancer) mutations. It is tempting to believe that cell takes advantage of the unusual properties of CTAG as points of high flexibility that might help to fold chromatin.

CONCLUSIONS

We present here an in-depth study of one of the most 'structurally speaking' polymorphic tetranucleotides found in B-DNA. The complete helical space of the CTAG tetranucleotide has been analysed by means of extensive molecular dynamics simulations and by data mining the Protein Data Bank, confirming its highly polymorphic behaviour at several helical parameters: shift, slide, twist and BI/BII. This confers to CTAG the possibility to exist in several different substates, being particularly flexible. We present here clear evidence that the type of substate displayed by CTAG in a given sequence context, and in conse-

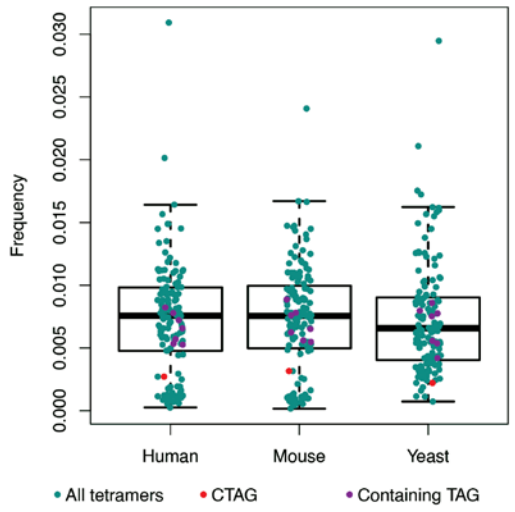


Figure 9. Frequency of each possible tetranucleotide in three different genomes. CTAG is marked in red, tetranucleotides containing TpApG (amber stop codon) are marked in violet and the rest are depicted in cyan.

quence its dynamics, is sequence dependent, and fine-tuned by next-to-nearest neighbours and beyond. Based on the concerted and correlated movements of bases and backbone torsions for the described multi-modal degrees of freedom, and driven by the mechanical limitations imposed by DNA's crankshaft motions, we were able to found a possible explanation on how structural information can travel almost half helical turn away from the central TpA step. This remote structural 'connection' allows the TpA step to 'feel' its sequence environment beyond the next-to-nearest neighbours, and eventually adopts a different substate if needed. Moreover, we found that previously described unconventional intra-molecular hydrogen bonds of the type C8H8-O3′ and C6H6-O3′ that link the movements of the bases with the torsions in the backbone, could be used as descriptors of such correlated motions. Finally, we established that although this highly flexible tetranucleotide is extremely under-represented in several genomes along the animal Kingdome, being mostly present in intergenic sequences, it has been preserved with a low rate of mutation implying a possible physical role for CTAG at the genomic level.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avançats) academia researcher. P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher.

Author contributions: The sequence library was designed by P.D.D. and A.B. Simulations were performed by A.B., with the assistance of J.W. and P.D.D. Analysis of the simulations was designed and performed by A.B., with all authors involved in assessing results and further discussions. D.B. did the genome-wide analysis, and A.H. helped with the data mining of PDB structures. M.O. and P.D.D. integrated all the results, discussed the analysis and wrote the manuscript with contributions from all the co-authors. The original idea of the project came from P.D.D. and M.O.

FUNDING

Spanish Ministry of Science [BFU2014-61670-EXP, BFU2014-52864-R]; Catalan SGR, Instituto Nacional de Bioinformática; European Research Council (ERC SimDNA); European Union's Horizon 2020 Research and Innovation Program [676556]; Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) (to M.O.); MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). Funding for open access charge: European Union's Horizon 2020 Research and Innovation Program [676556].

Conflict of interest statement. None declared.

REFERENCES

1. Wilkins,M.H.F., Stokes,A.R. and Wilson,H.R. (1953) Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature*, **171**, 738–740.
2. Franklin,R.E. and Gosling,R.G. (1953) Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–741.
3. Lucas,A.A., Lambin,P., Mairesse,R. and Mathot,M. (1999) Revealing the backbone structure of B-DNA from laser optical simulations of its X-ray diffraction diagram. *J. Chem. Educ.*, **76**, 378.
4. Kypr,J., Kejnovská,I., Renciuk,D. and Vorlícková,M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.*, **37**, 1713–1725.
5. Kato,M. (1999) Structural bistability of repetitive DNA elements featuring CA/TG dinucleotide steps and mode of evolution of satellite DNA. *Eur. J. Biochem.*, **265**, 204–209.
6. Kielkopf,C.L., Ding,S., Kuhn,P. and Rees,D.C. (2000) Conformational flexibility of B-DNA at 0.74 å resolution: d(CCAGTACTGG)2. *J. Mol. Biol.*, **296**, 787–801.
7. Maehigashi,T., Hsiao,C., Kruger Woods,K., Moulaei,T., Hud,N.V. and Dean Williams,L. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, **40**, 3714–3722.
8. Monchaud,D., Allain,C., Bertrand,H., Smargiasso,N., Rosu,F., Gabelica,V., De Cian,A., Mergny,J.-L. and Teulade-Fichou,M.-P. (2008) Ligands playing musical chairs with G-quadruplex DNA: A rapid and simple displacement assay for identifying selective G-quadruplex binders. *Biochimie.*, **90**, 1207–1223.
9. Radhakrishnan,I. and Patel,D.J. (1994) DNA Triplexes: Solution structures, hydration sites, energetics, interactions, and function. *Biochemistry*, **33**, 11405–11416.
10. Kaushik,M., Kaushik,S., Bansal,A., Saxena,S. and Kukreti,S. (2011) Structural diversity and specific recognition of four stranded G-quadruplex DNA. *Curr. Mol. Med.*, **11**, 744–769.
11. Dai,J., Carver,M. and Yang,D. (2008) Polymorphism of human telomeric quadruplex structures. *Biochimie.*, **90**, 1172–1183.
12. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
13. Dans,P.D., Danilāne,L., Ivani,I., Dršata,T., Lankaš,F., Hospital,A., Walther,J., Pujagut,R.I., Battistini,F., Gelpí,J.L. *et al.* (2016)

Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.*, **44**, 4052–4066.
14. Imeddourene,A. Ben, Xu,X., Zargarian,L., Oguey,C., Foloppe,N., Mauffret,O. and Hartmann,B. (2016) The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res.*, **44**, 3432–3447.
15. Ben Imeddourene,A., Elbahnsi,A., Guéroult,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations meet experiment to reveal new insights into DNA intrinsic mechanics. *PLOS Comput. Biol.*, **11**, e1004631.
16. Tian,Y., Kayatta,M., Shultis,K., Gonzalez,A., Mueller,L.J. and Hatcher,M.E. (2009) ^{31} P NMR investigation of backbone dynamics in DNA binding sites†. *J. Phys. Chem. B*, **113**, 2596–2603.
17. Zgarbová,M., Jurečka,P., Lankaš,F., Cheatham,T.E., Šponer,J. and Otyepka,M. (2017) Influence of BII backbone substates on DNA Twist: A unified view and comparison of simulation and experiment for all 136 distinct tetranucleotide sequences. *J. Chem. Inf. Model.*, **57**, 275–287.
18. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
19. Balaceanu,A., Pérez,A., Dans,P.D. and Orozco,M. (2018) Allosterism and signal transfer in DNA. *Nucleic Acids Res.*, **46**, 7554–7565.
20. Calladine,C.R., Drew,H.R., Luisi,B.F. and Travers,A.A. (2004) *Understanding DNA: The molecule and how it works*. Elsevier Academic Press, London and San Diego.
21. Dickerson,R.E. and Klug,A. (1983) Base sequence and helix structure variation in B and A DNA. *J. Mol. Biol.*, **166**, 419–441.
22. Fratini,A. V, Kopka,M.L., Drew,H.R. and Dickerson,R.E. (1982) Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *J. Biol. Chem.*, **257**, 14686–14707.
23. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
24. Cheatham,T.E., Cieplak,P. and Kollman,P.A. (1999) A modified version of the Cornell et al. Force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
25. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A., Orozco,M. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
26. Dršata,T. and Lankaš,F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102.
27. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V., Šponer,J. and Lankaš,F. (2013) Structure, stiffness and substates of the Dickerson-Drew dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
28. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
29. Dans,P.D., Ivani,I., Hospital,A., Portella,G., González,C. and Orozco,M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, 4217–4230.
30. Jiang,W., Phillips,J.C., Huang,L., Fajer,M., Meng,Y., Gumbart,J.C., Luo,Y., Schulten,K. and Roux,B. (2014) Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. *Comput. Phys. Commun.*, **185**, 908–916.
31. Salomon-Ferrer,R., Götz,A.W., Poole,D., Le Grand,S. and Walker,R.C. (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.
32. Lee,J., Cheng,X., Swails,J.M., Yeom,M.S., Eastman,P.K., Lemkul,J.A., Wei,S., Buckner,J., Jeong,J.C., Qi,Y. *et al.* (2016) CHARMM-GUI Input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.*, **12**, 405–413.
33. Páll,S., Abraham,M.J., Kutzner,C., Hess,B. and Lindahl,E. (2015) *Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS*. Springer, Cham, Stockholm, pp. 3–27.
34. Beveridge,D.L., Barreiro,G., Suzie Byun,K., Case,D.A., Cheatham,T.E., Dixit,S.B., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H. *et al.* (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA Oligonucleotides. I. research design and results on d(CpG) steps. *Biophys. J.*, **87**, 3799–3813.
35. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H. *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA Oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**, 3721–3740.
36. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2010) 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**, 299–313.
37. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2014) μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
38. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–2423.
39. Balaceanu,A., Pasi,M., Dans,P.D., Hospital,A., Lavery,R. and Orozco,M. (2017) The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28.
40. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M. *et al.* (2016) BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
41. Yuan,H., Quintana,J. and Dickerson,R.E. (1992) Alternative structures for alternating poly(dA-dT) tracts: the structure of the B-DNA decamer C-G-A-T-A-T-A-T-C-G. *Biochemistry*, **31**, 8009–8021.
42. Mack,D.R., Chiu,T.K. and Dickerson,R.E. (2001) Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.*, **312**, 1037–1049.
43. Case,D.A., Betz,R.M., Cerutti,D., Cheatham,T.E. III, Darden,T.A., Duke,R.E., Giese,T.J., Gohlke,H., Goetz,A.W., Homeyer,N. *et al.* (2016) AMBER 2016.
44. Berendsen,H.J.C., Grigera,J.R., Straatsma,T.P., Grigera,J.R., Straatsma,T.P., Berendsen,H., Grigera,J., Straatsma,T., Grijera,J., Berendsen,H.J.C. *et al.* (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
45. Smith,D.E. and Dang,L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
46. Dang,L.X. (1995) Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-Crown-6 Ether: A molecular dynamics study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
47. Dang,L.X. and Kollman,P.A. (1995) Free energy of association of the K+:18-Crown-6 complex in Water: A new molecular dynamics study. *J. Phys. Chem.*, **99**, 55–58.
48. Darden,T., York,D. and Pedersen,L. (1993) Particle mesh Ewald: An *N* ·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
49. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
50. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47–W55.
51. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
52. Schwarz,G. (1978) Estimating the dimension of a Model. *Ann. Stat.*, **6**, 461–464.
53. Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
54. Schilling,M.F., Watkins,A.E. and Watkins,W. (2002) Is human height bimodal? *Am. Stat.*, **56**, 223–229.

55. Ghosh,A. and Bansal,M. (2003) A glossary of DNA structures from A to Z. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **59**, 620–626.
56. Jolliffe,I.T. (1986) *Principal Component Analysis*. Springer-Verlag, NY.
57. Hotelling,H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–441.
58. Dršata,T. and Lankaš,F. (2013) Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 355–363.
59. Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
60. Murtagh,F. and Legendre,P. (2014) Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's Criterion? *J. Classif.*, **31**, 274–295.
61. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of Lineage-determining transcription factors Prime cis-Regulatory elements required for macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.
62. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
63. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

SUPPORTING INFORMATION

MODULATION OF THE HELICAL PROPERTIES OF DNA: NEXT-TO-NEAREST NEIGHBOUR EFFECTS AND BEYOND

Alexandra Balaceanu[1], Diana Buitrago[1], Jürgen Walther[1], Adam Hospital[1], Pablo D. Dans[1] and Modesto Orozco[1,2,*]

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain.
[2] Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain.

* To whom correspondence should be addressed: Prof. Modesto Orozco, Tel: +34 93 403 7155, Fax: +34 93 403 7157, Email: modesto.orozco@irbbarcelona.org.

SUPPORTING METHODS

The choice of sequences. We built a library of 40 different 16 bp oligomer sequences with a middle $d(CpTpApG)_2$ that cover the entire hexanucleotide space featuring a XpCpTpApGpX sequence pattern (X stands for any nucleotide) as well as all possible pyrimidine(Y)/purine(R) combinations at the octanucleotide level in several (>3) repeats.

System preparation and MD simulations. All the sequences were prepared with the leap program of AMBERTOOLS 16 (1) and simulated using pmemd.cuda code (2). Following the ABC protocol (3), canonical duplexes were generated using Arnott B-DNA fiber parameters (4), and solvated by a truncated octahedral box with a minimum distance of 10 Å between DNA and the closest face of the box.

Simulations were run using parmbsc1 force-field, SPC/E water model (5) and 150 mM concentration of K^+Cl^- salt using Smith/Dang parameters (6–8). Systems were optimized and equilibrated as described in our previous works, and simulated for at least 500 ns and up to 10 μs in the NPT ensemble, using Particle-Mesh Ewald

corrections (2, 9) and periodic boundary conditions. SHAKE was used to constrain bonds involving hydrogen (10), allowing 2 fs integration step. All the trajectories and the associated analysis are accessible in the BigNAsim portal: https://mmb.irbbarcelona.org/BIGNASim/.

Analysis of Molecular Dynamics trajectories. All the trajectories were processed with the *cpptraj* module of the AMBERTOOLS 16 package (1), and the NAFlex server (10) for standard analysis. DNA helical parameters and backbone torsion angles were measured and analysed with the CURVES+ and CANAL programs (11), following the standard ABC conventions (3). The CANION module from Curves+ (12) was used to determine the position of cations in curvilinear cylindrical coordinates for each snapshot of the simulations with respect to the instantaneous helical axis. We obtained and analysed the ion distribution in one- (R, D, A) and two-dimensional (RA, DA, DR) curvilinear cylindrical coordinates at the central tetranucleotide sequence. Duplexes were named following the Watson strand (*e.g.* CTAG stands for (CTAG)·(CTAG)). The letters R, Y and X stand for a purine a pyrimidine or any base respectively, while X·X and XX represent a base pair and base-pair step respectively. Base pairs flanking the CTAG were denoted using two dots to represent the central tetrad (e.g. R··Y).

The Essential Modes of generic TpA in helical space. We performed Principal Component Analysis (PCA) of the 18 intra- and inter- base-pair parameters that define all degrees of freedom of the central TpA step in a rigid-base model. Before calculating the covariance matrix in helical space, its entries had to be made dimensionally uniform, so all rotational degrees of freedom were scaled by a factor of 10.6 (13). The covariance was calculated from the joint equilibrated trajectories of all 40 sequences taken at every 100 ps. The first 3 Principal Components, which explain ~60% of the total variance, have their largest projections on a subset of 8 of the original 18 helical parameters. These 3 PCs were used to perform multidimensional clustering in the essential helical space using the mclust package of R. The clustering is performed using the optimal model according to Bayesian Information Criterion (BIC) for an expectation-minimization (EM) algorithm initialized by hierarchical clustering for parameterized Gaussian mixture models.

Distributions of helical parameters that guide specific sequence dependence. The helical parameters that showed the highest variability across trajectories of different sequences were identified using Principal Component Analysis (PCA) of the 18 intra- and inter base pair parameters that define all degrees of freedom of the central TpA step in a rigid-base model. The first 3 Principal Components, which explain ~60% of the total variance have their largest projections on a subset of 8 of the original 18 helical parameters. The Bayesian Information Criterion (BIC) (14, 15) was used, limiting the analysis to either two or three components to determine the number of normal functions needed to meaningfully represent the appearance

of possible substates in the shift, slide, roll and twist 1D distributions of the joint trajectory of all sequences. The normal distributions obtained from the BIC decomposition were compared to the distributions of the same parameters obtained after the multivariate clustering (into 3 clusters) of the first 3 PCs.

From the eight parameters identified from the PCA as accounting for the most variance, six are non-collinear in the essential helical space, namely the shift, slide and twist of TpA bps, the buckle and propeller twist of dT and the buckle of dA. The distributions of the subset of these 6 parameters were used to evaluate the similarity between central TpA steps in different oligonucleotide sequences using the Kullback-Leibler (KL) divergence theorem. For each pair of oligomers we calculated the symmetrized values of the KL divergence and then applied hierarchical cluster analysis using Ward's clustering criterion (16), where the dissimilarities are squared before cluster updating (17) in order to identify specific sequence effects on TpA helical space flexibility.

The 4-state model of TpA dynamics. The 3D and 2D distributions of these three parameters and their paired combinations, respectively, in the meta-trajectory have also been calculated and they show a clear preference of the TpA to occupy one of four states in the Shift-Slide-Twist space. In fact, the states of the 3 helical parameters that display polymorphisms are highly inter-dependent, as shown in the 2- and 3- dimensional distribution plots. The 3 most populated states in the twist-slide-shift space, when considering the entire meta-trajectory of all oligonucleotides, are: High Twist/Positive Slide/Negative Shift (HPN), High Twist/Positive Slide/Positive Shift (HPP), and Low Twist/Negative Slide/Zero Shift (LNZ). In order to capture and better understand these effects, we filtered the meta-trajectory into 3 sub-trajectories corresponding to the 3 states, removing all frames that did not belong to any of these. We compared the distribution of helical parameters beyond the next-to-nearest neighbours (octanucleotide level) in both directions ("-" sign for moving towards the 5' direction on the Watson strand and "+" sign for the 3' direction) between the 3 substate-trajectories and found significant effects in the neighbouring shift, slide and twist. We also compared up to the octanucleotide level, backbone torsions, sugar puckering, and glycosidic torsions.

Breaking down the twist, slide and shift contributions to the distal sequence effects, we calculate the Pearson's correlations of these parameters at TpA to the helical parameters at one and two levels away from TpA in each direction and the point biserial correlations to the backbone torsion (zeta – categorized in trans and gauche-), sugar pucker (categorized into South and North) and glycosidic torsion (categorized into Anti and High Anti).

Equilibrium distributions of inter base pair helical parameters at the TpA step vary beyond next-to-nearest neighbours. BIC (Bayesian Information Criterion) was used to distinguish between the normal (one Gaussian) or multi-normal (a mixture of two or more Gaussians) nature of the distributions of TpA helical parameters (14, 15).

Since for each individual trajectory, the BIC decomposition assign the same number of Gaussians (1, 2 and 3) in the respective helical parameters (roll, twist/slide and shift, respectively) and the peaks of the distributions are consistent thought the set of oligomers, we compare the propensities of each Gaussian of the individual trajectories with the total average propensity per peak, assigning them to one of three ranges: mean – sd, mean + sd and within this interval, in order to identify large deviations in population imposed by sequence.

Correlation between twist and zeta states. As previously analysed in depth for the CpG case, we found strong correlations between the twist state and the BI/BII backbone state at the 3' side of the TpA step on both Watson and Crick strands. The backbone state was defined by discretizing the zeta torsion sub-states into trans (180 ± 40 degrees – associated with a backbone in BII), gauche positive (60 ± 40 degrees – extremely infrequent) and gauche negative (300 ± 40 degrees – associated with a backbone in BI). Just like in the CpG case, a low twist state was found to usually be coupled with BII transitions at both 3' junctions.

Correlation between twist and C-H··O3' hydrogen bond. Relying on strong evidence from previous studies (18, 19) of almost perfect correlation between backbone state and the formation of base to backbone hydrogen bonds, we looked at the correlation between twist state at the TpA step and hydrogen bond formation beyond the next-to-nearest neighbours. We found, as expected, a dependency of 3' side adjacent bond formation to twist state that perfectly mirrors that of the backbone state. But we also discovered an insightful sequential anti-correlation of bond formation from one step to the next that is also highly dependent on sequence, which favours the formation of one or the other.

Stacking and Base-pairing strength. In order to estimate the strength of stacking at the TpA step we calculated a Stacking Factor based on the distance between the centres of mass of DT and DA, and the angle between the two planes of the bases, defined as (20):

$$\xi = \frac{r_M}{S(\alpha)}$$

$$S(\alpha) = e^{-\alpha} + e^{-(\alpha-\pi)^4} + 0.1e^{-(\alpha-0.5\pi)^4}$$

where r_M is the distance between the two centres of mass and α the angle between the base planes. We calculated the Stacking Factors separately for the major 3 of the 4 states in twist/slide/shift space defined above to determine the stabilizing factors of the highly preferred states.

Database Analysis of structural features. We retrieved high resolution (< 3Å) structures of double stranded DNA containing the CTAG tetrad and distinguished between the protein-bound and free DNA structures. We compared helical parameter distributions and components of BIC analysis between the database structures and out results. We paid special attention to the sequence context bias found in the database and performed the comparison to the meta-trajectory from simulations containing the same hexanucleotide environments centred at TpA.

Database Analysis of genomic properties. Prevalence of CTAG in the genomes of *H. sapiens* (hg19), *E. coli* (NC_000913.3) and *S. cerevisiae* (sacCer3) was computed, finding low occurrence compared to other tetranucleotides (less than 0.5% in the three species). Occurrences of this tetranucleotide were then mapped, using Homer software (21), to the annotated regions of each organism obtained from UCSC and compared to the overall frequency of each annotation type. CTAG is enriched at intergenic regions in *H. sapiens* and *E. coli*, but not in *S. cerevisiae* probably due to the low number of intergenic regions in this organism (less than 2.5% compared to more than 20% in the other two). To evaluate resilience to mutation, the frequency of mutations for each tetranucleotide (normalised by tetranucleotide frequency) along the genome in 30 different cancer types (22) was computed. SNPs in human genome were retrieved from Ensembl Variation database (23) and were mapped to each tetranucleotide to compute normalized SNP frequency per tetranucleotide.

SUPPORTING TABLES

Table S1. Sequence library used to study CTAG polymorphisms, number of replicas and simulation time.

Num.	Sequence	Simulation time	Num.	Sequence	Simulation time
1	CGTCGGCTAGCCGAGC	500 ns	21	CGGAGACTAGACTCGC	500 ns
2	CGTCTCCTAGGAGAGC	500 ns	22	CGGAGACTAGCCTCGC	500 ns
3	CGAAAACTAGAAAAGC	500 ns	23	CGGAGACTAGGCTCGC	6 µs
4	CGAAAACTAGTTTTGC	500 ns	24	CGGAGACTAGTCTCGC	6 µs
5	CGATATCTAGATATGC	500 ns	25	CGGAGCCTAGACTCGC	500 ns
6	CGTATACTAGTATAGC	2 x 500 ns	26	CGGAGCCTAGCCTCGC	2 x 500 ns
7	CGGGGGCTAGGGGGGC	500 ns	27	CGGAGCCTAGGCTCGC	500 ns
8	CGGGGGCTAGCCCCGC	500 ns	28	CGGAGGCTAGACTCGC	500 ns
9	CGGCGCCTAGGCGCGC	500 ns	29	CGGAGGCTAGCCTCGC	6 µs
10	CGCGCGCTAGCGCGGC	500 ns	30	CGGAGTCTAGACTCGC	2 x 500 ns
11	CGTCTACTAGAGAGGC	500 ns	31	CGCTAGCTAGCTAGGC	4 x 500 ns
12	CGTCTACTAGCGAGGC	2 x 500 ns	32	CGATATCTAGAAATGC	2 µs
13	CGTCTACTAGGGAGGC	6 µs	33	CGGAGCCTAGAATCGC	2 µs
14	CGTCTACTAGTGAGGC	2 x 500 ns	34	CGGCGCCTAGGGGCGC	2 µs
15	CGTCTCCTAGAGAGGC	2 x 500 ns	35	CGGAGGCTAGCATCGC	2 µs
16	CGTCTCCTAGCGAGGC	500 ns	36	CGAAAACTAGTATAGC	2 µs
17	CGTCTCCTAGGGAGGC	500 ns	37	CGCTAGCTAGCGAGGC	2 µs
18	CGTCTGCTAGAGAGGC	6 µs	38	CGTCTGCTAGACAGGC	2 µs
19	CGTCTGCTAGCGAGGC	9 µs	39	CGAATCCTAGATAAGC	2 µs
20	CGTCTTCTAGAGAGGC	500 ns	40	CGGACACTAGCGTCGC	2 µs

Table S2. Pearson correlation coefficients of Shift, Slide and Twist at TpA with flanking bps parameters and selected backbone torsions up to next-to-nearest neighbours.

		Shift at TA	**Slide at TA**	**Twist at TA**		**Shift at TA**	**Slide at TA**	**Twist at TA**
-2	Shift	0.06	0.002	0.025	zetaW	-0.067	-0.063	-0.123
	Slide	0.157	0.149	0.206	zetaC	-0.471	-0.286	-0.421
	Rise	-0.052	-0.022	-0.086	phaseW	-0.130	-0.023	-0.073
	Tilt	0.086	0.031	0.051	phaseC	-0.061	-0.079	-0.110
	Roll	0.001	0.043	0.038	chiW	0.018	0.002	0.025
	Twist	0.089	0.051	0.021	chiC	-0.074	-0.042	-0.057
-1	Shift	-0.607	-0.149	-0.257	zetaW	-0.454	-0.098	-0.217
	Slide	-0.298	0.089	-0.094	zetaC	0.753	0.295	0.536
	Rise	0.028	-0.089	-0.109	phaseW	-0.425	0.006	-0.105
	Tilt	-0.12	0.057	-0.11	phaseC	0.111	0.102	0.090
	Roll	0.002	0.178	0.157	chiW	-0.140	-0.027	-0.058
	Twist	-0.223	-0.263	-0.453	chiC	0.107	0.173	0.153
Central TpA step								
+1	Shift	-0.607	0.192	0.306	zetaW	-0.736	0.340	0.589
	Slide	0.201	0.098	-0.078	zetaC	0.456	-0.166	-0.260
	Rise	0.017	-0.08	-0.114	phaseW	-0.157	0.130	0.103
	Tilt	-0.104	-0.047	0.12	phaseC	0.431	-0.045	-0.144
	Roll	-0.045	0.176	0.173	chiW	-0.206	0.186	0.170
	Twist	0.232	-0.25	-0.455	chiC	0.166	-0.022	-0.053
+2	Shift	0.185	-0.084	-0.148	zetaW	0.547	-0.332	-0.487
	Slide	-0.251	0.195	0.271	zetaC	0.023	-0.023	-0.061
	Rise	0.09	-0.04	-0.103	phaseW	0.020	-0.072	-0.076
	Tilt	0.156	-0.091	-0.125	PhaseC	0.085	-0.004	-0.054
	Roll	0.012	0.044	0.039	chiW	0.019	-0.012	-0.018
	twist	-0.095	0.079	0.067	chiC	-0.067	0.006	0.024

Table S3. Number and frequency of unique occurrences of hexanucleotides containing central CTAG in the PDB database.

Type	Hexanucleotide Context	No. structures	Frequency
Naked DNA structures	G··C	15	0.54
	A··T	5	0.18
	T··A	3	0.11
	C··G	2	0.07
	T··T	2	0.07
	T··C	1	0.04
Protein-DNA complexes	A··G	30	0.31
	G··A	30	0.31
	T··A	11	0.11
	A··T	8	0.08
	G··G	7	0.07
	C··G	5	0.05
	A··A	2	0.02
	G··C	2	0.02
	A··C	1	0.01
	T··G	1	0.01

SUPPORTING FIGS.

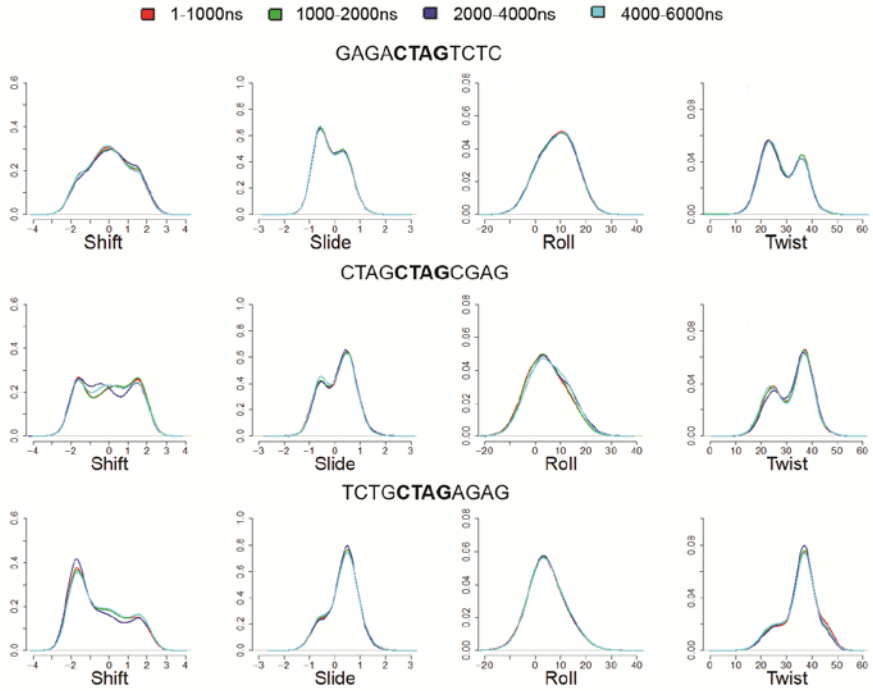


Fig. S1. Normalized frequencies of the shift, slide, roll and twist helical parameters for 3 selected sequences, whose trajectories were extended to 6 μs to check for convergence. Four distributions were computed for each helical parameter using segments of 1,000 or 2,000 ns.

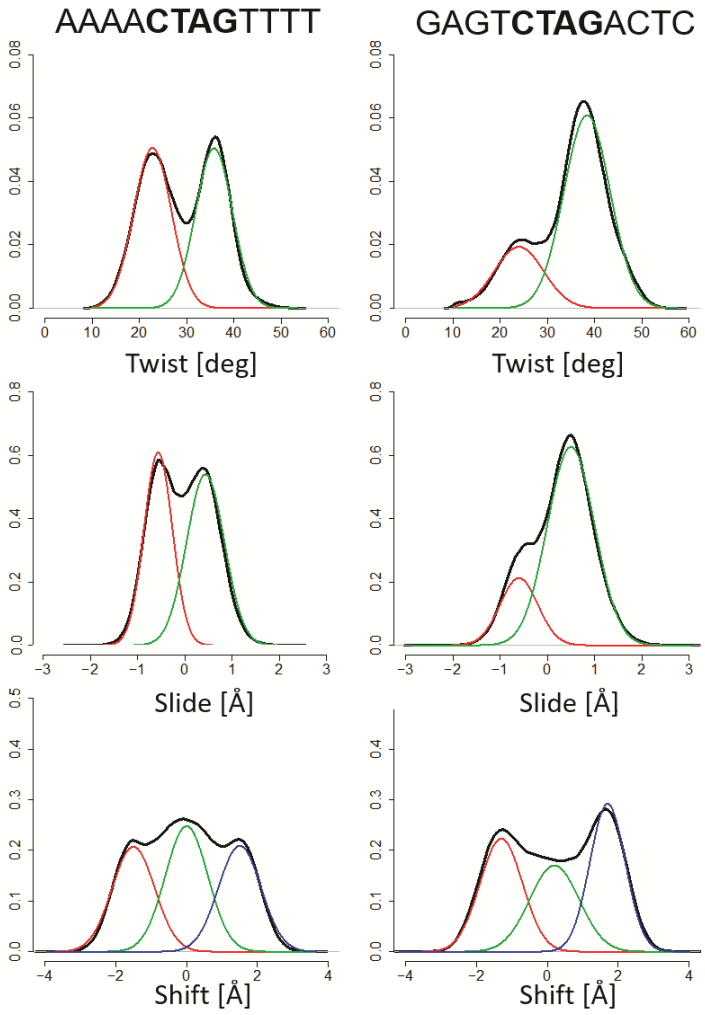


Fig. S2. Normalised frequencies of the shift, slide, and twist helical parameters for 2 selected sequences showing clear next-to-nearest neighbour effects, which could be appreciated from the change in the relative populations of the bi- and tri-normal distributions.

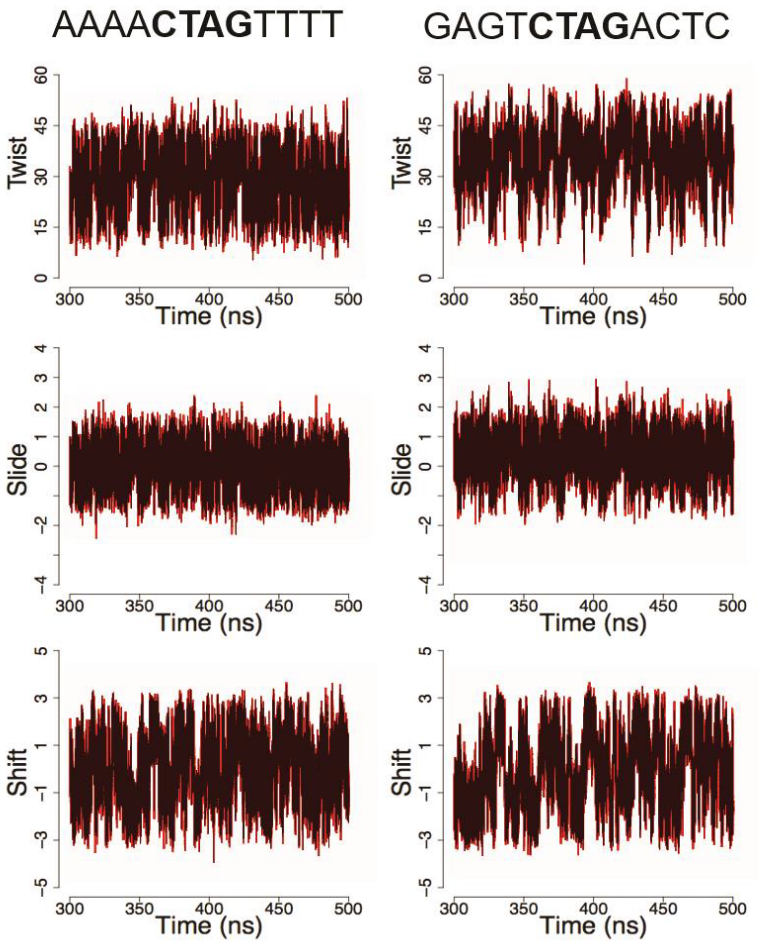


Fig. S3. Time evolution (500 ns) of shift, slide and twist for two selected sequences, showing the fast and reversible inter-conversion between high and low substates.

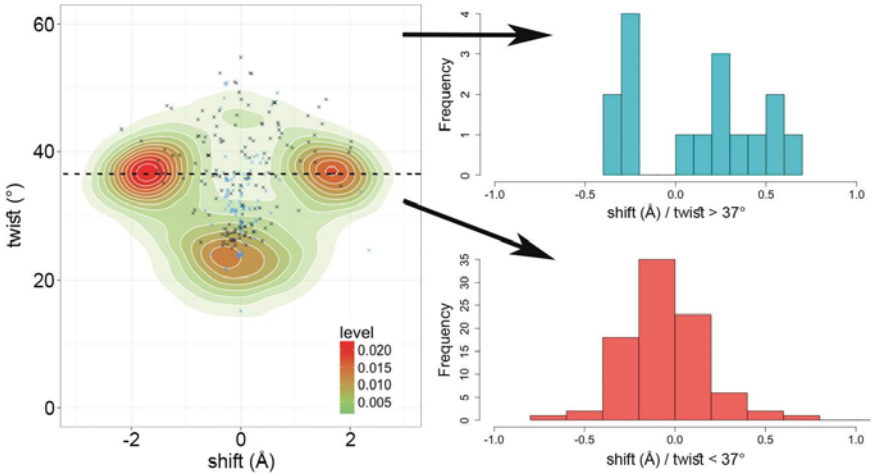


Fig. S4. 2D counts in the shift-twist plane from MD simulations at the central TpA step. In the 2D density plots experimental structures from the PDB (see Supp. Methods) were added as black crosses (Protein-DNA complexes), or blue crosses (isolated DNA). We divided the plane between high twist (> 37°), and low twist (< 37°) and analysed the shift distribution for these two cases.

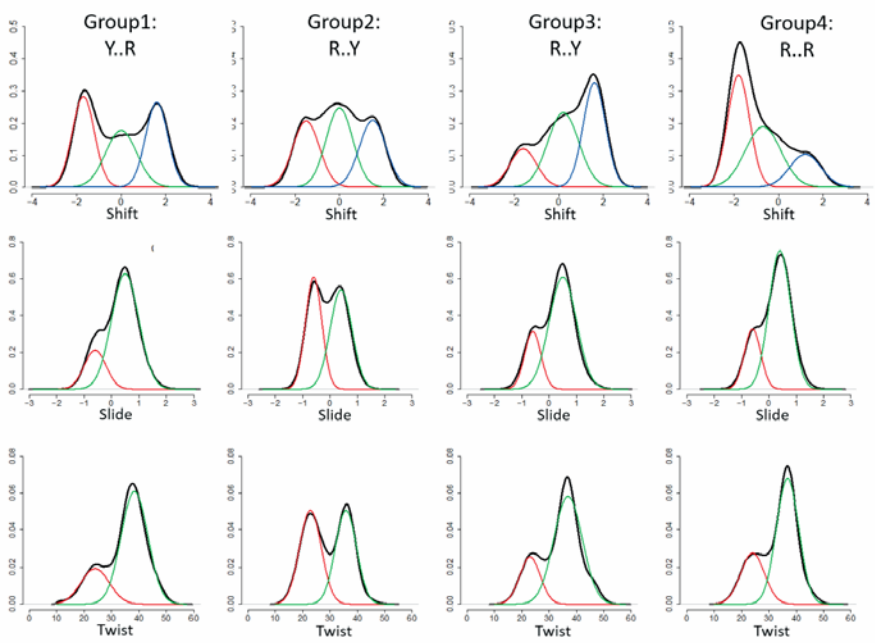


Fig. S5. Normalized frequencies for shift, slide and twist (black line), and the BIC decomposition in Gaussians (red, green, and blue lines), showing the behaviour of the clusters obtained in the dendrogram of Fig. 5.

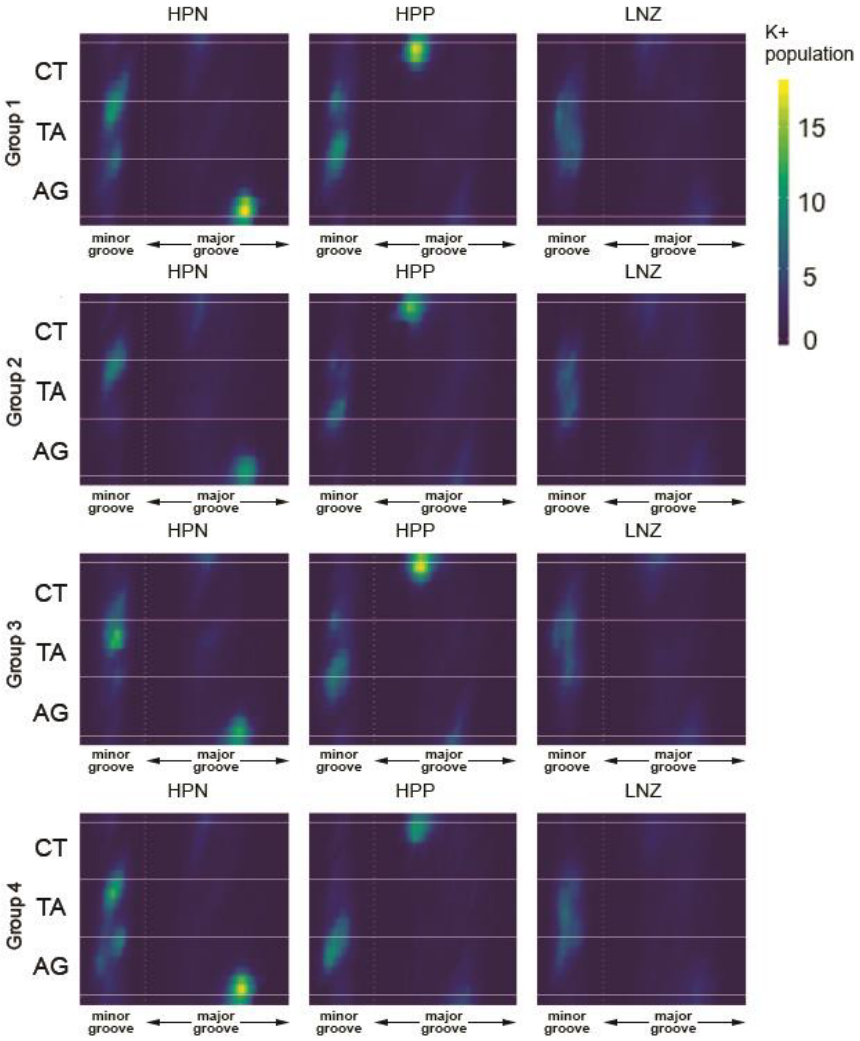


Fig. S6. Population of K+ ions inside the major and minor groove for bps CpT, TpA, and ApG in each of the three major states based on twist/slide/shift values at TpA.

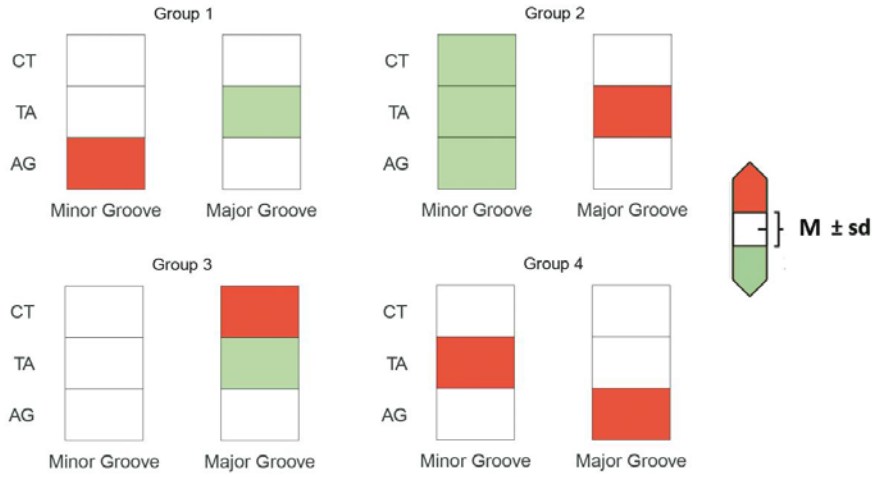


Fig. S7. Relative ion populations of cluster representatives in the minor and major groove at the CTAG tetranucleotide. Comparison to the global average ion populations per region.

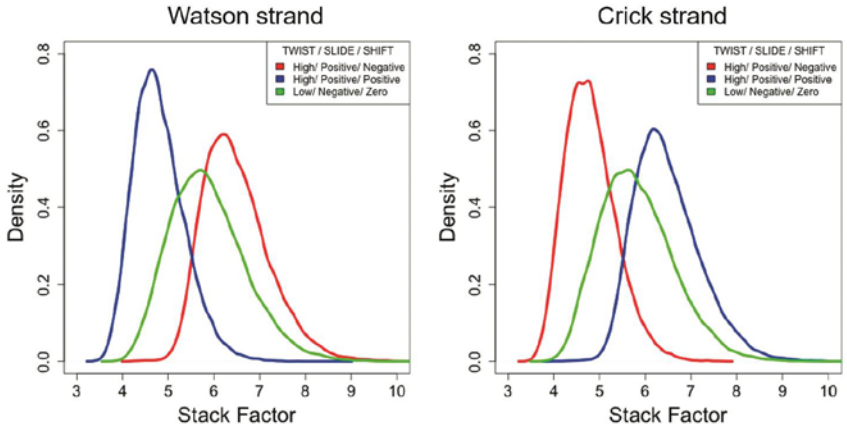


Fig. S8. Distributions of stacking coordinate at the TpA step on both Watson (left) and Crick (right) strands in the three main configurations of the bps in helical space.

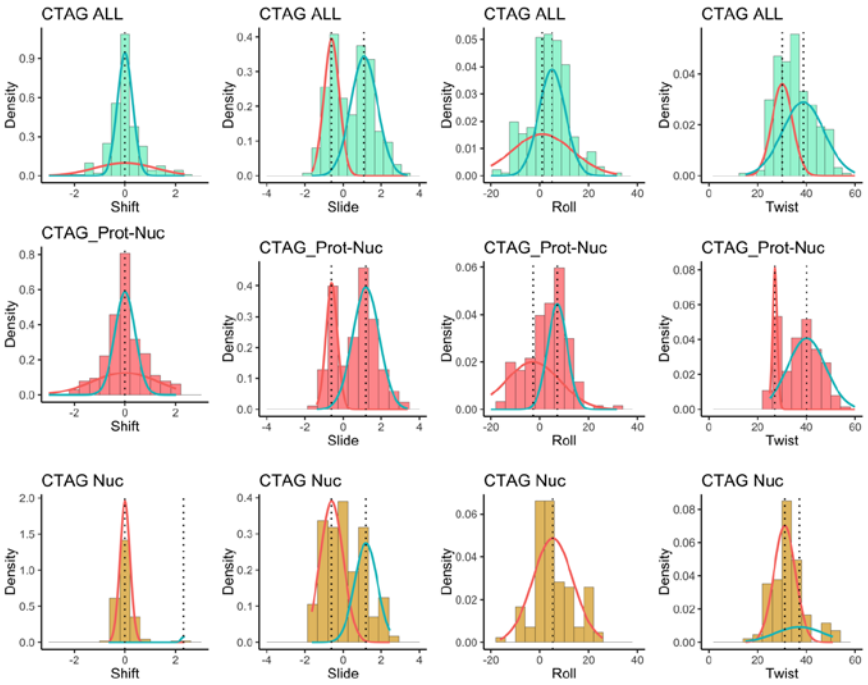


Fig. S9. Normalized frequencies of shift, slide, roll and twist at TpA obtained from the data mining of the PDB for all structures containing CTAG according to BIC analysis: all DNA (FIRST ROW), Protein-DNA complexes (SECOND ROW), and isolated DNA structures (THIRD ROW).

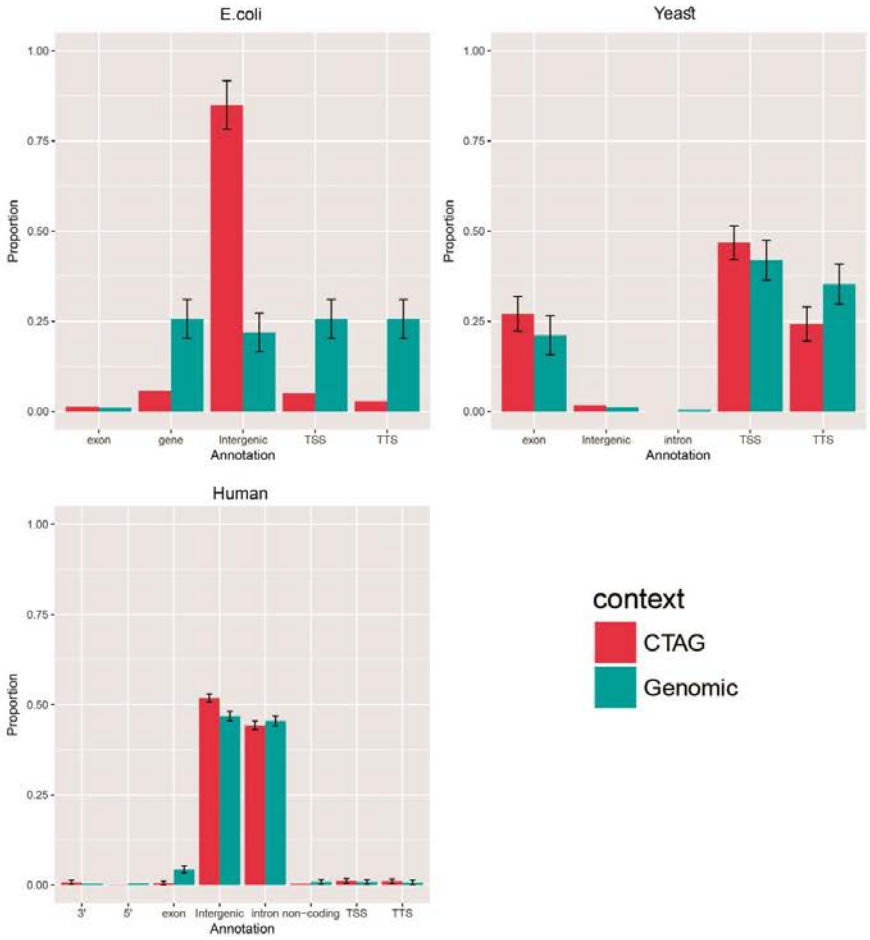


Fig. S10. Occurrence of CTAG in different genomic regions. Length of each annotation type is shown to evaluate significance of enrichment per region type.

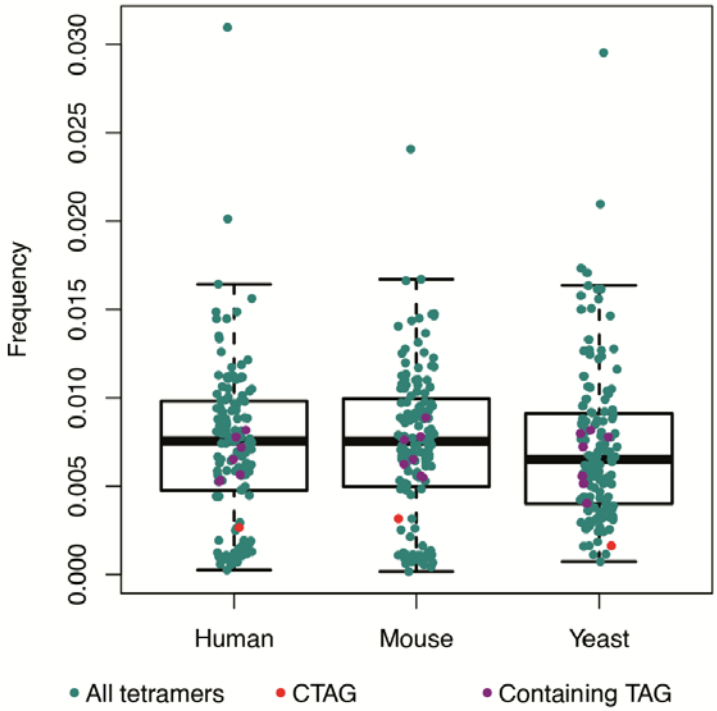


Fig. S11. Frequency of each possible tetranucleotide in 3 different genomes. CTAG is marked in red, tetranucleotides containing TpApG (all but the amber stop codon) are marked in violet, and the rest are depicted in cyan. Note that this analysis doesn't includes exons.

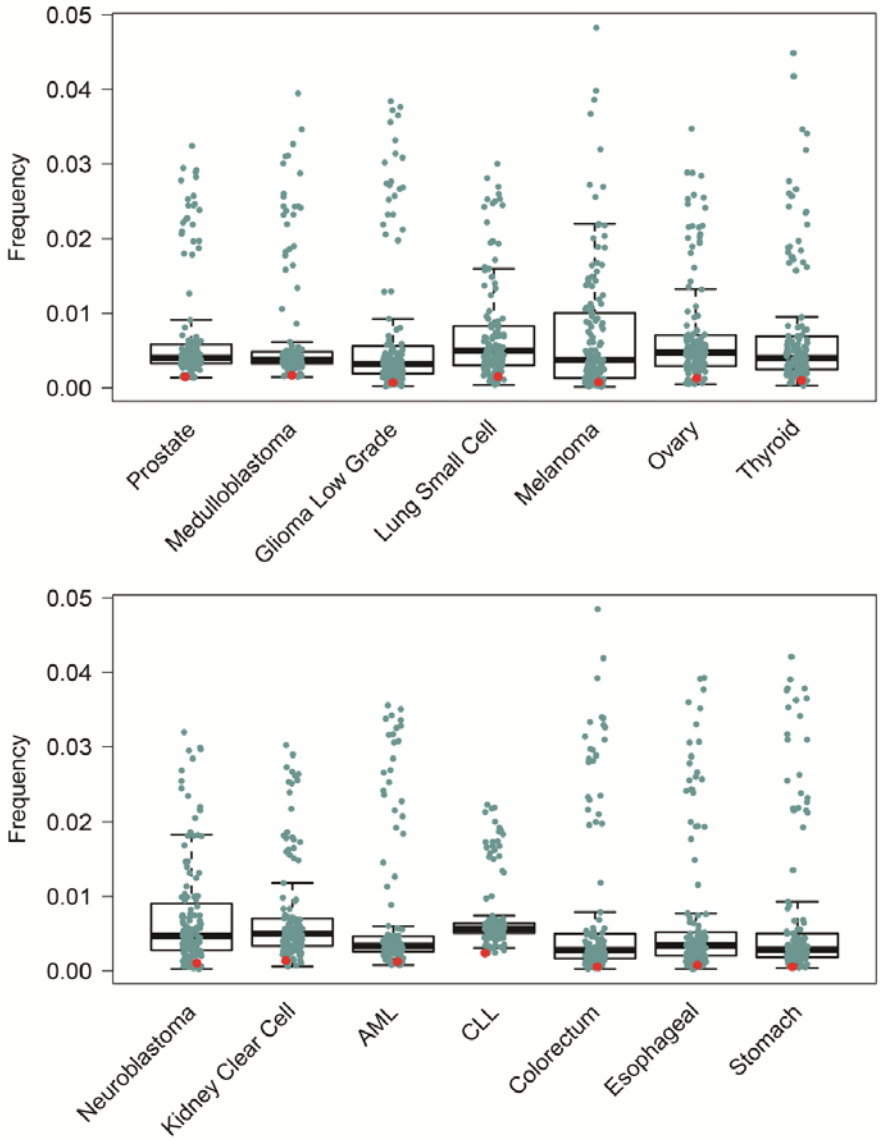


Fig. S12. Frequency of mutations for each tetranucleotide along the genome for several cancer types, normalised by genome-wide tetranucleotide occurrence. CTAG is marked in red.

SUPPORTING REFERENCES

1. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein,L.X. and P.A.K. (2016) AMBER 2016.
2. Le Grand,S., Götz,A.W. and Walker,R.C. (2013) SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.*, **184**, 374–380.
3. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
4. Arnott,S. and Hukins,D.W.L. (1973) Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. *J. Mol. Biol.*, **81**, 93–105.
5. Berendsen,H.J.C., Grigera,J.R., Straatsma,T.P., Grigera,J.R., Straatsma,T.P., Berendsen,H., Grigera,J., Straatsma,T., Grijera,J., Berendsen,H.J.C., *et al.* (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
6. Smith,D.E. and Dang,L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
7. Dang,L.X. (1995) Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
8. Dang,L.X. and Kollman,P.A. (1995) Free Energy of Association of the K+:18-Crown-6 Complex in Water: A New Molecular Dynamics Study. *J. Phys. Chem.*, **99**, 55–58.
9. Darden,T., York,D. and Pedersen,L. (1993) Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
10. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
11. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
12. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–23.
13. Maddocks,T. and Lankaš,F. (2013) Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 355–363.
14. Schwarz,G. (1978) Estimating the Dimension of a Model. *Ann. Stat.*, **6**, 461–464.
15. Kass,R.E. and Raftery,A.E. (1995) Bayes Factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
16. Ward,J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *J. Am.*

Stat. Assoc., **58**, 236–244.

17. Murtagh,F. and Legendre,P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.*, **31**, 274–295.
18. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
19. Balaceanu,A., Pasi,M., Dans,P.D., Hospital,A., Lavery,R. and Orozco,M. (2017) The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**.
20. Jafilan,S., Klein,L., Hyun,C. and Florián,J. (2012) Intramolecular Base Stacking of Dinucleoside Monophosphate Anions in Aqueous Solution. *J. Phys. Chem. B*, **116**, 3613–3618.
21. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.
22. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A. V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
23. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

2. A helical coarse grain model of B-DNA dynamics and its web implementation

Based on the acquired knowledge on DNA dynamics we decided to develop a helical CG model implementing it in a stand-alone and web application which is freely distributed. The development of a helical coarse grain model involves different choices: the level of resolution, the type of Hamiltonian to sample the conformations and the parameters on which the calculations are based. We chose as level of resolution the base pair as smallest unit as in this case movements at the base pair step level are limited to three translations and three rotations (shift, slide, rise, tilt, roll, twist), which at the expense of some loss of resolution, drastically simplifies the calculation and the parameterization of the model (10, 11). The choice of the energy function to sample the helical states was until now a harmonic Hamiltonian assuming that under normal conditions the distributions of inter base pair coordinates are Gaussian. Within this assumption the energy of the DNA can be easily described by means of a stiffness matrix and a deformation vector indicating the deviation of an inter base pair coordinate from its equilibrium value (see Section 2 in Chapter II). However, we could conclude from the previous study of the whole tetranucleotide space, many of the inter base pair parameter distributions are not behaving Gaussian (80% of the inter base pair distributions of all tetranucleotides). For this reason, we implemented a new Hamiltonian inspired by empirical valence bond theory (12), which considers explicitly the different conformational substates of the tetranucleotides. The parametrization of the conformational substates for each tetranucleotide was done in a systematic manner by dividing the MD trajectory into corresponding structures belonging to a certain substate using machine learning approaches. The sampling of the extended nearest neighbor helical CG model via a Monte Carlo algorithm resulted in structure ensembles showing very similar global and local sequence-dependent dynamics as in MD. Comparison to experimental structures from PDB yields a good agreement in RMSd/bp and inter base pair parameters and the high computational efficiency of the CG helical DNA model allows the treatment of DNA segments at time scales up to five orders of magnitude faster than conventional atomistic MD and offers simulations of long DNA stretches at unprecedented detail not reachable by atomistic MD. The algorithm is implemented in a simple web interface (http://mmb.irbbarcelona.org/MCDNAlite/) and as a

stand-alone package (http://mmb.irbbarcelona.org/MCDNAlite/standalone) enabling easy access for potential users to the model.

We developed the web interface even further to allow direct online sampling and subsequent analysis of DNA equilibrium conformations via the extended nearest neighbor CG model. In this webserver, the user is given the possibility to simulate – apart from unrestrained B-DNA dynamics – DNA in a constrained environment such as supercoiled DNA or DNA coated with proteins. The trajectories (at base pair resolution or using a pseudo-atomistic reconstruction) can be downloaded and/or subjected to a large variety of analysis in the server. The server is accessible at http://mmb.irbbarcelona.org/MCDNA/.

2.1 Extended nearest neighbor helical coarse grain model (Publication 3)

This work shows the development of a helical CG model (MC-eNN from now on) that produces results comparable to those of atomistic MD, but at a fraction of computational cost (see Figure 25 for the workflow).

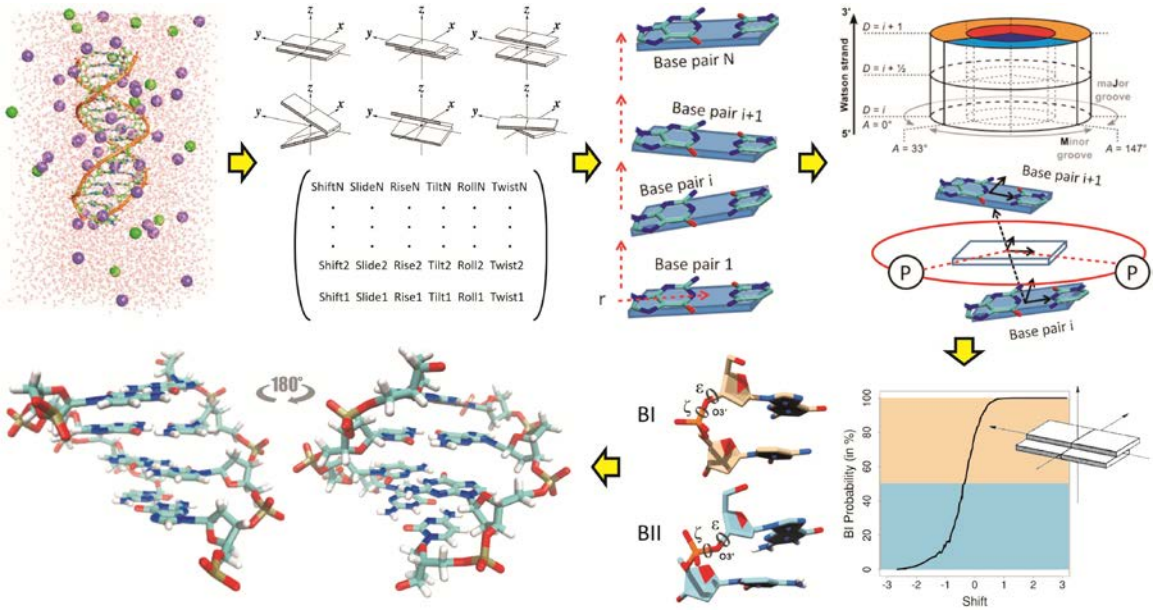


Figure 25. Workflow of the MC-eNN helical CG model.

The analysis of all unique tetranucleotides showed different helical states for several tetranucleotides, a phenomenon not able to capture by a standard harmonic model which

assumes all inter base pair parameter distributions to be Gaussian with only one single helical state. The implementation of a new Hamiltonian to correctly sample the inter base pair conformational substates makes use of parameters derived from a previous MD study of all unique tetranucleotides using parmbsc1 force field. The energy function of the extended nearest neighbor model is motivated by valence bond theory, converging to the standard harmonic model in the limit of a single helical state. The different helical substates were obtained by deconvoluting the inter base pair parameter distributions of each tetranucleotide into several harmonic distributions by means of machine learning approaches such as PCA in helical space and unsupervised clustering. We found that most tetranucleotides can be represented by 3-5 helical states (Figure 3 in the following publication) and in less than 10% of all tetranucleotides more than 5 helical states are needed. The extended nearest neighbor model is coupled to a Metropolis Monte Carlo sampling algorithm in the inter base pair parameter space where sampled configurations (examples see Figure 26) can be shown in an all-atom representation and backbone torsions were reconstituted using the correlations between inter base pair coordinates and backbone states (BI or BII) found in the previous study of the analysis of all tetranucleotides (Figure 22).

To test the CG model we compare the sampled conformations (in atomistic detail) against atomistic MD trajectories of different sequence length and context. An analysis of 10 different 18mers reveals similarity indices higher than 0.8 when comparing the MC-eNN method versus MD (see Figure 5 in the following publication), much higher than the cross-similarity indices. Local inter base pair distributions obtained from MC-eNN calculations are impossible to differentiate from those derived from atomistic MD simulations (see Figure 6 in the following publication), even in the cases where the inter base pair parameters are correlated in a highly non-linear manner. We tested the MC-eNN performance against the longest naked DNA duplex in the BigNASim database of 56 base pairs in length. Apart from high similarity in essential dynamics (a Boltzmann-weighted similarity index close to 90%) and very similar average RMSd/bp (0.09 Å x bp for MD and 0.11 Å x bp for MC-eNN), groove dimensions and many other subtle structural details such as the sequence-dependent backbone substate population are well reproduced (see Figure 27).

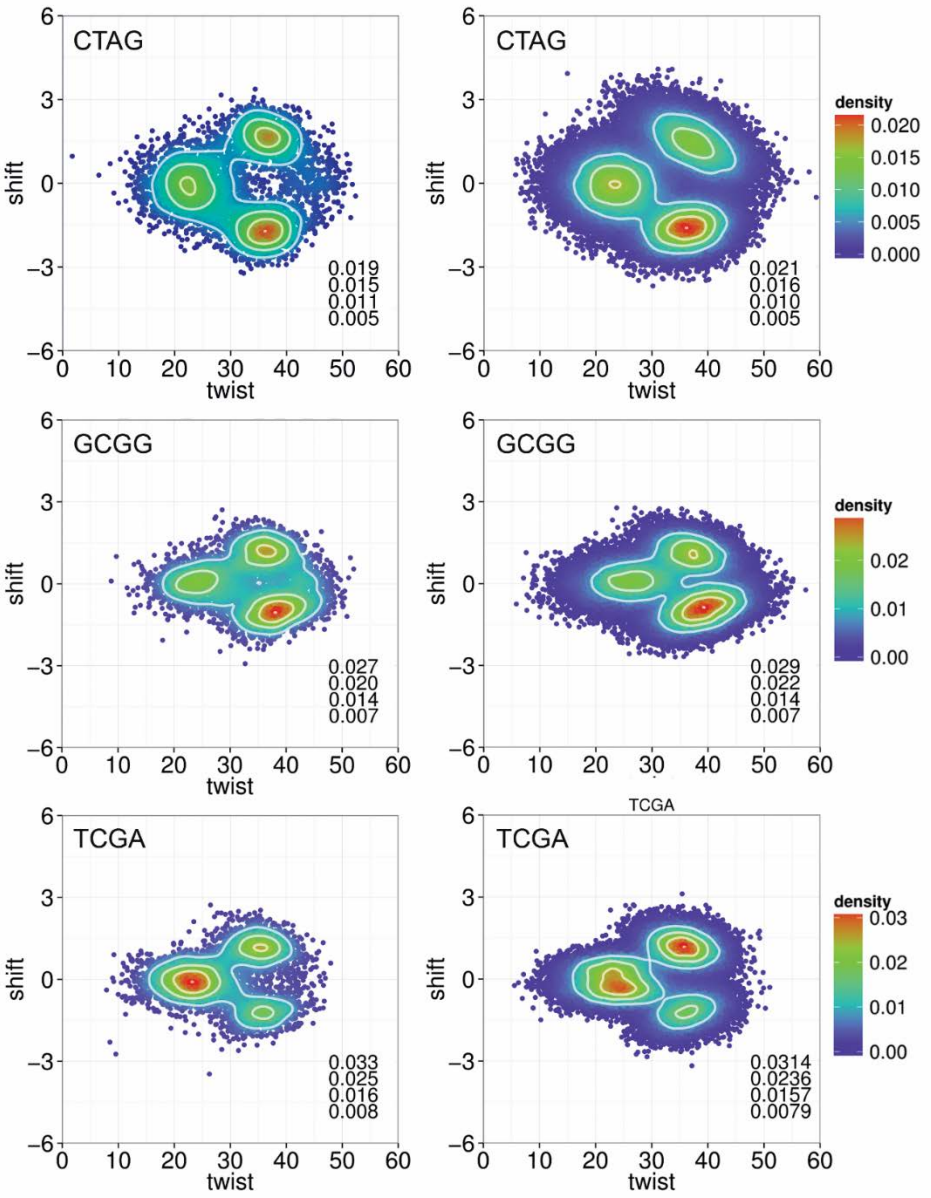


Figure 26. Bi-dimensional inter base pair parameter maps of Twist-Shift of three tetramers CTAG (top), GCGG (middle) and TCGA (bottom) of MD simulations of the parmbsc1-ABC data set (left) and MC-eNN simulations (right).

We performed an exhaustive comparison of MC-eNN configurations with highly-resolved experimental structures (X-ray or NMR) from the PDB data base (see Supplementary Table 5 in the following publication). We find good agreement in terms of the average inter base pair parameters and an average RMSd of around 0.3 Å x bp is very close to those found in atomistic MD trajectories.

The performance of the MC-eNN algorithm is such that to obtain converged inter base pair distributions atomistic MD simulation of a 56-mer duplex would require more than 500 days in a 64-core cluster while to obtain equivalent sampling with MC-eNN would only require 12 minutes outperforming MD by a factor of $\sim 10^5$ (see Supplementary Figure S12 in the following publication).

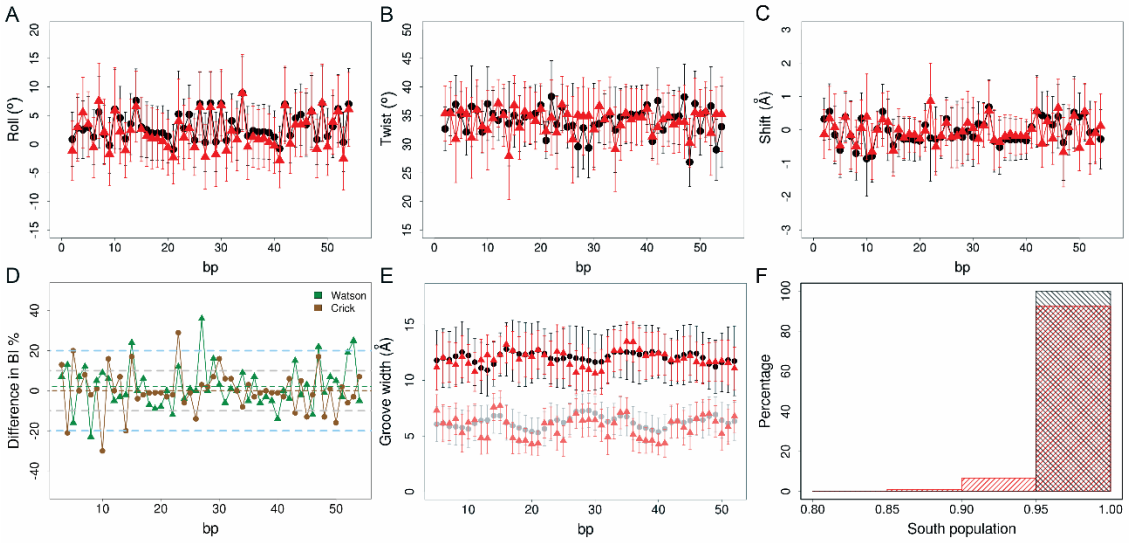


Figure 27. Comparison of MC-eNN (black) and MD simulations (red) of the longest naked DNA duplex in the BigNASim database (56 bp in length, sequence see Suppl. Table S4 in the following publication).

In summary, the new mesoscopic model for the representation of structure and dynamics of naked DNA structures allows an accurate representation of complex polymorphisms in DNA while maintaining its mathematical elegant description and computational efficiency. It is implemented in simple tools that can be used by non-experts (http://mmb.irbbarcelona.org/MCDNAlite for the web implementation and http://mmb.irbbarcelona.org/MCDNAlite/standalone for the stand-alone version) aiming to serve as an easy-to-use model to obtain a more complete picture of DNA structures.

Publication:

Jürgen Walther, Pablo D. Dans, Alexandra Balaceanu, Adam Hospital, Genís Bayarri and Modesto Orozco; A multi-modal coarse-grain model of DNA flexibility mappable to the atomistic level, (submitted)

A MULTI-MODAL COARSE-GRAIN MODEL OF DNA FLEXIBILITY MAPPABLE TO THE ATOMISTIC LEVEL

Jürgen Walther[1], Pablo D. Dans[1], Alexandra Balaceanu[1], Adam Hospital[1], Genís Bayarri[1] and Modesto Orozco[1,2]*

We present a new coarse-grained method for the simulation of duplex DNA. The algorithm uses a generalized multi-harmonic model that can represent any multi-normal distribution of helical parameters, avoiding thus caveats of current mesoscopic models of DNA simulation. The method has been parameterized from accurate parmbsc1 molecular dynamics simulations of all unique 4-mer sequences of DNA embedded in long duplexes and takes advantage of the correlation between helical states and backbone configurations to derive atomistic representations of DNA. The algorithm, which is implemented in a simple web interface and in a standalone package reproduces with a high computational efficiency the structural landscape of long segments of DNA untreatable by atomistic molecular dynamics simulations.

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology.
[2] Department of Biochemistry and Biomedicine. The University of Barcelona. Barcelona, Spain.

* Correspondence to Prof. Modesto Orozco: modesto.orozco@irbbarcelona.org

INTRODUCTION

Under physiological conditions DNA behaves like a polymeric entity whose properties are dependent on the underlying sequence. Experimental approaches to the determination of sequence-dependent physical properties of DNA are impaired by their inability to deal with long and flexible polymers, which has fueled the development of theoretical simulation techniques (1), among them atomistic molecular dynamics (MD), a method that after recent improvements in force-fields (2, 3) has shown extreme accuracy in describing the structural and dynamic properties of a variety of DNA structures (4–10). Unfortunately, the computational cost of MD simulation scales (roughly) with the 3rd power of the length of the duplex, and a simple 100 bp duplex would require a simulation box containing more than 10^7 water molecules, a system for which reaching reasonable simulation times is nearly impossible.

Coarse grain (CG) methods are a cost-effective alternative to simulate very long segments of DNA, approaching the chromatin scale. Summarizing, two families of CG methods have been developed (1, 11–14): the first ones (Cartesian CG) are based on reducing the atomistic representation of the nucleotides to a few beads whose interactions are defined by empirical potentials and whose movements are followed by means of (typically) Langevin-Brownian MD algorithms (15–17). The second family of methods (helical CG) reduces the degrees of freedom in DNA by considering the nucleobases or the base pairs (bp) as rigid planes whose movements are defined by three rotations and three translations. In this second family of methods the sampling is typically obtained by means of Monte Carlo (MC) simulation techniques. While the Cartesian CG methods have the advantage of universality, for physiological DNAs, helical CG methods are probably more efficient as helical coordinates are better suited to describe the essential movements of DNA (12, 13).

Three crucial choices must be taken in defining a helical CG model. The first one is the level of resolution: nucleobases or base pairs. In nucleobase-resolution scheme the CG model should account for the movement of each nucleobase with respect to

three neighbors in a simple base pair step (bps) (the paired one, one located at the 3', and one at 5' in the opposite strand), which sums up to 6^3 degrees of freedom per nucleobase. By combining nucleobase (intra base pair) and base pair step (inter base pair) helical coordinates the number of degrees of freedom can be significantly reduced (18–20). Simpler and more popular (21, 22) are helical-CG methods that represent the DNA at the base pair level. In this case movements at the base pair step level are limited to three translations and three rotations (shift, slide, rise, tilt, roll, twist), which at the expense of some loss of resolution, drastically simplifies the calculation and the parameterization of the model.

The second important choice in building a helical CG model is the nature of the Hamiltonian (energy function) used to describe the dependence between the energy of the system and the change in helical coordinates. Most CG models rely on the use of a harmonic Hamiltonian (1, 12, 13, 18–22), which assumes that under normal conditions the distributions of helical coordinates (at either nucleobase or base pair level) are Gaussian. Within this assumption the energy of the DNA can be easily described by means of a stiffness matrix and a deformation vector indicating the deviation of a helical coordinate from its equilibrium value (21). For the most common base pair resolution model this means that the energy is computed as shown in equation 1:

$$E(X) = \sum_{j=1}^{N} \frac{1}{2} K_j \Delta X_j^2 \qquad (1)$$

where E is the energy, N is the number of bps, K_j is the 6x6 stiffness matrix for bps j, and ΔX_j is the 6-dimension deformation vector ($\Delta X_j = X_j - X_j^0$), with X_j and X_j^0 being the current conformation vector of bps at a given point of the ensemble and the equilibrium vector respectively.

The last choice in the definition of a helical CG model is the origin of the parameters (stiffness matrix and the equilibrium vector X_j^0 used to compute the deformation vector) defining the energy function. Original models developed by Olson & Zhurkin (21) used a nearest-neighbor (NN) scheme, where parameters for the ten unique

bps were derived by inspection of the helical geometries of bps found in databases of crystal structures of DNA-protein complexes. Further refinements used MD simulation of different DNA duplexes containing the ten-unique bps as source of parameters (22, 23). More recently, as the shortcomings of the NN scheme became evident, new harmonic models relying on inter base pair parameters adapted to all the different tetranucleotides emerged (1, 6, 24), with the corresponding parameters being fitted from atomistic MD simulations. These models showed a good ability to reproduce the conformational space of DNA duplexes, but were limited by two fundamental problems: i) they were parameterized from the parmbsc0 force-field (2) which showed caveats in the representation of certain characteristics of the helix and ii) they were based on the harmonic approximation, which is unable to reproduce multimodality shown both experimentally and theoretically in the distribution of inter base pair coordinates of certain bps (4, 6, 7, 25, 26).

We present here an evolution of the helical CG model which assumes a novel multi-normal model which accounts for the non-Gaussian nature of some inter base pair deformations and considers a flexible extended nearest neighbor model (eNN model), which reproduces very well the impact of remote neighbors in the definition of the deformability of bps. Parameters (stiffness and equilibrium values per state and shifting values between states) were derived from atomistic MD simulations using parmbsc1 force-field and state-of-the-art simulation procedures. Sampling is obtained by means of a highly efficient Metropolis Monte Carlo algorithm. The method has been implemented in a server (http://mmb.irbbarcelona.org/MCDNAlite/) which incorporates tools that, taking advantage of correlations between helical states and backbone conformation (25, 27) allows the atomistic-level reconstitution of the DNA at the nucleobase and backbone level. The method produces MC ensembles difficult to distinguish from atomistic MD trajectories with a fraction of computational cost and reproduces well known experimental structures.

THE ALGORITHM

Hamiltonian definition: A recent analysis of the dynamics of the 136 unique tetranucleotides of B-DNA performed by the ABC consortium (25) revealed that 80% of the 816 (136x6) unique inter base pair distributions cannot be correctly described using a single normal distribution (http://mmb.irbbarcelona.org/miniABC/ ;(25)). As described elsewhere (4) in many cases the peaks of the fitted normal distributions are close, and a single unimodal function can reasonably describe the real distribution. However, in 4% of the cases at least a bimodal distribution must be used to obtain a reasonable fit to the real distribution. Bimodality can be seen in slide (several tetranucleotides containing the central d(GpG) step), shift (typically in a few tetranucleotides containing d(YR) central step), and twist (very often in tetranucleotides containing central d(CG) or d(AG) steps). Certain tetranucleotides, such as d(CdTdAdG) show especially complex distributions (26) impossible to describe by a single Gaussian. In summary, the normality assumption on which the harmonic model is based should be revisited for more realistic representations of DNA flexibility.

We propose here a new Hamiltonian inspired by empirical valence bond theory (28), where we assume that the distribution of inter base pair parameters (shift, slide, rise, tilt, roll, twist) underlies a Boltzmann-averaged combination of Gaussian distributions. The Hamiltonian leading to such a distribution can be derived as shown in equation 2:

$$E(X) = -k_B T \sum_{j=1}^{N} \ln \sum_{i=1}^{n} e^{-\frac{1}{k_B T}\left(\frac{1}{2}K_{ij}\Delta X_{ij}^{\ 2}+E_{ij}\right)} \qquad (2)$$

where k_B is the Boltzmann constant, T is the temperature, N is the number of bps, n is the number of states in which the distribution of inter base pair parameters of a given bps (in its sequence environment) can be decomposed (see below), K is the stiffness matrix associated to the state i in step j; ΔX is the deformation vector (with equilibrium values dependent on step j and state i) and E_{ij} is the relative energy of state i at bps j (shifting values between states). Note that for a single unimodal distribution eq. 2 leads to the classical harmonic model shown in eq. 1. Also note

that due to sequence end effects single state dimer stiffness parameters are used for the first and last bps.

Definition of the states: Eq. 2 implies that the energy is computed from a set of stiffness matrices and deformation vectors which are not only dependent on the step, but also on the state. In principle, if there are m states for each inter base pair distribution, we should expect m^6 states at the bps level (*i.e.* for bimodality m=2 we could expect 64 different stiffness matrices and equilibrium vectors for each bps). Fortunately, the number of unique helical states is smaller as some inter base pair parameters are correlated and others show a purely uninormal-unimodal distribution. To assign in a systematic manner the number of states to describe a given bps we process µs-long parmbsc1 MD simulation of a large number of duplexes (see Table 1) containing the 136 unique tetranucleotides (data can be downloaded from http://mmb.irbbarcelona.org/BigNASim/ (29)). To this end, we transform the original inter base pair coordinates of the central bps of each tetranucleotide in a new set of dimensionless parameters using Lankaš transformation (30); see eq. 3:

$$\gamma_i^* = \delta\gamma_i \ + (1-\delta)10.6\gamma_i \qquad (3)$$

where γ and γ^* are normal and dimensionless inter base pair parameters and δ is a Heaviside step function equal to 1 if γ is a translational parameter (measured in Å) and is equal to 0 when it is a rotational parameter (measured in degree).

Principal component analysis (PCA) is then performed to reduce the coordinatespace where a certain number of components (those explaining at least 80% of variance) are kept (usually 3). Original trajectories projected in this reduced space are subjected to clustering following a Gaussian finite mixture model (31). The MD ensemble is then divided into several sub-ensembles for which the equilibrium vector (X_0) is determined. The covariance matrix in the original inter base pair parameter space is defined and inverted (22) to obtain the stiffness matrix specific for a given state of a bps in a certain tetranucleotide environment. Finally, all the

harmonic models defining the global energetics of the tetranucleotide are combined by using eq. 2.

Monte Carlo simulations. Simulation of the movements of the DNA at the CG level were performed using eq. 2 (or for comparison eq. 1) implemented in a Monte Carlo (MC) sampling algorithm, where movements in the inter base pair parameter space are attempted and accepted or not based on the Metropolis algorithm. For each MC move one to four inter base pair parameters are randomly selected to be modified. The strength of the change is determined by two values: a scaling factor which is dependent on the diagonal entry of the stiffness matrix of the inter base pair parameter and which is scaled to guarantee ~40 % acceptance rate. The output of a MC run is a long file of 6xNxT (N number of bps, T number of snapshots) inter base pair coordinates, which can be partially or totally transformed into Cartesian coordinates as described below. The sampling algorithm is implemented in a simple web interface (http://mmb.irbbarcelona.org/MCDNAlite) and ready to download as a stand-alone version via the web interface (http://mmb.irbbarcelona.org/MCDNAlite/standalone).

Atomic detail reconstitution. The inter base pair coordinates collected from the MC algorithm above were transformed to derive Cartesian representations of the DNA (Figure 1), as in many cases this is the level of detail required to understand DNA functionality. For a given set of inter base pair coordinates the positions of the phosphates were derived from helical axis by using Lavery's rules (see Figure 1; (32)). Atomistic coordinates of the nucleobases were derived using the SCHNArP algorithm (33), and backbone torsions were reconstituted using the correlations between inter base pair coordinates and backbone states (BI or BII) found in a recent ABC study (25). Thus, for each tetranucleotide the inter base pair coordinate showing the highest correlation with the backbone state is used as a classifier of the backbone state (typically shift; see Suppl. Table S1). The accuracy of the backbone state prediction is typically in the range of 80-90% (see Suppl. Figure S1). Average BI and BII backbone conformations for each of the 16 dimers were extracted from the meta-trajectory of all the occurrences of the dimers in a recent ABC simulation set (see Table 1) and fit to the nucleobase position defined by the inter base pair

coordinates (see Figure 1). A short restrained steepest descent optimization relaxes mismatched local geometries without altering state definition. The mesoscopic MC-eNN ensemble using full atomistic reconstruction can be analyzed with any common MD analysis tool (links to NaFlex (34) are included in the web interface), which highly increases the usability of the model.

Data and analysis tools. Original trajectories were obtained in previous works using parmbsc1 force-field (3) and standard simulation protocols used by the ABC consortium ((6), individual simulation times at least 1 μs; data deposited at BigNASim (29) database; ID 'miniABC_K'). DNA inter base pair parameters, groove widths and backbone torsion angles were measured and analyzed with the CURVES+ and CANAL programs (32, 35). Principal component analysis (PCA) in Cartesian space was done using PCASUITE (http://mmb.pcb.ub.es/software/pcasuite/pcasuite.html). Essential dynamics of simulated trajectories were obtained using the Boltzmann's averaged absolute similarity index (36). BIC (Bayesian Information Criterion) was used to determine the normal (one Gaussian) or multi-peaked nature of the distributions of inter base pair parameters (see Suppl. Methods and (37, 38)). For multi-peaked distributions we used an extension of the Helguerro's theorem (39, 40) to distinguish those cases where the Gaussians are very close (unimodal) from those where they are significantly separated. Clustering was done using the mclust library (41) in R 3.1.2. The same software package was used to perform all the statistic studies in the manuscript.

RESULTS AND DISCUSSION

The inter base pair parameter space from MD simulations. All of the 136 tetranucleotides and 80% of the 136x6 inter base pair distributions can be classified as multi-peaked, but fortunately, only 20 % of the tetranucleotides and 4 % of individual inter base pair distributions are multi-modal based on Helguerro's theorem. However, these numbers mask the complexity of the coupling between inter base pair coordinates. This is illustrated by inspection of normalized bi-dimensional distributions (Figure 2 for examples), which show the existence of 4

major scenarios: i) the inter base pair parameters are uncorrelated and show uninormal distributions leading to clear 2D Gaussian distributions, ii) the two parameters show unimodal distributions, but are correlated leading to ellipsoidal shaped distributions, iii) at least one of the two parameters is double-peaked resulting in two hotspots in the bi-dimensional map and finally, iv) multiple peaks in two inter base pair parameters and correlation between them lead to a complex bi-dimensional probability distribution. Certainly, by moving to higher dimensions more complex probability distributions impossible to represent by combining 1D distribution would be encountered.

To define unambiguously the number of states required to define the preferentially sampled regions we performed a clustering algorithm (see Methods), finding that most tetranucleotides can be represented by 3-5 clusters (Figure 3). The need to use more than 5 clusters is found in less than 10% of the cases (Figure 3), but those tetranucleotides where a single state is enough to represent the sampling are even less common. As expected from previous studies (6, 25, 26), shift and twist are the main drivers for the multiplicity of states (see Suppl. Table S2). Note that no assumption on unimodality is made for the derivation of the different states, which means that an inter base pair parameter distribution of an individual state may be classified as multimodal. However, when Bayes-Helguerro's analysis is done at the state level, in only 0.8% of the clustered distributions (3192 in total) unimodality is not satisfied and overall multi-normality decreases from 80% to 20%. This means that the dimension reduction and clustering process outlined here reduces dramatically the problem of multi-normality and multi-modality (see examples in Figure 3 and Suppl. Figure S2) and produces a robust protocol to define the number of states where a harmonic behavior is granted, the basic assumption required to use eq. 2.

Equilibration and convergence of Monte Carlo Simulations. Before analyzing the performance of the eNN method we evaluate the expected length of the simulation required to obtain reasonably converged ensembles. To this end we performed several MC simulations (room temperature) of duplexes of random sequence and lengths ranging from 10 to 1,000 base pairs using Arnott's fiber data to generate the

starting structures. As Arnott's parameters are known to overestimate twist by 1-2 degrees (42) we can evaluate the performance of the MC method to relax and equilibrate an incorrect structure. Results in Figure 4A (and Suppl. Figure S3) indicate that for the most sensitive parameter (the number of helical turns) equilibration is achieved when the number of collected configurations equals the length of the oligomer multiplied by 200 (for other parameters such as end-to-end distance convergence is faster, *i.e.* around 100 x length). Thus, for the largest oligomer considered here (1,000 bp) equilibration is achieved after 100,000-200,000 Monte Carlo steps. For oligomers of a size compatible with atomistic MD simulations (~ 50 bp) equilibration is so fast that it is not visible in the plots (Suppl. Figure S3).

Once the rules for the equilibration time were clear we evaluated the length of the ensemble required to obtain converged distributions of local and global DNA properties. Results in Figure 4B (and Suppl. Figure S4) show that in general good sampling for sensitive global parameters such as the helical turns is obtained after a reasonably small number of configurations selected after equilibration (around 10,000-20,000 configurations). Irrespectively of the length of the duplex convergence in local geometry takes from 10,000 to 40,000 configurations depending on the complexity of the tetrad accessible inter base pair parameter space (see examples in Suppl. Figure S5 and S6). When comparison is possible, MC-convergence is faster than that obtained from MD simulations (see Suppl. Figure S6 and discussion below).

MC-eNN calculations reproduce well atomistic MD trajectories. We compare ensembles obtained for several medium-sized DNA duplexes (Suppl. Table S3) using our MC-eNN protocol and 0.5-2 μs long atomistic MD simulations (using parmbsc1 force-field). Figure 5 shows that MC and MD trajectories for the same sequence are nearly indistinguishable. Auto-similarity indexes (diagonal in Figure 5) are always larger than cross-similarity index (for a common set of equal atoms) which indicates that the MC-eNN method reproduces very well the sequence-specific details of the deformability of DNA. Local (Figure 6) inter base pair distributions obtained from MC-eNN calculations are impossible to differentiate from those derived from

atomistic MD simulations, even in those cases where the inter base pair probability distributions are correlated in a highly non-linear manner, impossible to capture by a standard harmonic model (see Suppl. Figure S7). To test the limit of the method we compared MC-eNN and MD ensembles for the longest naked DNA duplex in the BigNASim database (56 bp in length, see Suppl. Table S4). The essential dynamics obtained from MC and MD samplings are nearly indistinguishable(absolute similarity index of 0.88; see Suppl. Figure S8) and the same level of agreement is found when looking to sequence-dependent inter base pair properties (Figure 7A-C and Suppl. Figure S9). In addition, even local and fine details, such as compensatory changes in neighboring steps, or the inter base pair distributions at highly structural polymorphic sites are well captured by the MC-eNN model.

The reconstitution protocol provides reasonable backbone conformations, leading to "atomistic" reconstitutions that are hard to distinguish from the atomistic MD simulations. For example, for the 56-mer duplex the RMSd (using all heavy atoms as reference) of the ensemble *vs* the MD-averaged structure is around 0.09 Å x bp, while the RMSd increases to only 0.11 Å x bp when the MC-eNN ensemble is compared with the MD-averaged structure. Groove dimensions and many other subtle structural details such as the distribution of BI/BII states or the puckering of the sugar are well reproduced by the method (Figure 7D-F) reflected by an average difference in groove widths between MC-eNN and MD of (0.28 $\pm$0.68) Å and a linear correlation coefficient of 0.85 of BI population along the sequence of MC-eNN vs MD, significantly higher than when MD is compared with NMR experiments (0.45 in average, (43)). The difference in backbone populations of MC-eNN and MD (1.1 ± 10.8)% and lies within the experimental accuracy of backbone state determination (10%; (43)) in more than 70% of the cases, compared to 53% when MD and experiment are compared (43). Both MC-eNN and MD experience a South *vs.* North pucker population of 0.95-1.00 in over 90% of the cases with overall mean Phase angle of P = (161 ± 19)° for MC-eNN compared to P = (149 ± 30)° in MD. In summary, it seems that the "atomistic" structures derived from MC-eNN calculations are accurate enough as to be used to discuss specific protein-binding to the DNA.

MC-eNN calculations reproduce well experimental structures. We performed an exhaustive comparison of MC-eNN ensembles with experimental (X-Ray or NMR) structures in PDB (Figure 8, Suppl. Figures S10 and S11 and Suppl. Table S5). Our structures at T=300 K show average RMSd around 0.3 Å x bp (using all heavy atoms as reference) from the known experimental structure, a value that is close to those found in atomistic MD trajectories performed at the same temperature (see Suppl. Table S5), and not far from the RMSd generated by thermal noise (around 0.1 Å x bp, see previous section). The performance of the MC-eNN calculations is such that we can detect regions where experimental structures might need to be revisited. For example, large compensatory twist oscillations likely originated from the refinement protocol (1DN9, 1HQ7 in Suppl. Figure S11D-E), or regions where anomalous inter base pair parameter values (low Roll in last bps and very high twist in bps 4 for PDB id 2JYK in Figure 8B) occur.

Computational performance. The MC-eNN method is very efficient from a computational point of view. To obtain converged complex inter base pair distributions (see Figure 4B and Suppl. Figure S3 and Suppl. Figure S5) atomistic MD simulation of a 56-mer duplex (~550,000 atoms) would require more than 500 days in one of our 64-core cluster (400 ns of trajectory), while to obtain equivalent sampling (as determined from the convergence rate) would require only 12 minutes in the same machine using the MC-eNN method outperforming MD by a factor of $\sim 10^5$ (see Suppl. Figure S12). The difference in computer performance between MD and MC-eNN calculations increases for larger duplexes, as the cost of MD simulations scales with the third power of the length of the DNA, while MC-eNN simulation time increases only linearly with the length of the duplex. Furthermore, contrary to atomistic MD, MC-eNN scales perfectly with the number of processors, which facilitates its use in supercomputers.

The MC-eNN webserver. The MC-eNN simulation method is distributed as a stand-alone executable version for MacOS and Linux systems (see Supplementary Information; source code is available upon request), but it is also accessible as a webserver http://mmb.irbbarcelona.org/MCDNAlite/ (the stand-alone version can be downloaded via the webserver

http://mmb.irbbarcelona.org/MCDNAlite/standalone) which requires just the sequence of the duplex as input and provides as output a limited number of alternative conformations, selected to capture the most probable configurations according to the states at tetranucleotide level. All results can be viewed directly in the web interface and downloaded for further local analysis. A direct link in the webserver to our NAFlex tool (34) constitutes a user-friendly way for deeper online analysis of the DNA structures.

CONCLUSIONS

We present a new mesoscopic model for the representation of the structure and dynamics of naked DNA structures, which integrates all the information acquired from the analysis of B-DNA dynamics from the latest efforts published by the ABC consortium. The method maintains the simple bps model, but tackles rigorously the multi-modality of inter base pair distributions and their dependence on nearest neighbors, allowing an accurate representation of complex polymorphisms in DNA. The mesoscopic ensembles provided by our algorithm can be transformed to atomistic models of DNA with a high accuracy even in local details, something beyond the expectations of a mesoscopic model. The method is extremely efficient, making it possible to simulate long fibers of DNA that will be unreachable for atomistic MD simulation in the next decades. It is implemented in simple tools that can be used by non-experts aiming to obtain a more complete picture of DNA than that derived from the inspection of canonical average structures.

ACKNOWLEDGEMENTS

We thank the ABC consortium for the atomistic MD simulations used to derive the parameters and all the colleagues there for many discussions on the simulation engine. We are also indebted to Prof. R. Lavery for Curves+ (35) and for his phosphate-location approach. We also thank Dr. F. Battistini for providing MD simulations to benchmark the MC-eNN method and for inspiring discussions. J.W. and A.B. are La Caixa PhD fellows (UB and IRB Barcelona, Spain). P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema

Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher. M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avançats) researcher.

AUTHOR CONTRIBUTIONS

M.O and J.W designed the method, J.W. developed and coded the method and did all the analysis with support of P.D.D.. A.B. provided average backbone states of all nucleobases. J.W. and A.H. developed the backend of the webserver while G.B. was responsible for setting up the webpage. J.W., P.D.D., and M.O. discussed the analysis and wrote the manuscript with contributions from all the co-authors. M.O. directed the work.

REFERENCES

1. Dans,P.D., Walther,J. and Gómez,H. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45. https://doi.org/10.1016/J.SBI.2015.11.011

2. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.*, **92**, 3817–3829. https://doi.org/10.1529/biophysj.106.097782 http://www.ncbi.nlm.nih.gov/pubmed/17351000

3. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58. https://doi.org/10.1038/nmeth.3658

4. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring

polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–78.
https://doi.org/10.1093/nar/gks884
http://www.ncbi.nlm.nih.gov/pubmed/23012264

5. Balaceanu,A., Pasi,M., Dans,P.D., Hospital,A., Lavery,R. and Orozco,M. (2017) The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28.
https://doi.org/10.1021/acs.jpclett.6b02451

6. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–83.
https://doi.org/10.1093/nar/gku855
http://www.ncbi.nlm.nih.gov/pubmed/25260586

7. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–20.
https://doi.org/10.1093/nar/gku809
http://www.ncbi.nlm.nih.gov/pubmed/25223784

8. Dans,P.D., Ivani,I., Hospital,A., Portella,G., González,C. and Orozco,M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, gkw1355.
https://doi.org/10.1093/nar/gkw1355
http://www.ncbi.nlm.nih.gov/pubmed/28088759

9. Dans,P.D., Danilāne,L., Ivani,I., Dršata,T., Lankaš,F., Hospital,A., Walther,J., Pujagut,R.I., Battistini,F., Gelpí,J.L., *et al.* (2016) Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.*, **44**, 4052–4066.
https://doi.org/10.1093/nar/gkw264

http://www.ncbi.nlm.nih.gov/pubmed/27084952

10. Balaceanu,A., Pérez,A., Dans,P.D. and Orozco,M. (2018) Allosterism and signal transfer in DNA. *Nucleic Acids Res.*, **46**, 7554–7565. https://doi.org/10.1093/nar/gky549

11. Dršata,T. and Lankaš,F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102. https://doi.org/10.1088/0953-8984/27/32/323102 http://www.ncbi.nlm.nih.gov/pubmed/26194779

12. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193. https://doi.org/10.1016/j.sbi.2008.01.005 http://www.ncbi.nlm.nih.gov/pubmed/18304803

13. Orozco,M., Pérez,A., Noy,A. and Luque,F.J. (2003) Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.*, **32**, 350–364. https://doi.org/10.1039/B207226M

14. Gómez,H., Walther,J., Darré,L., Ivani,I., Dans,P.D. and Orozco,M. (2017) Chapter 7. Molecular Modelling of Nucleic Acids. In.pp. 165–197. https://doi.org/10.1039/9781788010139-00165

15. Dans,P.D., Zeida,A., Machado,M.R. and Pantano,S. (2010) A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics. *J. Chem. Theory Comput.*, **6**, 1711–1725. https://doi.org/10.1021/ct900653p

16. Ouldridge,T.E., Šulc,P., Romano,F., Doye,J.P.K. and Louis,A.A. (2013) DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic Acids Res.*, **41**, 8886–8895.

https://doi.org/10.1093/nar/gkt687

17. Freeman,G.S., Hinckley,D.M., Lequieu,J.P., Whitmer,J.K. and de Pablo,J.J. (2014) Coarse-grained modeling of DNA curvature. *J. Chem. Phys.*, **141**, 165103.
https://doi.org/10.1063/1.4897649

18. Petkevičiūtė,D., Pasi,M., Gonzalez,O. and Maddocks,J.H. (2014) cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.*, **42**, e153–e153.
https://doi.org/10.1093/nar/gku825

19. Dršata,T., Zgarbová,M., Špačková,N., Jurečka,P., Šponer,J. and Lankaš,F. (2014) Mechanical Model of DNA Allostery. *J. Phys. Chem. Lett.*, **5**, 3831–5.
https://doi.org/10.1021/jz501826q
http://www.ncbi.nlm.nih.gov/pubmed/26278756

20. Lankaš,F., Gonzalez,O., Heffler,L.M., Stoll,G., Moakher,M. and Maddocks,J.H. (2009) On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, **11**, 10565.
https://doi.org/10.1039/b919565n

21. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 11163–8.
https://doi.org/10.1073/PNAS.95.19.11163
http://www.ncbi.nlm.nih.gov/pubmed/9736707

22. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.*, **85**, 2872–2883.
https://doi.org/10.1016/S0006-3495(03)74710-9

23. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2010) 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**, 299–313.
https://doi.org/10.1093/nar/gkp834
http://www.ncbi.nlm.nih.gov/pubmed/19850719

24. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H., *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**, 3721–40.
https://doi.org/10.1529/biophysj.105.067397
http://www.ncbi.nlm.nih.gov/pubmed/16169978

25. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Lavery,R., Maddocks,J.H. and Orozco,M. (2019) THE PHYSICAL PROPERTIES OF B-DNA BEYOND CALLADINE-DICKERSON RULES. *Submitted*.

26. Balaceanu,A., Buitrago,D., Walther,J., Hospital,A., Dans,P.D. and Orozco,M. (2019) Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.*, 10.1093/nar/gkz255.
https://doi.org/10.1093/nar/gkz255

27. Zgarbová,M., Jurečka,P., Lankaš,F., Cheatham,T.E., Šponer,J. and Otyepka,M. (2017) Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for All 136 Distinct Tetranucleotide Sequences. *J. Chem. Inf. Model.*, **57**, 275–287.
https://doi.org/10.1021/acs.jcim.6b00621

28. Colizzi,F. and Bussi,G. (2012) RNA Unwinding from Reweighted Pulling Simulations. *J. Am. Chem. Soc.*, **134**, 5173–5179.

https://doi.org/10.1021/ja210531q

29. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M., *et al.* (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
https://doi.org/10.1093/nar/gkv1301
http://www.ncbi.nlm.nih.gov/pubmed/26612862

30. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V, Sponer,J. and Lankaš,F. (2013) Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.

31. DAY,N.E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.
https://doi.org/10.1093/biomet/56.3.463

32. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–23.
https://doi.org/10.1093/nar/gkv080
http://www.ncbi.nlm.nih.gov/pubmed/25662221

33. Lu,X.J., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: Rebuilding program (SCHNArP). *J. Mol. Biol.*, **273**, 681–691.
https://doi.org/10.1006/jmbi.1997.1345
http://www.ncbi.nlm.nih.gov/pubmed/9356256

34. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47-55.
https://doi.org/10.1093/nar/gkt378

http://www.ncbi.nlm.nih.gov/pubmed/23685436

35. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–29.
https://doi.org/10.1093/nar/gkp608
http://www.ncbi.nlm.nih.gov/pubmed/19625494

36. Alberto Pérez,†,‡, José Ramón Blas,†, Manuel Rueda,†,§, Jose María López-Bes,‡, Xavier de la Cruz,†,|| and and Modesto Orozco*,†,§,⊥ (2005) Exploring the Essential Dynamics of B-DNA. 10.1021/CT050051S.
https://doi.org/10.1021/CT050051S

37. Schwarz,G. (1978) Estimating the Dimension of a Model. *Ann. Stat.*, **6**, 461–464.
https://doi.org/10.1214/aos/1176344136

38. Kass,R.E. and Raftery,A.E. (1995) Bayes Factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
https://doi.org/10.1080/01621459.1995.10476572

39. Schilling,M.F., Watkins,A.E. and Watkins,W. (2002) Is Human Height Bimodal? *Am. Stat.*, **56**, 223–229.
https://doi.org/10.1198/00031300265

40. DE HELGUERO,D.D.F. (1904) SUI MASSIMI DELLE CURVE DIMORFICHE. *Biometrika*, **3**, 84–98.
https://doi.org/10.1093/biomet/3.1.84

41. Fraley,C., Raftery,A.E., Scrucca,L., Brendan Murphy,T. and Fop,M. (2016) Gaussian Mixture Modelling for Model-Based Clustering,Classification, and Density Estimation [R package mclust version 5.4.2]. *R J.*, **8**, 205–233.

42. Arnott,S. and Hukins,D.W.L. (1972) Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.*, **47**, 1504–1509. https://doi.org/10.1016/0006-291X(72)90243-4

43. Ben Imeddourene,A., Elbahnsi,A., Guéroult,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLOS Comput. Biol.*, **11**, e1004631. https://doi.org/10.1371/journal.pcbi.1004631

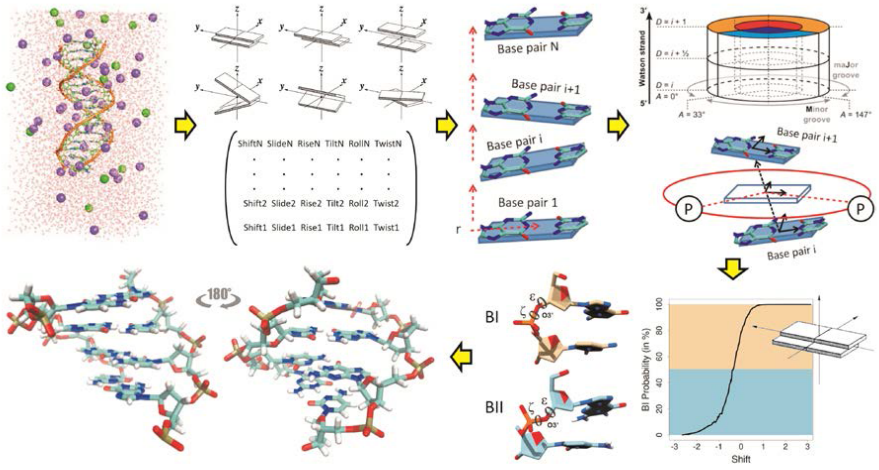


Figure 1. Workflow of the MC-eNN model. The model is parametrized by MD simulations of a sequence set of all unique 136 tetramers (see Table 1 for sequences). Monte Carlo sampling in the inter base pair parameter space based on the new Hamiltonian (see eq. 2) of a structure with N+1 base pairs yields a set of 6xNxT inter base pair coordinates (T is the number of structures sampled). For a single structure, atomistic coordinates of the nucleobases are derived using the SCHNArP algorithm (31) and the position of the phosphates relative to the helical axis using Lavery's rules (30) are determined. Using correlations of inter base pair parameters and backbone torsions the backbone states are classified to either BI or BII. For each central bps of a tetranucleotide the inter base pair coordinate showing the highest correlation with the backbone state is used as a classifier of the backbone state (see Suppl. Table S1 and Methods for more details). Average BI and BII backbone conformations for each of the 16 dimers were fit to the nucleobase position defined by the inter base pair coordinates. A short restrained steepest descent optimization relaxes mismatched local geometries resulting in the final structure (for more details see Methods).

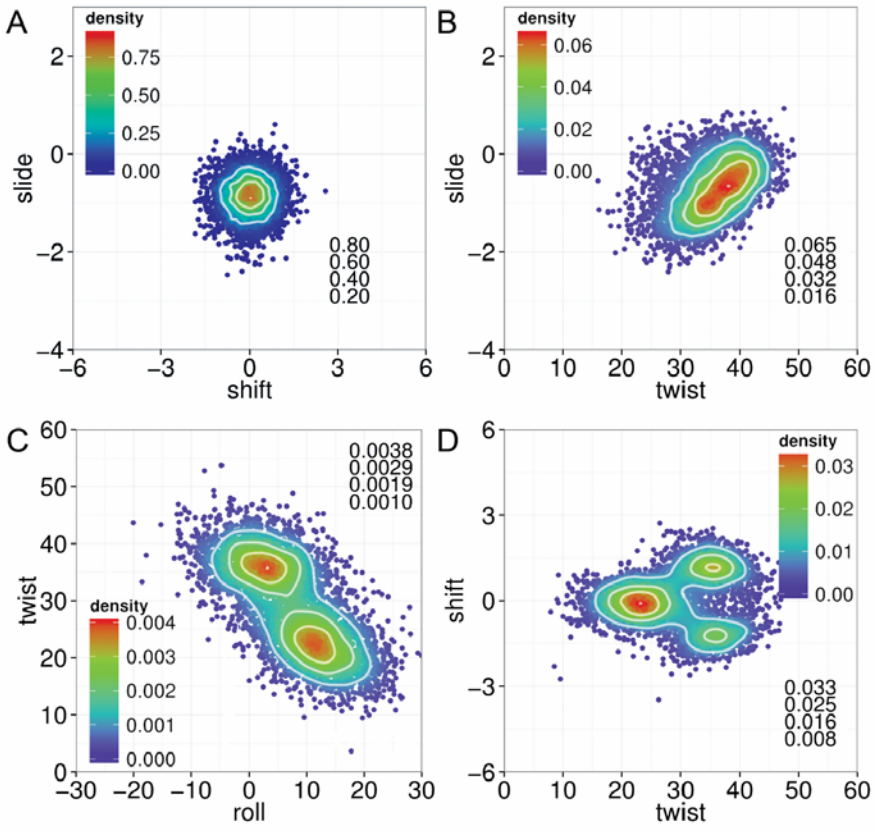


Figure 2. Examples of the four different scenarios of bi-dimensional inter base pair parameter distributions found in the BigNASim database. A) Two uncorrelated and uninormal distributions show Gaussian behaviour (tetramer AATT in MD simulation with BigNASim ID 'DDD_800ns'). B) Unimodal distributions which are correlated show elipsoidal shaped pattern (tetramer AAGC in MD simulation with BigNASim ID 'miniabc_K_12'). C) Two hotspots appear when at least one of the two parameters contains two separate peaks (third appearance of tetramer CTAG in MD simulation with BigNASim ID 'AGCT'). D) A complex multi-peaked bi-dimensional map is obtained when both inter base pair parameters are multimodal and correlated (tetramer TCGA in MD simulation with BigNASim ID 'miniabc_K_10'). The four isodensity lines equal to 100%, 75%, 50% and 25% of the maximum density and the corresponding values are shown in each plot.

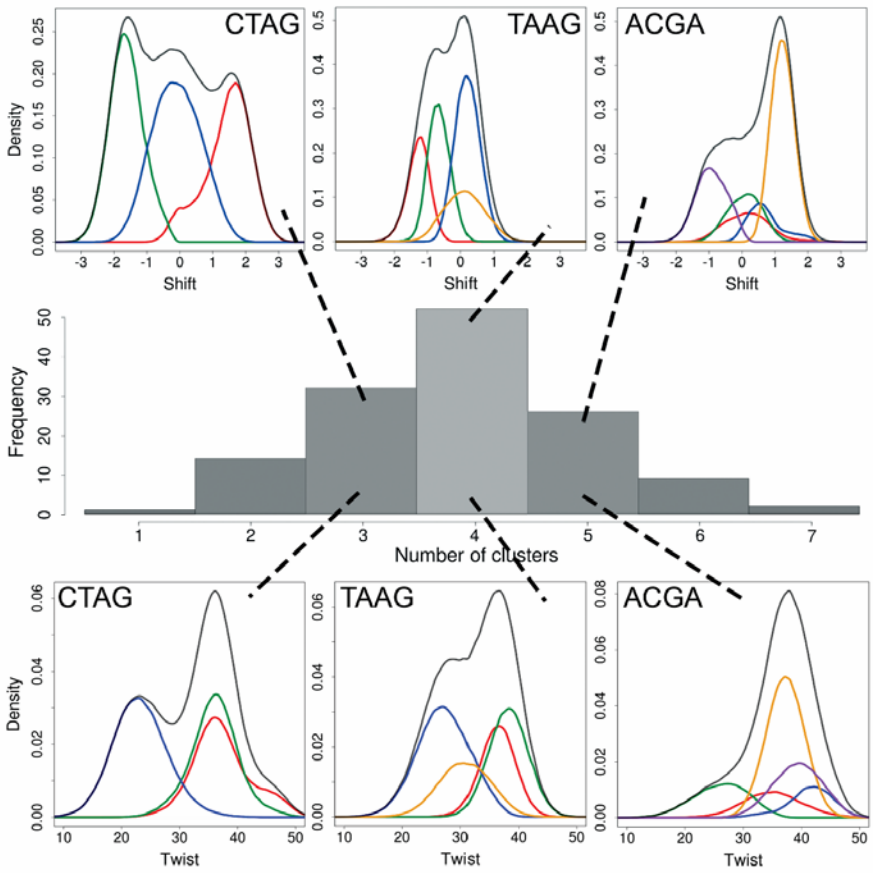


Figure 3. Histogram of the number of clusters to represent the six-dimensional inter base pair parameter space of the 136 unique tetramers (Middle). Examples for the division of inter base pair parameter distributions into multiple states for the most common number of clusters are shown for Shift (Top) and Twist (Bottom) for the tetramers CTAG (3 clusters), TAAG (4 clusters) and ACGA (5 clusters). The inter base pair parameter distributions (grey) are clustered into several distributions shown in green, blue and red for 3 clusters; green, blue, red and orange for 4 clusters and green, blue, red, orange and purple for 5 clusters.

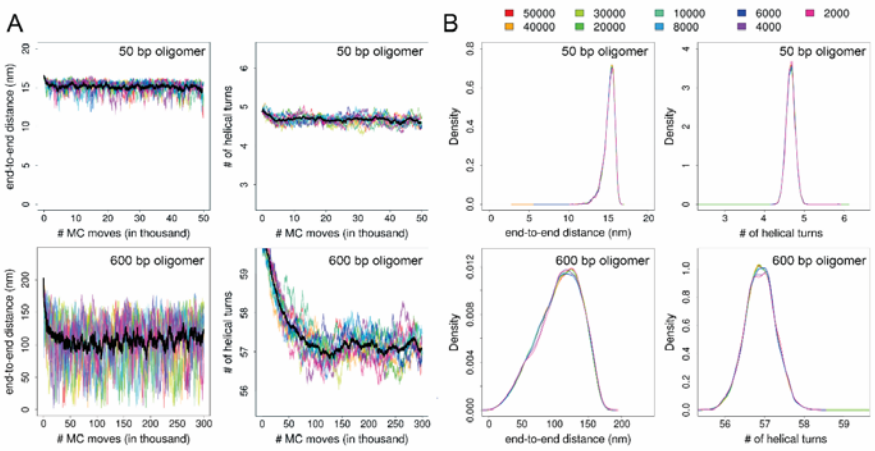


Figure 4. Equilibration and convergence of the MC-eNN simulation. A) Number of MC moves needed for fiber equilibration obtained by investigating end-to-end distance (left) and number of helical turns (right) of 10 individual simulations of a fiber of random sequence of 50 bp (top) and 600 bp (bottom) in length. The 10 individual simulations are shown in different colors and a black line illustrates the average of the 10 simulations. Equilibration is obtained when the number of Monte Carlo moves equals the length of the oligomer multiplied by 200 (see main text). B) Convergence rules were achieved by comparing the length of the ensemble needed to obtain converged distributions. Distributions of end-to-end distance (left) and number of helical turns (right) of 2,000-50,000 configurations of a fiber of random sequence of 50 bp (top) and 600 bp (bottom) in length show that a small number of configurations is sufficient for good sampling of sensitive global fiber parameters. Note: the maximum of the scale of the axis of end-to-end distance is calculated as 4Å x fiber length (in base pair).

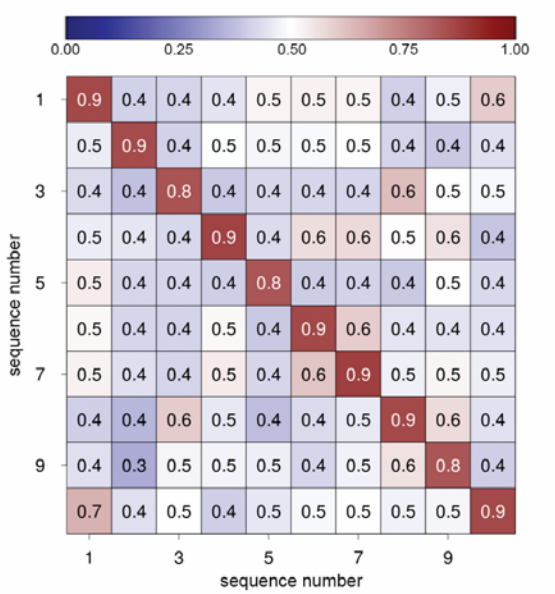


Figure 5. Essential dynamics between MC-eNN and MD. Boltzmann-weighted absolute similarity indices (34) of MC-eNN simulations (x-axis) versus MD simulations (y-axis) of 10 different sequences of 18 bp in length (see Suppl. Table S3) are calculated with the first 30 eigenvectors of the central 14 base pairs. Absolute similarity indices of simulations of the same sequence (auto similarity index) are computed using as reference all the heavy nucleobase atoms while when comparing two different sequences the heavy atoms common in all the nucleobases are used as reference (cross similarity index). The numbers in each square of the matrix are rounded to a single decimal number and each square of the matrix is color-coded according the color legend. Important to note is that the auto similarity indices (diagonal) are always higher than the cross-similarity indices when comparing MC-eNN and MD.

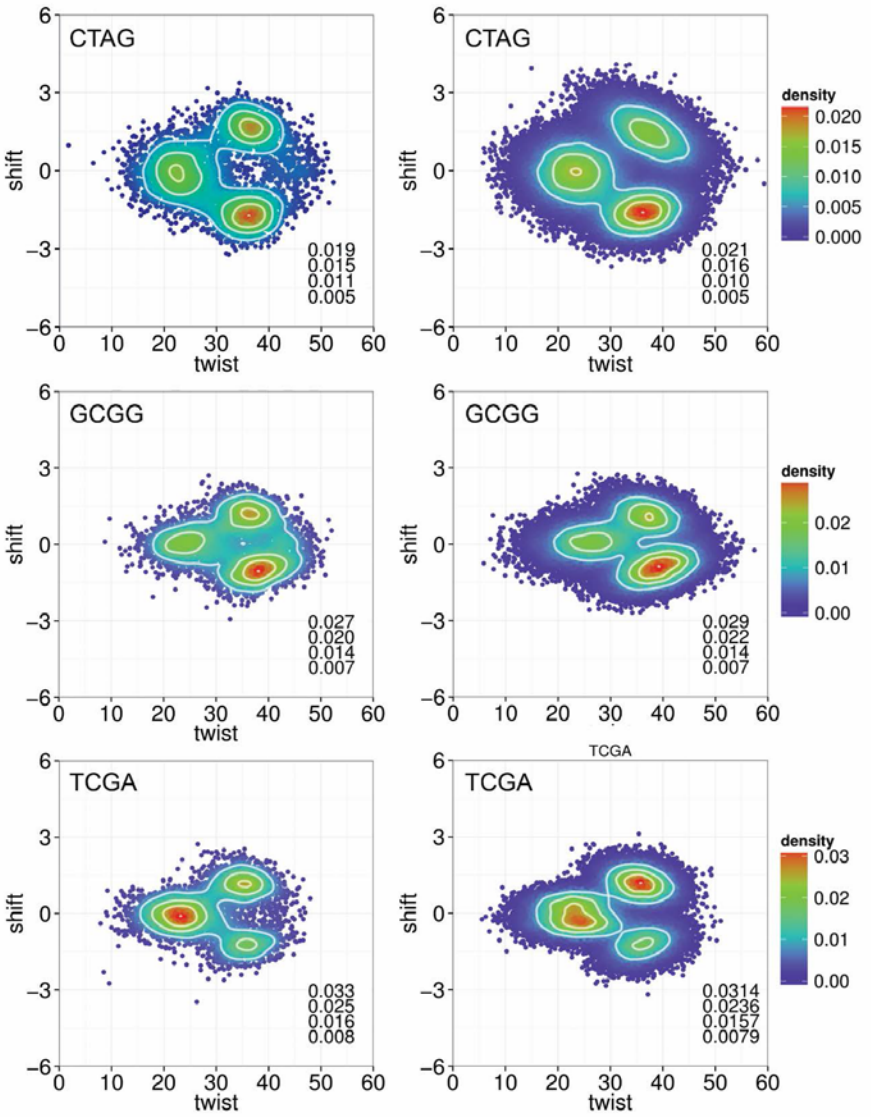


Figure 6. Bi-dimensional inter base pair parameter maps of Twist-Shift of three tetramers CTAG (top), GCGG (middle) and TCGA (bottom) of MD simulations of the parmbsc1-ABC data set (left) and MC-eNN simulations (right) of the same sequences (see Table 1). For each tetramer there is a different color legend. The four isodensity lines equal to 100 %, 75 %, 50 % and 25 % of the maximum density and the corresponding isodensity values are shown in the bottom right of each plot. The bi-dimensional inter base pair parameter distributions of MD and MC-eNN simulations

are indistinguishable even when correlated in a highly non-linear manner which is impossible to capture by a standard harmonic model.

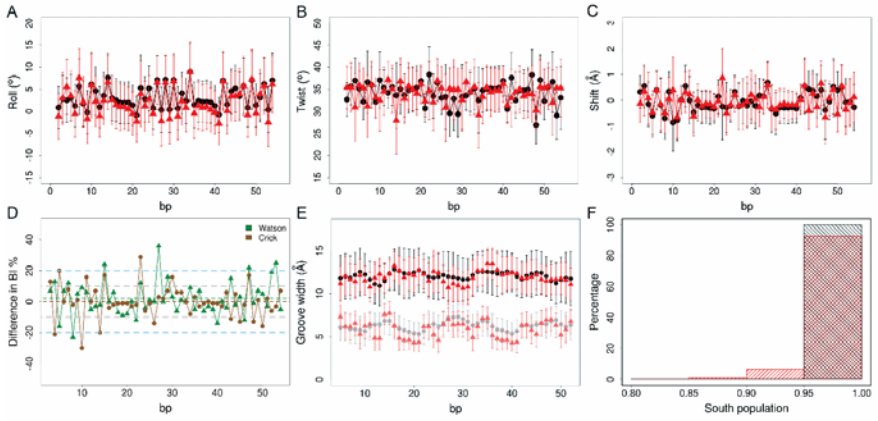


Figure 7. Comparison of MC-eNN (black) and MD simulations (red) of the longest naked DNA duplex in the BigNASim database (56 bp in length, sequence see Suppl. Table S4). A) Roll distribution in degrees, B) Twist distribution in degrees and C) Shift distribution in Angstrom of the central 53 bps. D) Difference in BI Percentage of backbone states of MC-eNN - MD in Watson (green) and Crick (brown) strand of the central 54 base pairs. The green and brown dashed line show the average difference in BI percentage in the Watson (2.2%) and Crick (0.1%) strand, the grey horizontal dashed lines illustrate the 10 % margin corresponding to the accuracy of determining the backbone state in NMR experiments (41) and blue horizontal dashed lines represent 20% difference in BI population similar to the average discrepancy of backbone state population estimations of MD simulations compared to NMR experiments (41). E) Major (top, in bold) and minor (bottom, transparent) groove width. F) Histogram of the population of South pucker (Phase angle of 120°-210°) of all the South/North (Phase angle of 340°-40°) pucker conformations of the central 54 base pairs. All the error bars of Fig7A-D represent the standard deviation.

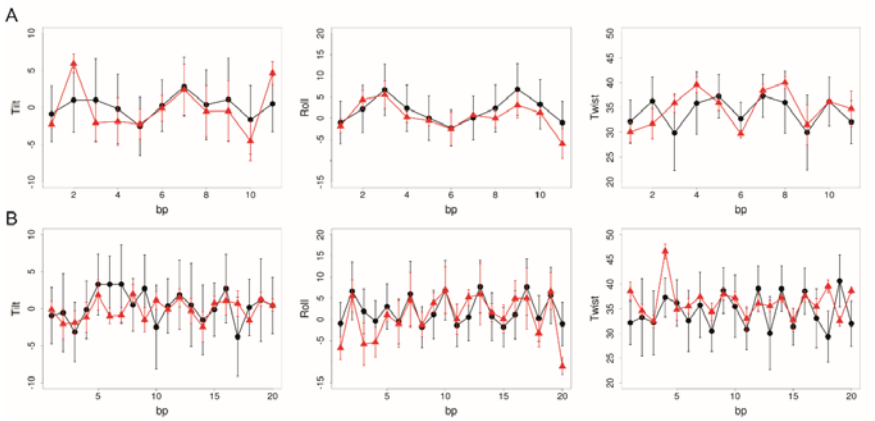


Figure 8. Comparison of the rotational inter base pair parameter distributions Tilt (left), Roll (middle) and Twist (right) of MC-eNN simulations (black) with experimental structures in PDB (red). Error bars represent the standard deviation of the MC-eNN simulation or the different models of the experiment, respectively. A) PDB id 1ILC (12 bp, resolved by NMR). B) PDB id 2JYK (21 bp, resolved by NMR). The translational inter base pair parameter distributions are compared in Suppl. Figure S9 and more examples are depicted in Suppl. Figure S10 (see Suppl. Table S5 for more details on the experimental structures).

Supplementary Information

A MULTI-MODAL COARSE-GRAIN MODEL OF DNA FLEXIBILITY MAPPABLE TO THE ATOMISTIC LEVEL

Jürgen Walther[1], Pablo D. Dans[1], Alexandra Balaceanu[1], Adam Hospital[1], Genís Bayarri[1] and Modesto Orozco[1,2]*

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology.
[2] Department of Biochemistry and Biomedicine. The University of Barcelona. Barcelona, Spain.

* Correspondence to Prof. Modesto Orozco: modesto.orozco@irbbarcelona.org

Supplementary Methods

Stand-alone version

The Stand-alone version of the MC-enNN model can be downloaded via http://mmb.irbbarcelona.org/MCDNAlite/standalone. The executables are available for Linux and MacOS systems (source code available upon request to the authors).

The stand-alone version calculates an ensemble of DNA configurations simulated via a Monte Carlo algorithm. The result is a set of DNA configurations which can be visualized according to the desired level of resolution. There are two options of resolution:

1) Only nucleobase atoms and the phosphate are reconstructed of each DNA structure. ("coarse grain")
2) Nucleobase atoms and the whole backbone are reconstituted ("atomistic")

Note that to obtain only information about the conformation of the nucleobases, the "coarse grain" option is sufficient, otherwise the "atomistic" option is recommended.

Once downloaded, extract the tar.gz file. In the folder there are two relevant executable "run_cg.sh" for option 1 and "run_atomistic.sh" for option 2.

The user needs to specify three parameters: the input sequence, the number of structures to be calculated and the output folder in which the results are saved. A sequence file 'Test_Sequence.dat' is available in the main folder to test the algorithm. Generally, sequence files have to be written in a single line with no spaces and all upper case letters using only 'A', 'C', 'G' or 'T'. The minimum sequence length is 5nt.

Execute 'coarse-grain' model. To execute the model with 'coarse grain' resolution following command has to be typed into the console (change directory in the console to the directory where the executable are located):

sh run_cg.sh <absolute path of sequence file> <# of structures to generate> <output folder>

Example:

sh run_cg.sh /home/Directory/Test_Sequence.dat 10 /home/test (adjust directory path as needed)

This command calculates 10 structures in the 'coarse-grain' resolution taking the sequence from '/home/Directory/Test_Sequence.dat'. The output is saved in '/home/test'.

The structures are saved as pdb in the output folder in the folder called 'output_pdb' and the bps coordinates of each structure are saved in the folder 'output_helpar' (the values in the first line in the file 'table_all_twis.dat' correspond to the Twist values of bps 1 to N of the first structure and so on).

Execute 'atomistic' model. To execute the model with 'atomistic' resolution two prerequisites have to be taken into account.

NOTE: Two prerequisites have to be fulfilled since freely available third party software is used for the atomistic backbone reconstruction.

1) R needs to be installed (https://www.r-project.org/). An R version equal or higher than 3.2.0 is necessary and the library 'bio3d' needs to be installed (this can be done via the command 'install.packages("bio3d")' after starting R). If this is not fulfilled the program cannot execute correctly.

2) Ambertools needs to be installed (http://ambermd.org/AmberTools.php). The environment variable AMBERHOME needs to be exported for the integrated tools to work correctly.

The model with 'atomistic' resolution executes the same way as the 'coarse grain' one by just changing the name of the executable (change directory in the console to the directory where the executables are located):

sh run_atomistic.sh <absolute path of sequence file> <# of structures to generate> <output folder>

Example:

sh run_atomistic.sh /home/Directory/Test_Sequence.dat 10 /home/test (adjust directory path as needed)

This command calculates 10 structures in the 'atomistic' resolution taking the sequence from '/home/Directory/Test_Sequence.dat'. The output is saved in '/home/test'.

IMPORTANT

(1) Depending on the version of R the function 'read.pdb()' from the bio3d package might not work properly. However, the function 'read.pdb2()' will work. If this is the problem change in the file Backbone_reconstruction/bkb_conf_helpars_function.R the name 'read.pdb' to 'read.pdb2'.

(2) For a new version of Ambertools, the file 'leaprc.ff14SB' might not be located in the folder '$AMBERHOME/dat/leap/cmd/leaprc.ff14SB'. The path in the file Backbone_reconstruction/rec_bkb.sh then needs to be changed from '$AMBERHOME/dat/leap/cmd/leaprc.ff14SB' to the correct path '$AMBERHOME/dat/leap/cmd/oldff/leaprc.ff14SB'.

The 'atomistic' structures are saved as a trajectory in AMBER format (out.mdcrd and out.top) in the output folder in the folder called 'output_pdb' and the bps coordinates of each structure are saved in the folder 'output_helpar' (the values in the first line in the file 'table_all_twis.dat' correspond to the Twist values of bps 1 to N of the first structure and so on).

Note: Due to its 'atomistic' resolution the trajectory 'out.mdcrd' can be analyzed by all common analysis tools for molecular dynamics simulations.

Execution time. The benchmarking of MC-enNN with MD simulations is shown in Supplementary Figure 11. On an ordinary laptop the 'atomistic' resolution model takes 1 hour to calculate 1,000 DNA structures of 18 bp in length while the 'coarse grain' resolution model lies in low minute range. Note that the execution time scales linearly with the number of structures. If the user wishes to simulate more structures with 'atomistic' resolution it is recommended to distribute the calculation on several cores. The time scales linearly with the number of cores since no communication between the cores is needed (10,000 structures would take 90 min on 8 cores).

Parallelization. To execute the MC-enNN model on several cores (for example 10,000 structures on 8 cores), the user needs to individually execute the program on each core simulating 10,000/8 = 1,250 structures each. After the simulation the generated trajectories can be combined using common MD analysis tools like Ambertools to obtain the final trajectory of 10,000 structures.

Number of structures. Even though the ensemble size has to be around 40,000 structures to fully recapitulate complex correlation between inter base pair parameters, a reduced size of

5,000-10,000 snapshots in the ensemble of DNA structures are usually already sufficient for general analysis purposes.

Bayesian Information Criterion (BIC), Bayes Factors, and the Helguerro's theorem. We used the BIC methodology to determine the optimal number of Gaussian function needed to fit a given distribution. This is done by finding the set of parameters that minimizes the BIC values (the model with the lower BIC is chosen) according to (1, 2):

$$-2\ln p(x \mid k) \approx BIC = -2\ln(L) + k\ln(n)$$

Where x are the observed data, k is the number of free parameters to be estimated, and $p(x|k)$ is the probability of the observed data given the number of parameters, or, in other words, the likelihood of the parameters given the dataset. L is the maximized value of the likelihood function for the estimated model, and n is the number of data points in x (the number of observations). In this work we limit the BIC to considering a maximum of two Gaussians, leading to the classification of each distribution as uninormal (fitted with one Gaussian) or binormal (fitted with a combination of two Gaussians). When a distribution is not classified as uninormal sometimes it is referred to as multi-normal or multi-peaked in the manuscript.

The Bayes Factors, that can be extracted from the BIC analysis, were used to determine the strength of the evidence in favour of the model chosen by BIC (see (3) for a detailed discussion). This lead to a third classification labelled as "insufficient evidence", when either of the two models determined with BIC (uninormal or binormal) couldn't be statistically supported.

Finally, when there was sufficient evidence to favour a binormal fitting, we used an extension of the Helguerro's theorem (4, 5) to define the modality of the distribution and distinguish the cases where the two peaks of the two fitted Gaussians are close together from those where they are significantly separated. This is the most important distinction in terms of understanding DNA dynamics. In the first case, for practical purposes, the use of a single Gaussian distribution may often be justified to represent the data (the overall distribution may be interpreted as binormal-unimodal), while it cannot be in the second (binormal-

bimodal distributions). When a distribution is not classified as unimodal sometimes it is referred to as multi-modal in the manuscript.

Supplementary Tables

Table S1. Classification of BI/BII backbone state. Percentage of all 136 unique tetramers in Watson and Crick strand showing the highest correlation of an inter base pair parameter with the backbone state.

Shift	Slide	Twist	Twist 5'
95.2	3.3	1.1	0.4

Table S2. Drivers of clustering of all 136 tetramers (in %). The inter base pair parameter of a tetramer with the highest range of mean values of its clustered distributions (using non-dimensionalization of the inter base pair parameters according to eq. 3) is termed driver of clustering of this tetramer.

Shift	Slide	Rise	Tilt	Roll	Twist
61.0	10.3	1.5	2.2	4.4	20.6

Table S3. Sequences of 18 bp used to compare essential dynamics between MC-enNN and MD (see Figure 5). MD simulations were 1-2 μs in length using the parmBSC1 force field. The ID corresponds to the name in the BigNASim database (http://mmb.irbbarcelona.org/BigNASim/).

Seq. number	ID	Watson strand (5'-3' direction)
1	1r4i	CCAGAACATCAAGAACAG
2	1zgw	GCAAATTAAAGCGCAAGA
3	AGCG	GCCGAGCGAGCGAGCGGC
4	AGCT	GCCTAGCTAGCTAGCTGC
5	CGTG	GCTGCGTGCGTGCGTGGC
6	lks1	GCCTATAAACGCCTATAA
7	lks2	CTAGGTGGATGACTCATT
8	lks3	CACGGAACCGGTTCCGTG
9	lks4	GGCGCGCACCACGCGCGG
10	muTCGA	GCGATCGATCGATCGAGC

Table S4. Sequences of 56merTIP3P and 42mer. The ID corresponds to the name in the BigNASim database (http://mmb.irbbarcelona.org/BigNASim/)

ID	Watson strand (5'-3' direction)
Nl15	ATGGATCCACTGATACTACGACCAGAACATGATGTTCTCA
56merTIP3P	CGCCGGCAGTAGCCGAAAAAATAGGCGCGCGCTCAAAAAAATGCCCCATGCCGCGC

Table S5. PDB structures of B-DNA duplexes used to compare with the MC-enNN method. Resolution and RMSd/bp are given in Å. The RMSd of the MC-enNN ensemble with the average PDB structure was calculated considering all heavy atoms of the duplex without the flanking base pair on both ends.

PDB Code	Method	Resolution	Number models	RMSd/bp with MC-enNN	Sequence (Watson strand, 5'-3' direction)
1ILC	X-ray	2.2	3	0.38	ACCGAATTCGGT
2JYK	NMR	-	10	0.25	ACAGCTTATCATCGATCACGT
1NAJ	NMR	-	5	0.37	CGCGAATTCGCG
5F9I	X-ray	3	2	0.25	CCAATAATCGCGATTATTGG
424D	X-ray	2.7	1	0.38	ACCGACGTCGGT
1DN9	X-ray	2.2	1	0.41	CGCATATATGCG
1HQ7	X-ray	2.1	1	0.39	GCAAACGTTTGC

Supplementary Figures

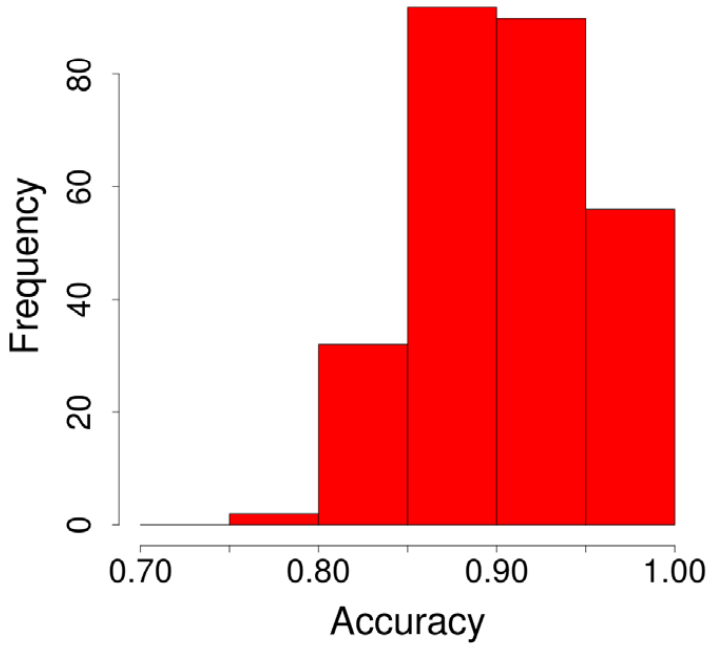


Figure S1. Accuracy of the backbone state (BI and BII) prediction of the nucleobases of the central base pair by the inter base pair parameter showing the highest correlation with the backbone states (see Suppl. Table S1 for the distribution of inter base pair parameters with highest correlation) for all unique 136 tetramers in Watson and Crick strand. In more than 85% of the cases the accuracy is above 0.85, the lowest accuracy is 0.78.

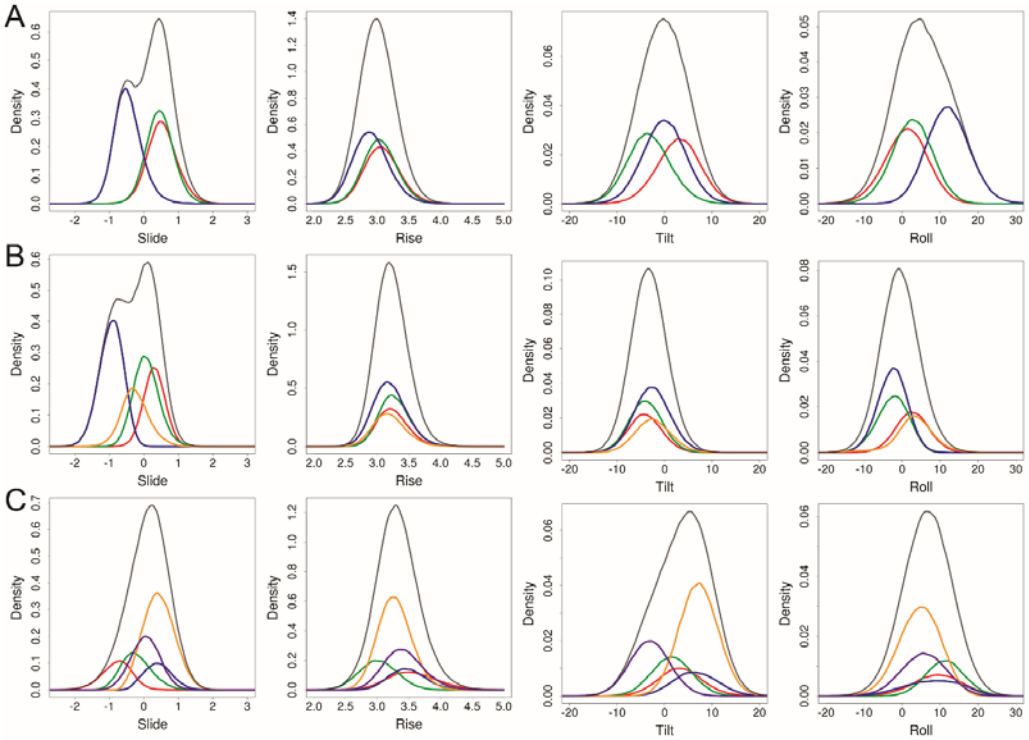


Figure S2. Examples for the division of inter base pair parameter distributions (Slide, Rise, Tilt and Roll; Shift and Twist are shown in Figure 3 of the main text) into multiple states for the most common number of clusters. A) CTAG (3 clusters). B) TAAG (4 clusters). C) ACGA (5 clusters). The inter base pair parameter distributions is shown in grey, the clustered distributions in green, blue and red for 3 clusters; green, blue, red and orange for 4 clusters and green, blue, red, orange and purple for 5 clusters.

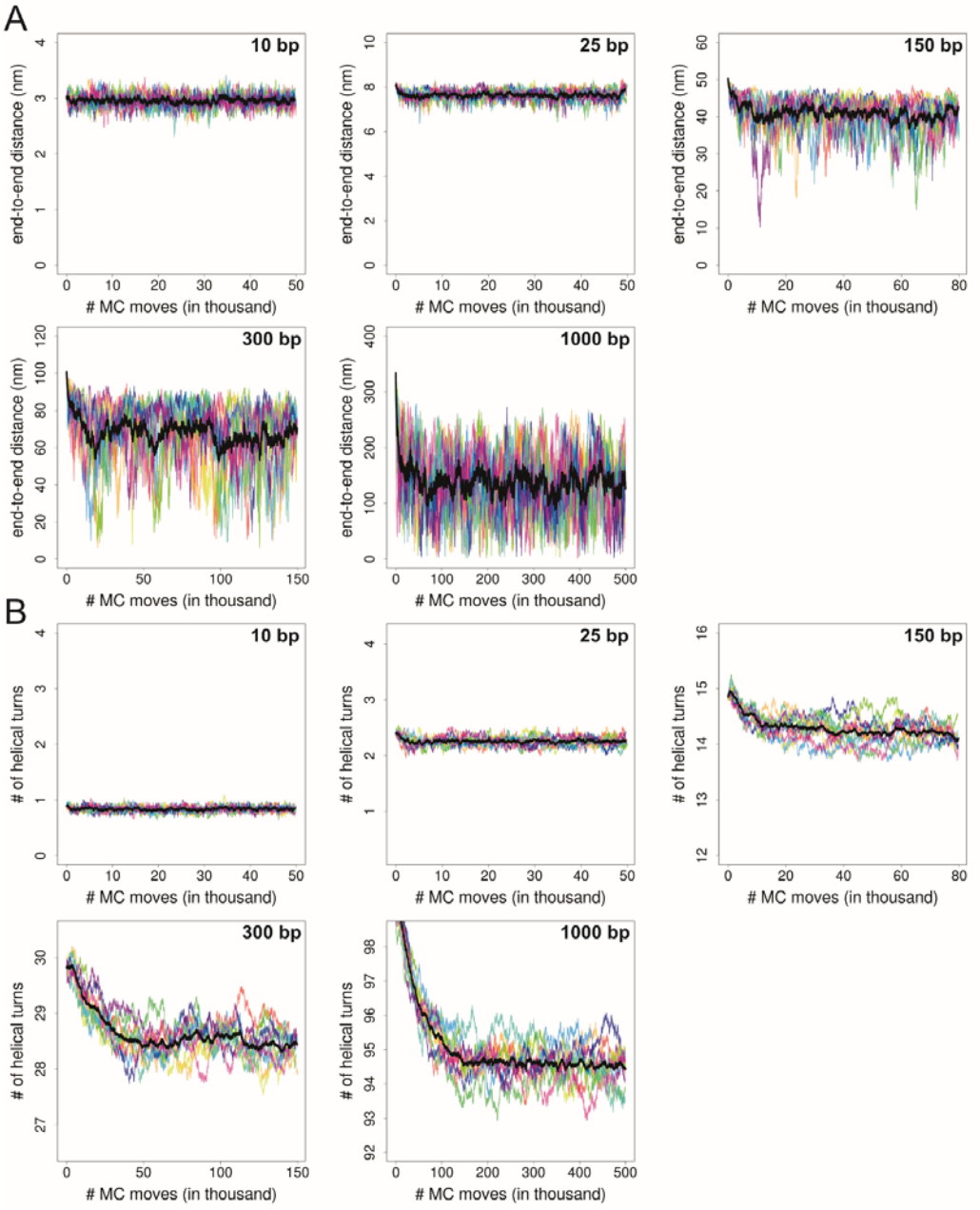


Figure S3 Equilibration of MC-enNN simulation. Number of MC moves needed for fiber equilibration of 10 individual simulations of a fiber of random sequence of 10, 25, 150, 300 and 1000 bp in length (results for fibers of 50 bp and 600 bp are shown in Figure 4A) by investigating two parameters: A) End-to-end distance. B) Number of helical turns.

The 10 individual simulations are shown in different colors and a black line illustrates the average of the 10 simulations. Note: the maximum of the scale of the axis of end-to-end distance is calculated as 4Å x fiber length (in base pair). The scale of the axis of # of helical turns differs for the fiber of 1,000 bp in length for better visualization.

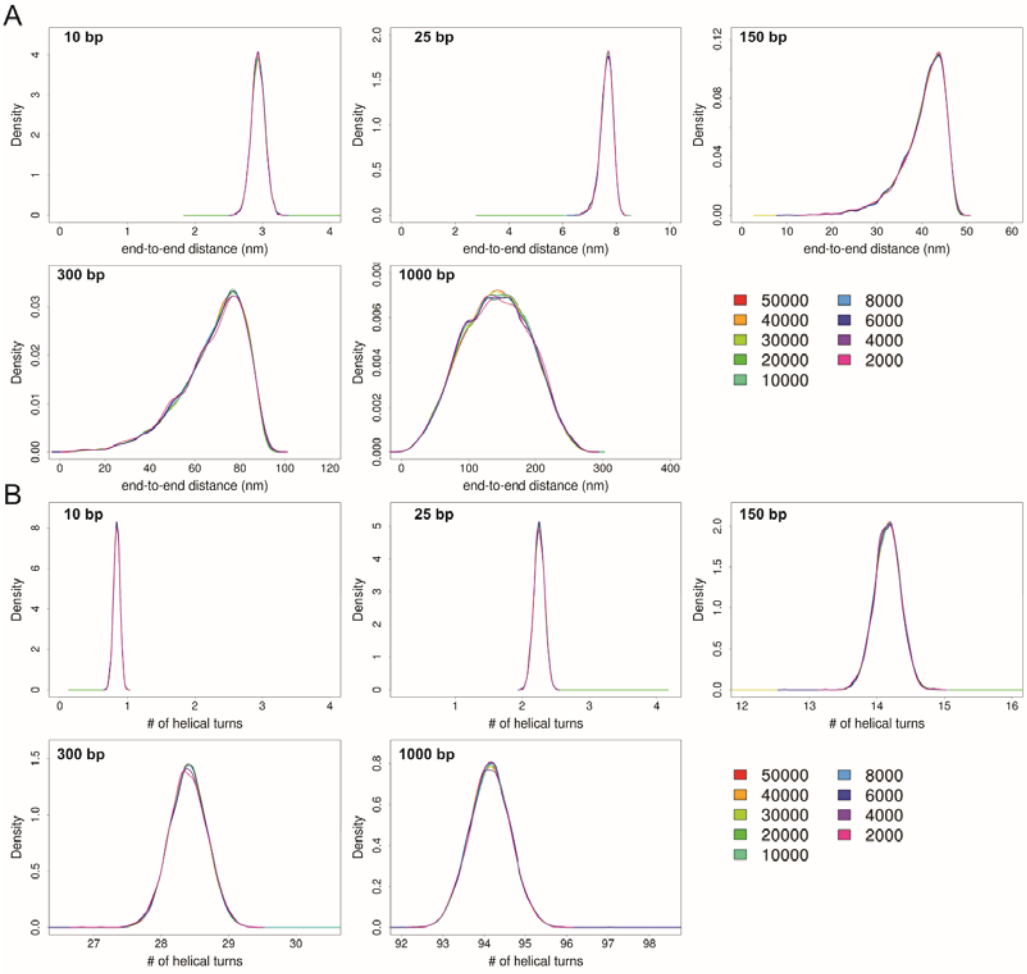


Figure S4. Convergence of MC-enNN simulation. Length of the ensemble needed to obtain converged distributions of a fiber of random sequence of 10, 25, 150, 300 and 1000 bp in length (results for fibers of 50 bp and 600 bp are shown in Figure 4B). A) End-to-end distance. B) Number of helical turns. Note: the maximum of the scale of the axis of end-to-end distance is calculated as 4Å x fiber length (in base pair). The scale of the axis of # of helical turns differs for the fiber of 1,000 bp in length to fit the scale as in Suppl. Figure S3B.

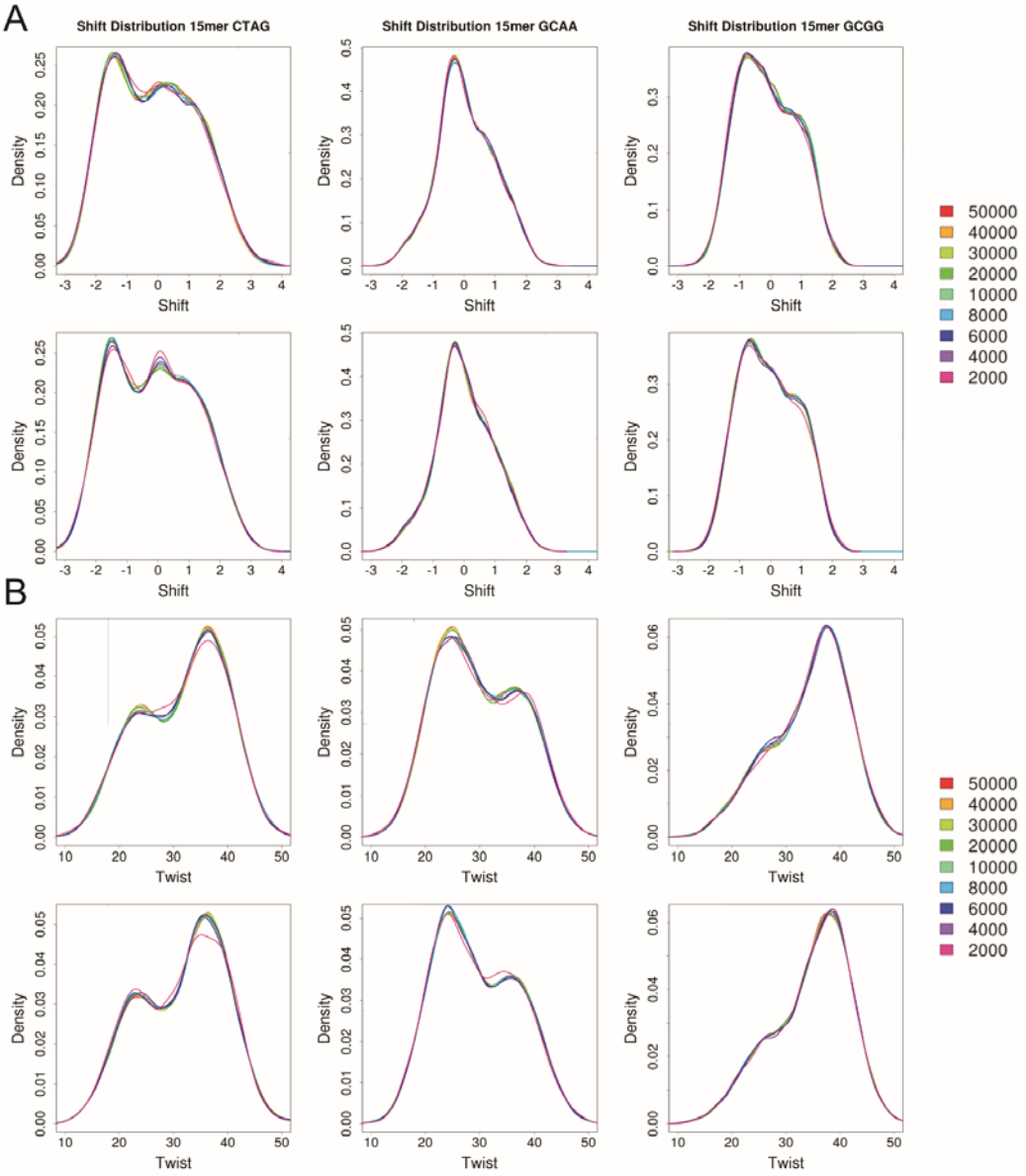


Figure S5. Dependence of convergence of complex inter base pair distributions of MC-enNN simulation on sequence length. A) Shift distribution of CTAG (left), GCAA (middle) and GCGG (right) of a simulation of a random sequence of 15 bp (top) and (B) 150 bp in length containing those three tetramers. B) Twist distribution of CTAG (left), GCAA (middle) and GCGG (right) of a simulation of a random sequence of 15 bp (top) and (B) 150 bp in length containing those three tetramers.

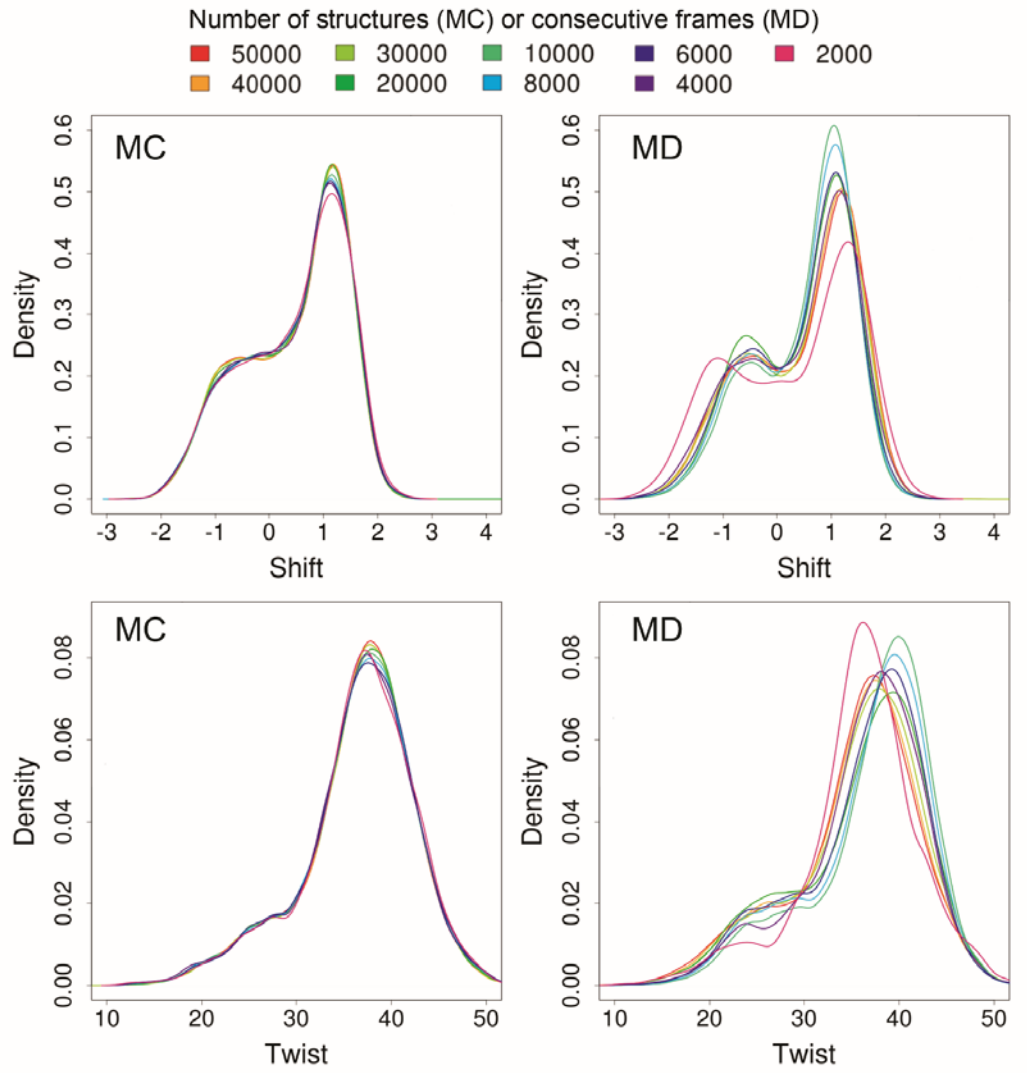


Figure S6. Comparison of convergence of the complex Shift (top) and Twist (bottom) distributions of the tetramer ACGA of MC-enNN simulation (left) with MD simulation (right) of a sequence of 40 bp in length. (BigNASim ID Nl15 in Suppl. Table S4). The number of structures in the MC ensemble are compared against the number of consecutive frames (1 frame every 10 ps) in the MD simulation.

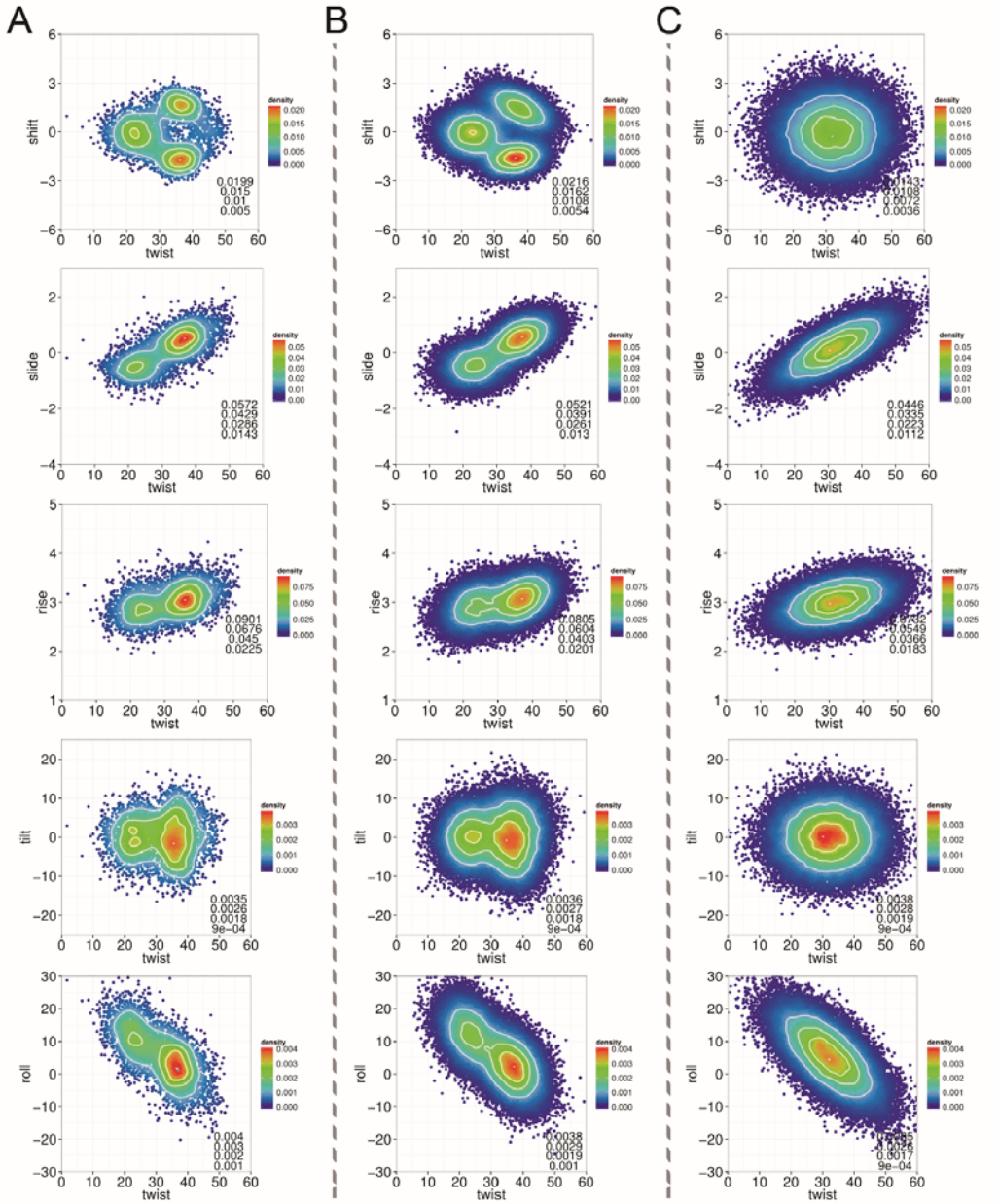


Figure S7. Bi-dimensional inter base pair parameter maps of Twist vs. the other five parameters of the tetramer CTAG of a simulation of sequence 13 of Table1 in the main text. A) MD simulation. B) MC-enNN simulation. C) MC-enNN simulation with one state per tetramer ('Standard harmonic approach', see eq. 2 in the manuscript). Note: The color legends in a single row are in the same scale and each row has a different color legend. The

four isodensity lines equal to 100 %, 75 %, 50 % and 25 % of the maximum density and the corresponding isodensity values are shown in the bottom right of each plot.

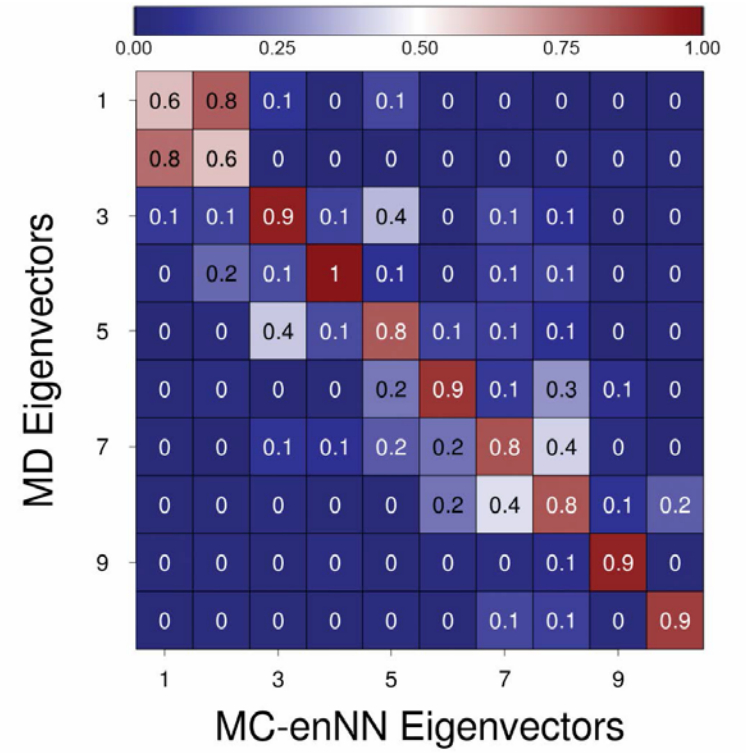


Figure S8. Matrix of inner product of the first 10 EV's of MC-enNN (x-axis) and MD (y-axis) simulations of a sequence of 56 base pairs in length (see Suppl. Table S4). As reference the heavy nucleobase atoms of the central 52 base pairs were used for PCA. Note: The numbers in each square of the matrix are rounded to a single decimal number and each square of the matrix is color-coded according the color legend.

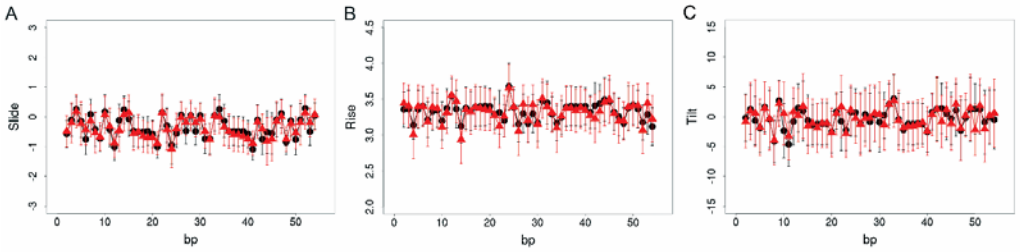


Figure S9. Comparison of MC-enNN (black) and MD simulations (red) of the longest naked DNA duplex in the BigNASim database (56 bp in length, sequence see Suppl. Table S4). A) Slide distribution in Angstrom, B) Rise distribution in Angstrom and C) Tilt distribution in degrees of the central 53 bps.

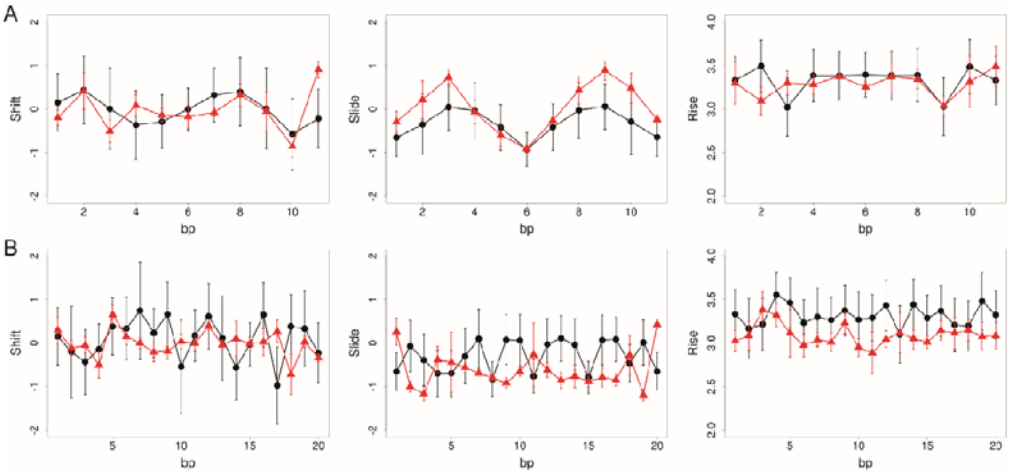


Figure S10. Comparison of the translational inter base pair parameter distributions Tilt (left), Roll (middle) and Twist (right) of MC-enNN simulations (black) with experimental structures in PDB (red). Error bars represent the standard deviation of the MC-enNN simulation or the different models of the experiment, respectively. A) PDB id 1ILC (12 bp, resolved by NMR). B) PDB id 2JYK (21 bp, resolved by NMR). The rotational inter base pair parameter distributions are compared in Figure 8 of the main text (see Suppl. Table S5 for more details on the experimental structures).

Supplementary References

1. Schwarz,G. (1978) Estimating the Dimension of a Model. *Ann. Stat.*, **6**, 461–464. https://doi.org/10.1214/aos/1176344136

2. Kass,R.E. and Raftery,A.E. (1995) Bayes Factors. *J. Am. Stat. Assoc.*, **90**, 773–795. https://doi.org/10.1080/01621459.1995.10476572

3. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–78. https://doi.org/10.1093/nar/gks884 http://www.ncbi.nlm.nih.gov/pubmed/23012264

4. Schilling,M.F., Watkins,A.E. and Watkins,W. (2002) Is Human Height Bimodal? *Am. Stat.*, **56**, 223–229. https://doi.org/10.1198/00031300265

5. DE HELGUERO,D.D.F. (1904) SUI MASSIMI DELLE CURVE DIMORFICHE. *Biometrika*, **3**, 84–98. https://doi.org/10.1093/biomet/3.1.84

2.2 Web Implementation of the helical coarse grain model (Publication 4)

To give the user a more complete experience of the MC-eNN model we decided to implement it in a large-scale computational environment where simulations and subsequent analysis of the sampled configurations are directly carried out without any interference needed by the user. The user only needs to specify the DNA sequence, if the structure at the free energy minimum or a set of equilibrium conformations should be the output and few parameters related to the type of simulation (workflow see Figure 28).

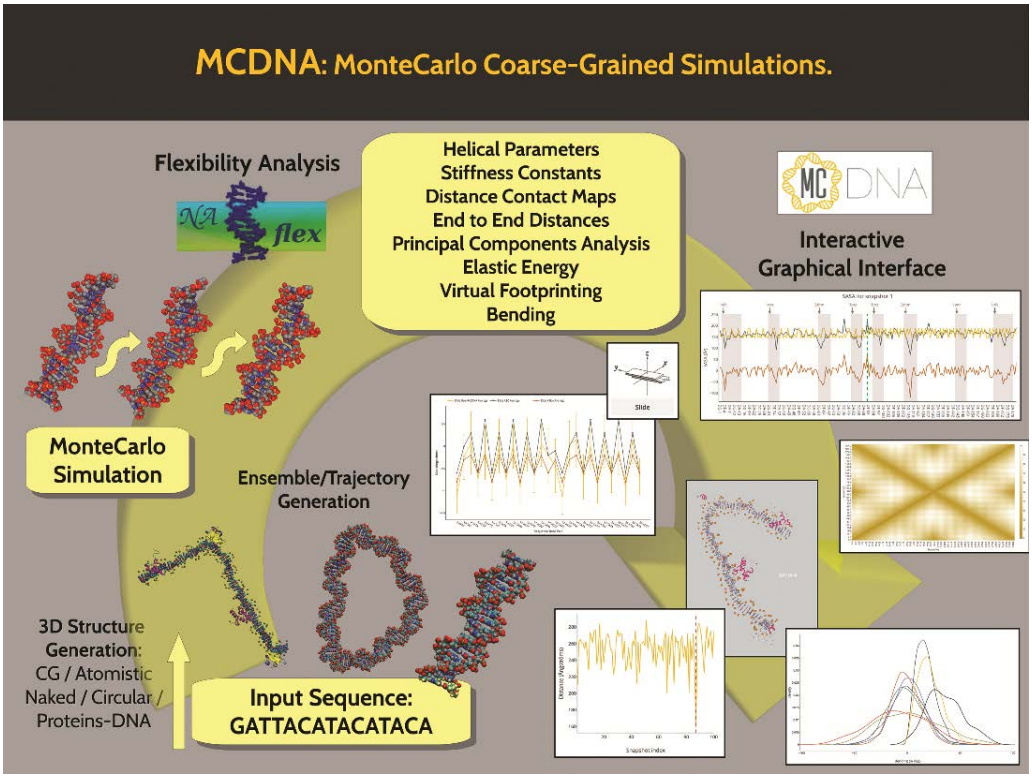


Figure 28. General workflow of the MCDNA webserver.

The webserver, named MCDNA, offers the user simulation of free B-DNA, circular DNA and protein-coated DNA. The core of the sampling method is via the above described method, however in constrained environments additional care has to be taken. Circular DNA is simulated as free DNA glued together at the end and the sampling algorithm is based on recursive stochastic closure (RSC; (13)) Monte Carlo moves in the inter base pair space (see Supplementary Methods in the following publication). Different degrees of supercoiling can be introduced by the user to

mimic superhelical density found inside the cell. Concerning simulations of protein-coated DNA, we did an exhaustive search of protein-DNA complexes in the PDB data bank and imposed the helical structure on the DNA where the protein is bound. The user only needs to provide the PDB code of the protein(s) and the position(s) of DNA where the protein(s) is(are) placed. Alternatively, the user can scan for the region(s) of DNA which are better shaped to adopt the bioactive conformation (see Figure 29).

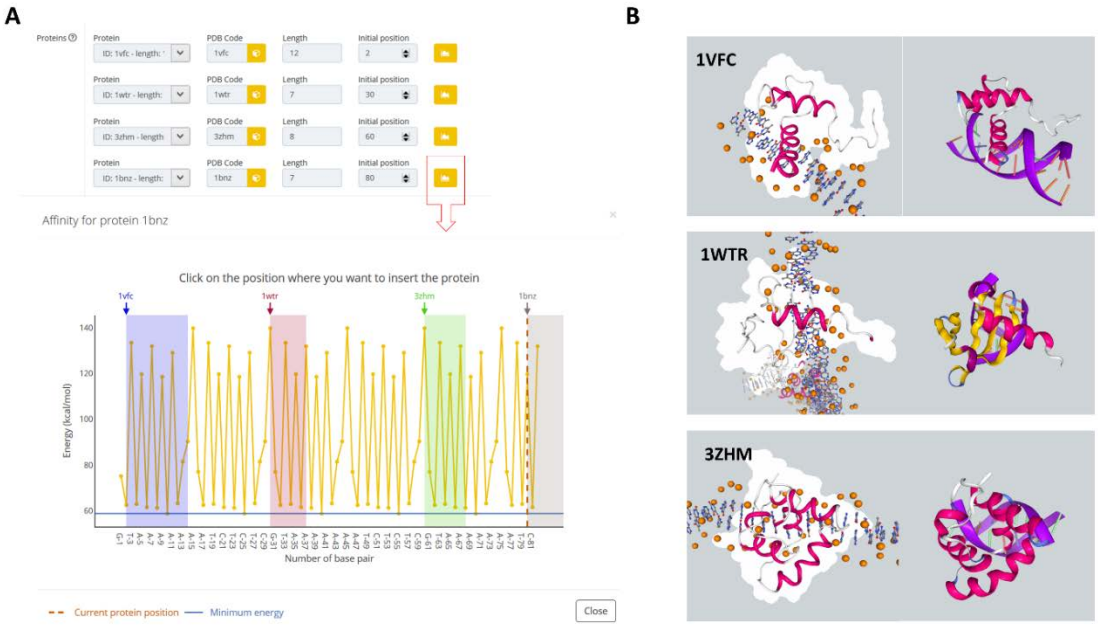


Figure 29. Details on the placement of the proteins along the fiber. A: Fragment of the input form for the Protein-DNA method. A yellow box is attached to every input protein, offering the possibility to launch a protein affinity process to identify the most favorable regions of the sequence to position the protein structure. In order to avoid possible overlaps, the proteins already included in the fiber are highlighted in colored rectangles, taking into account the length of the sequence recognized by the protein. B: Examples of modeled protein-DNA complexes from MC_DNA. Modeled complexes (left), and especially the DNA fragment orientation, can be compared to the original PDB crystal (right).

The output information of MCDNA provides a summary about all the input parameters chosen for the simulation together with an interactive visualization of the generated ground state structure and trajectory. A more in-depth analysis provides a full description of DNA flexibility. The inventory of analyses performed within the server includes helical parameters, stiffness energy constants, distance contact maps (for DNA and proteins), end-to-end distances, DNA bending, circular descriptors, elastic energies and virtual DNA footprinting. Results are presented in a very

intuitive and friendly interface, exploiting interactivity when possible (see section below). A guided tour for each analysis tool helps the user to get started navigating through the analysis section.

Several examples of potential use of the tool can be viewed in the webserver (see Supplementary Figures S1-3 for inputs and a selection of some outputs in Supplementary Figures S4-6 in the following publication). The first example is the simulation of a free DNA duplex of 30 base pairs in length for which the user can explore general structural and dynamic features, groove geometries and end-to-end distances to evaluate circularization propensity offering and interactive interface that displays the structure together with and end-to-end distance plot. By navigating through the different structures the user can easily have a 3D view from the most extended to the most bent structure (see Supplementary Figure S4 in the following publication). A second example shows the simulation of a DNA minicircle, where the distance matrix highlights that long-range contacts exist in the presence of super-helicity and in a third example the simulation of a protein-coated DNA fiber, accessibility of the DNA fiber to nucleases can be tested via in silico footprinting and generated conformations can be examined for DNA-mediated protein-protein contacts. The server is accessible at https://mmb.irbbarcelona.org/MCDNA/.

Publication:

Jürgen Walther, Adam Hospital, Genís Bayarri, Felipe Cano, Marco Pasi, Victor López-Ferrando, J. Lluís Gelpí, Pablo D. Dans and Modesto Orozco; MC_DNA: A web server for the detailed study of the structure and dynamics of DNA and chromatin fibers, (in preparation)

MC_DNA: A web server for the detailed study of the structure and dynamics of DNA and chromatin fibers

Jürgen Walther[1&], Adam Hospital[1&], Genís Bayarri[1], Felipe Cano[1], Marco Pasi[2], Victor López-Ferrando[3], J. Lluís Gelpí[3,4], Pablo D. Dans[1] and Modesto Orozco[1,4*]

juergen.walther@irbbarcelona.org, adam.hospital@irbbarcelona.org, genis.bayarri@irbbarcelona.org, felipe.cano@estudiant.upc.edu, marco.pasi@ens-cachan.fr, victor.lopez.ferrando@bsc.es, gelpi@ub.edu, pablo.dans@irbbarcelona.org, modesto.orozco@irbbarcelona.org

We present MC_DNA, a new web tool for the three-dimensional simulation of free DNA and medium-sized chromatin fibers. The program implements a Monte Carlo algorithm based on a mesoscopic model, using a tetramer-dependent base-pair step model fitted to reproduce parmbsc1 atomistic molecular dynamics (MD) simulations. The Monte Carlo ensembles can be projected to the atomistic level of resolution and processed to obtain *quasi*-time-dependent trajectories. The method provides ensembles of quality comparable to those obtained from atomistic MD, but at a tiny fraction of the computational cost, allowing to study systems much larger than those explored by atomistic MD. The trajectories (at atomistic or bp resolution levels) can be downloaded and/or subjected to a large variety of analysis in the server. All the tools are implemented in a friendly web interface where the user needs to specify only the DNA sequence, its topology (linear or circular) and whether the DNA fiber is free or protein-bound. The tool uses state-of-the-art technologies such as i) Open Nebula cloud infrastructure with virtual machines deployed on demand for computations, ii) WebGL-programmed NGL molecular viewer and the javascript plotly library for interactive plots, and iii) noSQL-MongoDB for storage. The server is accessible at http://mmb.irbbarcelona.org/MCDNA/.

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology.
[2] École normale supérieure (ENS) Paris-Saclay. Paris. France.
[3] Barcelona Supercomputing Center.
[4] Department of Biochemistry and Biomedicine. The University of Barcelona. Barcelona, Spain.

[&] Equally contributing authors.
* Correspondence to Prof. Modesto Orozco: modesto.orozco@irbbarcelona.org

MD simulations. Once collected, the coarse-grained ensemble (defined in a hybrid helical/Cartesian space) can be downloaded or analyzed "in situ" using a variety of tools. Furthermore, it can be converted into an atomistic ensemble for additional analysis or for generating inputs for further simulations. Both coarse-grained and atomistic ensembles can be transformed into a pseudo time-series (mimicking an MD trajectory) which can be visualized as an animation in an NGL viewer. The tool presented here is free and accessible without restrictions at (http://mmb.irbbarcelona.org/MCDNA/).

SIMULATION ENGINE

MC_DNA energy functional. The server uses a novel mesoscopic algorithm where the energy of the system is represented by a combination of short- (eq. 1) and long-range interactions (eq. 2). The short-range interactions are computed assuming that the distribution of rotational and translational parameters (rise, slide, shift, twist, roll, and tilt) in the helical space can be expressed as a Boltzmann-averaged combination of Gaussian distributions, which can be transformed into a simple (free) energy functional:

$$F_{short}(X) = -k_B T \ln \sum_{i=1}^{n} e^{-\frac{1}{k_B T}(K_i \Delta X_i^2 + E_i)} \qquad \text{(eq. 1)}$$

where the sum extends for all the Gaussians required to fit the probability distribution (*i.e.* the number of individual minima in the energy space), k_B is the Boltzmann constant, T is the absolute temperature, K_i is a 6x6 stiffness matrix obtained by inverting the covariance matrix centered at minimum i (11, 12), ΔX_i^2 is the displacement of the given conformation (X) with respect to the minimum energy conformation of minimum i (X_i^0), and E_i is the relative energy of minimum i with respect to the most stable minimum. Note that in the case of a unimodal-normal distribution (13, 14) the equation above reduces to the Olson-Zhurkin's functional: $F_{short}(X) = K_i \Delta X_i^2$ (11), which has been implemented in the current version of the server. Parameters defining the energy were obtained from µs-long parmbsc1 atomistic MD simulations of all unique bps in all the different tetramer environments (*i.e.* a next-nearest-neighbor model with 136 different unique tetramers), as determined by the ABC consortium (Dans *et al.*, to be published, data available at http://mmb.irbbarcelona.org/BigNASim (10), the set of sequences is shown in Supplementary Table 1).

In the current implementation of the server the treatment of long-range interactions is very simple (eq. 2), and is used only to avoid collision of two base-pairs:

$$E^{LJ} = 4\varepsilon^* \sum_{i,j} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \qquad \text{(eq. 2)}$$

where σ (σ = 13 Å) and ε* (ε* = k_BT) are the Lennard Jones parameters and r_{ij} stands for the distance of the center of base pair i with the center of base pair j. To speed up calculations for linear DNA, this term is considered only after a check if values of the bending of the fiber suggest that the structure could potentially overlap. E^{L-J} is calculated only if r_{ij} is below 20 nm. To not double count steric clashes, the long-range term is evaluated for every n^{th} bp (with n = 7 in the current implementation of the server).

MC_DNA sampling. Ensembles are obtained by using as Metropolis movements the three rotations and three translations of bps at a randomly selected number of places in the fiber. The positions of the phosphates are determined by taking Lavery's MD-averaged phosphate positioning relative to the helical axis (15). Regions occupied by proteins are assumed to be rigidified using the bps parameters of the DNA from the protein-DNA complexes in the PDB (Protein Data Bank) and accordingly, are not included as part of the active space in the generation of random movements. A computationally efficient version of Minary and Levitt's recursive stochastic closure algorithm (see Suppl. Methods and (16)) was used for the sampling of conformational space of mini-circles.

From coarse-grained ensembles to trajectories. The Metropolis algorithm guarantees that the collected snapshots define a Boltzmann ensemble, but the Markov chain derived from an MC run does not define a time-dependent series, something useful for obtaining dynamic pictures of DNA flexibility like those provided by MD-derived animations. Time-dependent pseudo-trajectories were built by re-ordering MC snapshots to guarantee that the step-to-step RMSd (the RMS deviation between snapshots at time t and t+Δt) follows a Maxwell-Boltzmann (MB) distribution for the entire trajectory. Accordingly, a randomly selected MC snapshot is taken as the 1^{st} step in the pseudo-trajectory and used as a reference to select the 2^{nd} one. The process is repeated until no snapshot matches the expected target step-to-step MB RMSd distribution. At this stage, we used a backtracking algorithm resuming the search back on time. This "revisiting" procedure is repeated many times until a minimum number of MC snapshots are left out of the pseudo-trajectory.

Either MC ensembles or pseudo-trajectories can be manipulated to gain atomistic detail of the DNA. For this purpose, the original mesoscopic "coarse-grained" trajectories (base-pair rotational and translational parameters for each base pair and phosphate positions in the Cartesian space) are transformed by using the SCHNArP algorithm (17) to generate the atomistic coordinates of the bases, taking as reference equilibrium geometries of the nucleobases (17). Finally, the backbone is generated using TLEAP Amber building capabilities (18) followed by a short steepest descent optimization to relax mismatched local geometries. Trajectories obtained at a given level of resolution can be downloaded or subjected to a variety of analysis as described below.

Performance of the simulation engine. We have analyzed the ability of the MC_DNA to reproduce atomistic MD simulations by comparing first the global characteristics of several 18-

mer duplexes already simulated at the atomistic level (parmbsc1 force-field) by the ABC consortium for more than 1 μs under physiological conditions (Dans *et al.*, to be published, data available at http://mmb.irbbarcelona.org/BigNASim/). As shown in Figure 2, MC_DNA simulations reproduce very well both general and detailed characteristics of the atomistic trajectories for all the duplexes studied, including the very long ones, where any error in the calibration of the model would result in dramatic errors. Good results are also obtained in the simulation of mini-circles, even for the ones with high torsional stress (Figure 3A). Finally, both global and local details of DNA-protein complexes are well represented (Figure 3B,C), suggesting that the method can be safely used to study long chromatin fibers. In summary, MC_DNA provides reasonable ensembles of DNAs (very often difficult to distinguish from atomistic MD structures) for a fraction of the computational cost of the MD simulations. For example, the 0.4 μs long atomistic trajectory of the 56-mer used here (see Figure 2A-D) required 650,000 CPU hours in the Tier-0 MareNostrum IV supercomputer, and an atomistic trajectory of just 13 ns of the 260-mer mini-circle (19) used 40,000 CPU hours in the Tier-0 Archer supercomputer. The MC_DNA calculations for these same systems were done in 3 minutes (56-mer) and 7 hours (260-mer mini-circle) using a medium-range desktop computer.

SERVER IMPLEMENTATION AND USAGE

Technology. MC_DNA is a web portal implemented with Slim PHP micro-framework (https://www.slimframework.com/) following a Model-view-controller (MVC) architectural pattern, supported by a MongoDB noSQL database (https://www.mongodb.org). NGL molecular viewer (20) is used to visualize 3D structures and trajectories, and plotly javascript package (https://plot.ly/) for modern data visualization was chosen to display all the analysis plots. Jobs are queued using SGE manager (http://gridscheduler.sourceforge.net/), and served in an on-demand processing model performed by Virtual Machines automatically deployed in an Open Nebula (21) OneFlow cloud environment for multi-tiered applications.

Input information. Starting from a DNA nucleotide sequence of up to 500 bases (in the current implementation), MC_DNA offers the possibility to study three types of nucleic acid systems: free linear DNA, circular DNA, and protein-bound DNA (see above). The user should select the desired level of resolution in the output (coarse-grained or atomistic) and the operations to be performed (generate a single structure or an ensemble/trajectory, number of structures, flexibility analyses, etc; see examples below). Depending on the type of DNA considered (free, protein-bound or circular), additional input parameters are needed. For example, the linking number is required for circular DNAs, and the identification code in the PDB (Protein Data Bank) of the protein(s) bound to DNA needs to be specified when simulating DNA-protein complexes.

The server offers the user the possibility to place the protein(s) at specific site(s) or scan the DNA to find the places where deforming the DNA to adopt the bioactive conformation is easier.

Upon submitting the job to the server, the user receives an URL address where he/she will find all the results of the simulation once it is completed. If the user provides an email address in the input form, he/she will be notified once the job is finished addressing him/her to the URL direction where the results of the job are shown/stored.

Output information. MC_DNA results are divided into three main sections:

- *Summary*: The summary section contains information about all the input parameters chosen for the job process, together with an NGL visualization of the generated structure and trajectory (if chosen).
- *Structure flexibility analysis / Trajectory flexibility analysis*: These two sections contain a set of flexibility analysis done on the generated structure and/or trajectory (if chosen). The list of analyses varies depending on the selected method and resolution and all together provide a full description of DNA flexibility. The inventory of analyses performed within the server includes helical parameters, stiffness energy constants, distance contact maps (for DNA and proteins), end-to-end distances, DNA bending, circular descriptors, elastic energies and virtual DNA footprinting. Results are presented in a very intuitive and friendly interface, exploiting interactivity when possible (see section below). A guided tour for each analysis tool helps the user to get started navigating through the analysis section.

Each of the results sections offers the possibility to download the specific analysis raw data in a compressed file for further analysis, or as a starting point for atomistic MD simulations using either local tools or our NAFlex server ((22), http://mmb.irbbarcelona.org/NAFlex/). Access to the web server is free, only an optional email address is requested to get a notification once the results are ready. Sample inputs and outputs are supplied to easily start getting familiar with the tool and its possibilities.

Examples of use. The server includes a few examples of the potential use of the tool to represent DNA ensembles (inputs in Suppl. Figures S1-3 and a selection of some of the outputs in Suppl. Figures S4-6). One prototypical example is the study of the dynamics of a medium-sized sequence of DNA (a 30-mer in the example, Suppl. Figure S1), for which the user can explore the general structural and dynamic features, the groove geometries (to explore possible binding pockets) and the end-to-end distances to evaluate the circularization propensity (Suppl. Figure 4A,B). For this last case, the server offers an interactive interface that displays the structure together with an end-to-end distance plot. Navigating through the top slider, the user can easily have a 3D view of the generated ensemble, from the most extended to the most bent structure (Suppl. Figure S4B). A second example demonstrates the potential of MC_DNA to

study the dynamics of a DNA mini-circle (see Suppl. Figure S2 for input), particularly the location of sharp distortions by analyzing the helical roll (compared to that expected from the ABC library and from our local database of X-Ray values, see Suppl. Figure S4C). The long-range contacts (Suppl. Figure S4D) highlight the existence of super-helicity (in this case a typical 8-shape pattern). In addition, the animation of the output trajectory in the server shows the population of the different conformations sampled for a given linking number. A final example illustrates the study of a short chromatin fiber, where the user first explores the best placement for proteins along the fiber (Suppl. Figure S5A), selecting optimal position(s) for the protein(s). The server models the complex by structurally superimposing the protein(s) on the protein-bound DNA conformation of the fiber (Suppl. Figure S5B). Then, the server analyses the obtained ensemble for accessibility of the DNA fiber to nucleases (in silico footprinting; see Suppl. Figure S6A), and for example, for the possibility of DNA-mediated protein-protein contacts (Suppl. Figure S6B). The server also allows the user to evaluate the suitability of the target DNA sequence to accommodate its structure to the "bioactive state" and how the distortion induced by the protein is distributed along the target DNA sequence (Suppl. Figure S6C).

ACKNOWLEDGEMENTS

We thank the ABC consortium for the atomistic MD simulations used to derive the parameters and all the colleges there for many discussions on the simulation engine. We thank Prof. Agnes Noy for the atomistic simulations on mini-circles, Prof. R. Lavery for Curves+ (23) and for his phosphate-location approach. We are also indebted to Prof. C. Laughton for his method to generate starting circular structures, and to J. Alcántara for setting up the hardware infrastructure behind the server. This work has been supported by the Spanish MINECO (BIO2015-64802-R; BFU2015-61670-EXP), the ERC Council (SimDNA 291433) and the H2020 projects BioExcel (675728) and MuG (676556). IRB Barcelona is recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain, GA SEV-2015-0500). P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SIN (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher.

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science (80-.).*, **316**, 1497–1502.
https://doi.org/10.1126/science.1141319
http://www.ncbi.nlm.nih.gov/pubmed/17540862

2. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721. https://doi.org/10.1038/nrg3788

3. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93. https://doi.org/10.1126/science.1181369 http://www.ncbi.nlm.nih.gov/pubmed/19815776

4. Cremer,M., Grasser,F., Lanctôt,C., Müller,S., Neusser,M., Zinner,R., Solovei,I. and Cremer,T. (2012) Multicolor 3D Fluorescence In Situ Hybridization for Imaging Interphase Chromosomes. In *Methods in molecular biology (Clifton, N.J.)*.Vol. 463, pp. 205–239. https://doi.org/10.1007/978-1-59745-406-3_15 http://www.ncbi.nlm.nih.gov/pubmed/18951171

5. Brunet,A., Tardin,C., Salomé,L., Rousseau,P., Destainville,N. and Manghi,M. (2015) Dependence of DNA Persistence Length on Ionic Strength of Solutions with Monovalent and Divalent Salts: A Joint Theory–Experiment Study. *Macromolecules*, **48**, 3641–3652. https://doi.org/10.1021/acs.macromol.5b00735

6. Dans,P.D., Walther,J. and Gómez,H. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45. https://doi.org/10.1016/J.SBI.2015.11.011

7. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58. https://doi.org/10.1038/nmeth.3658

8. Dans,P.D., Danilāne,L., Ivani,I., Dršata,T., Lankaš,F., Hospital,A., Walther,J., Pujagut,R.I., Battistini,F., Gelpí,J.L., *et al.* (2016) Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.*, **44**, 4052–4066. https://doi.org/10.1093/nar/gkw264 http://www.ncbi.nlm.nih.gov/pubmed/27084952

9. Dans,P.D., Ivani,I., Hospital,A., Portella,G., González,C. and Orozco,M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, gkw1355. https://doi.org/10.1093/nar/gkw1355 http://www.ncbi.nlm.nih.gov/pubmed/28088759

10. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M., *et al.* (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278. https://doi.org/10.1093/nar/gkv1301 http://www.ncbi.nlm.nih.gov/pubmed/26612862

11. Olson,W.K., Gorin,A.A., Lu,X.-J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci.*, **95**, 11163–11168. https://doi.org/10.1073/PNAS.95.19.11163 http://www.ncbi.nlm.nih.gov/pubmed/9736707

12. Lankaš,F., Šponer,J., Hobza,P. and Langowski,J. (2000) Sequence-dependent elastic properties of DNA. *J. Mol. Biol.*, **299**, 695–709. https://doi.org/10.1006/JMBI.2000.3781

13. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320. https://doi.org/10.1093/nar/gku809 http://www.ncbi.nlm.nih.gov/pubmed/25223784

14. Balaceanu,A., Pasi,M., Dans,P.D., Hospital,A., Lavery,R. and Orozco,M. (2017) The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28. https://doi.org/10.1021/acs.jpclett.6b02451

15. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–49. https://doi.org/10.1093/nar/gku504 http://www.ncbi.nlm.nih.gov/pubmed/24906882

16. Minary,P. and Levitt,M. (2010) Conformational optimization with natural degrees of freedom: a novel stochastic chain closure algorithm. *J. Comput. Biol.*, **17**, 993–1010. https://doi.org/10.1089/cmb.2010.0016 http://www.ncbi.nlm.nih.gov/pubmed/20726792

17. Lu,X.-J., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP). *J. Mol. Biol.*, **273**, 681–691. https://doi.org/10.1006/jmbi.1997.1345 http://www.ncbi.nlm.nih.gov/pubmed/9356256

18. Case,D.A., Cheatham,T.E., Darden,T., Gohlke,H., Luo,R., Merz,K.M., Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688. https://doi.org/10.1002/jcc.20290 http://www.ncbi.nlm.nih.gov/pubmed/16200636

19. Noy,A., Maxwell,A. and Harris,S.A. (2017) Interference between Triplex and Protein Binding to Distal Sites on Supercoiled DNA. *Biophys. J.*, **112**, 523–531. https://doi.org/10.1016/j.bpj.2016.12.034 http://www.ncbi.nlm.nih.gov/pubmed/28108011

20. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlić,A. and Rose,P.W. (2016) Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology - Web3D '16*. ACM Press, New York, New York, USA, pp. 185–186. https://doi.org/10.1145/2945292.2945324

21. Moreno-Vozmediano,R., Montero and Llorente,I.M. (2012) IaaS Cloud Architecture: From Virtualized Datacenters to Federated Cloud Infrastructures. *Computer (Long. Beach. Calif).*, **45**, 65–72. https://doi.org/10.1109/MC.2012.76

22. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47-55. https://doi.org/10.1093/nar/gkt378

http://www.ncbi.nlm.nih.gov/pubmed/23685436

23. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–29.
https://doi.org/10.1093/nar/gkp608
http://www.ncbi.nlm.nih.gov/pubmed/19625494

24. Pérez,A., Blas,J.R., Rueda,M., López-Bes,J.M., De La Cruz,X. and Orozco,M. (2005) Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.*, **1**, 790–800.
https://doi.org/10.1021/ct050051s
http://www.ncbi.nlm.nih.gov/pubmed/26641895

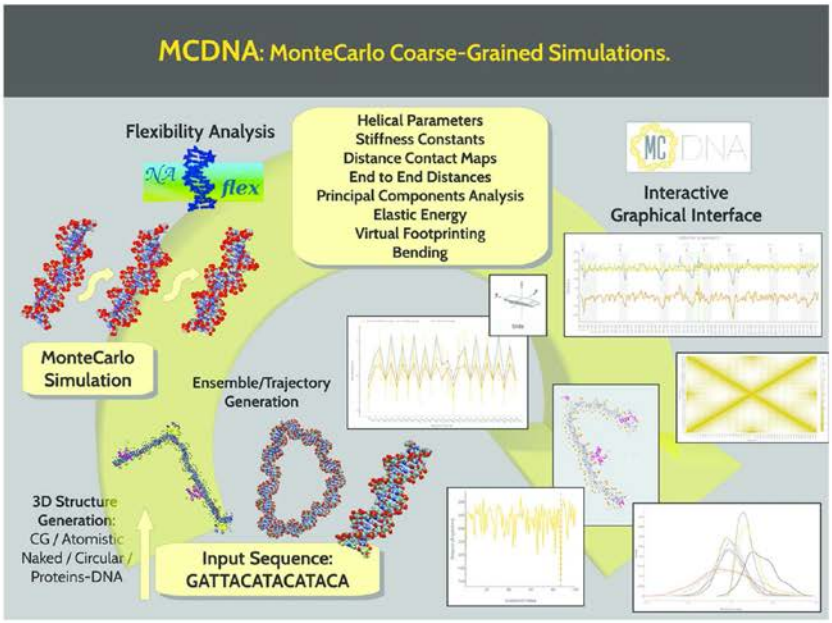
LEGENDS TO FIGURES

Figure 1. The general structure of the MC_DNA web server.

Figure 2. Examples of the performance of the MC_DNA simulation engine for different free DNA duplexes. (A) Mean and standard deviation of minor (left) and major (right) groove width of (atomistic-explicit solvent) MD (red) versus MC_DNA (blue) simulations along a 56-mer DNA duplex (sequence available at http://mmb.irbbarcelona.org/BigNASim (10) ID 'NAFlex_56merTIP3P'). The mean values of both simulations lie within 1 Å for each base-pair. (B) Distribution of total bending of all 5 bp (left) and 10 bp (right) segments of the 56-mer simulation. The overlap of the MC_DNA and MD distributions is shown in purple. (C) 2D scatter plot of all possible combinations of bps parameters of the 56-mer simulations capping the first and last base-pair. The overlap of the blue MC_DNA and red MD distributions is shown in purple. (D) The inner product of the first ten PCA eigenvectors (vertical MD, horizontal MC_DNA) of the 56-mer simulations. In each square, the value of the inner product is shown along with the color code (note the high values obtained in the diagonal). The Boltzmann's average absolute similarity index (calculated as the Boltzmann-weighted sum of the inner product of the first ten PCA eigenvectors (24)) between the MC_DNA and the MD trajectory is 0.88, indistinguishable from the value obtained when the 1st and 2nd part of the MD trajectory are compared. (E) Results for the central 12-mer of several 18-mer duplexes (sequences available in Suppl. Table 1). Top: helical turns; middle: total bending, and bottom: RMSD of the atomic coordinates of the bases with respect to the average structure of the MD trajectory. (F) Values of Tilt, Roll and Twist (from left to right) along the sequence of the 36mer duplex (sequence available at http://mmb.irbbarcelona.org/BigNASim (10) ID 'NAFlex_36merTIP3P').

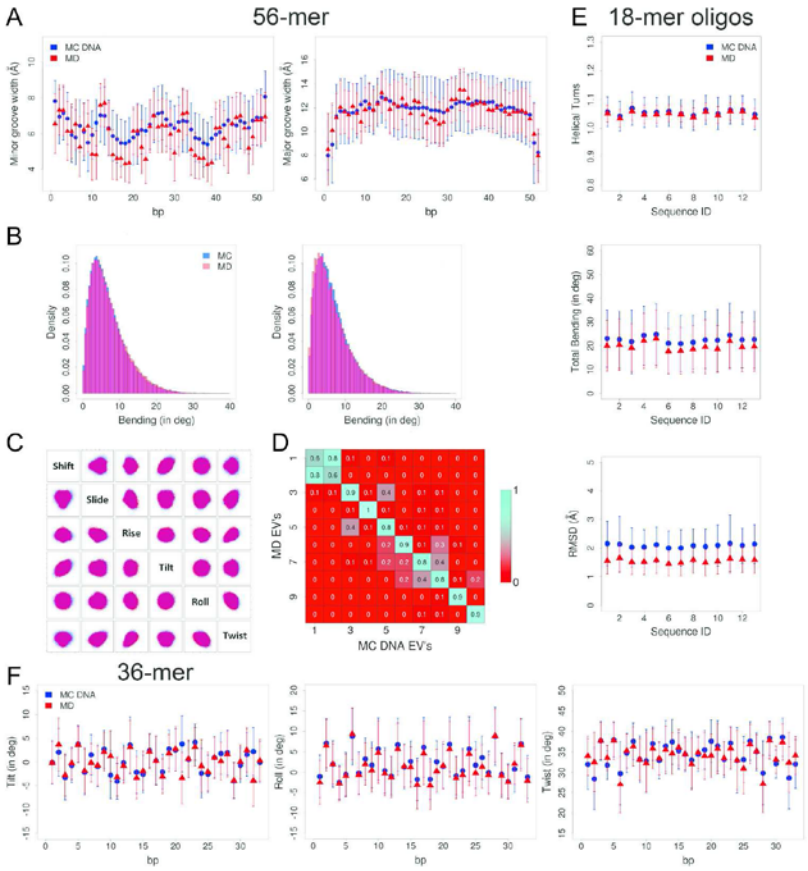
Figure 3. Examples of the performance of the MC_DNA simulation engine for circular DNA and for DNA-protein complexes. (A) (Top) Average and standard deviation of Writhe and Twist of the simulations of the 260-mer mini-circle at different ΔLk by MD (red) and by MC_DNA (blue). Twist and Writhe values of the MD simulations are taken from (19), Twist in the MC_DNA simulations is the cumulative sum of the Twist value of all base-pair steps, the Writhe then is determined as ΔLk-Twist. Representative structures (bottom) of MD and MC_DNA runs for the different ΔLk show a very similar pattern. (B) Average and standard deviation of Radius of Gyration (top) and RMSd (bottom) for the four protein-DNA complexes studied (1TRO, 2DGC, 3JXC, 3TQ6) with both MC_DNA (blue) and Molecular Dynamics (red). MD simulations used correspond to 2 μs long trajectories containing both protein and DNA. MC_DNA simulations used contain 500 snapshots generated with the MC_DNA web server, restricting the movements for the base pair steps in contact with the protein. (C) RMSd of Roll (left) and Twist (right) profiles obtained from MD trajectories, MC_DNA ensembles, and the crystal structure.

Note that the difference between MD and MC_DNA profiles is typically smaller than the difference between the MD simulation and the experiment.



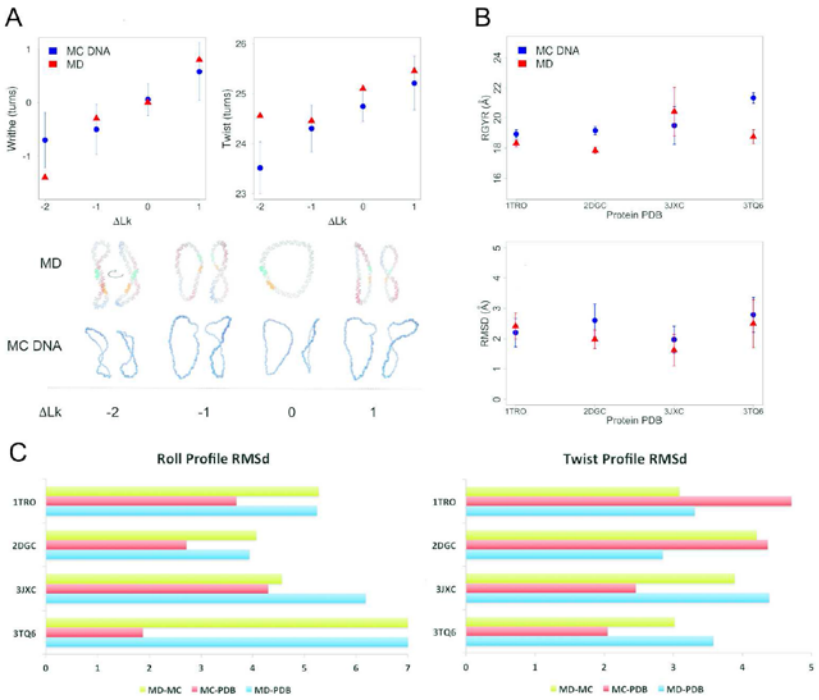
The general structure of the MC_DNA web server.

94x71mm (300 x 300 DPI)



Examples of the performance of the MC_DNA simulation engine for different free DNA duplexes. (A) Mean and standard deviation of minor (left) and major (right) groove width of (atomistic-explicit solvent) MD (red) versus MC_DNA (blue) simulations along a 56-mer DNA duplex (sequence available at http://mmb.irbbarcelona.org/BigNASim (10) ID 'NAFlex_56merTIP3P'). The mean values of both simulations lie within 1 Å for each base-pair. (B) Distribution of total bending of all 5 bp (left) and 10 bp (right) segments of the 56-mer simulation. The overlap of the MC_DNA and MD distributions is shown in purple. (C) 2D scatter plot of all possible combinations of bps parameters of the 56-mer simulations capping the first and last base-pair. The overlap of the blue MC_DNA and red MD distributions is shown in purple. (D) The inner product of the first ten PCA eigenvectors (vertical MD, horizontal MC_DNA) of the 56-mer simulations. In each square, the value of the inner product is shown along with the color code (note the high values obtained in the diagonal). The Boltzmann's average absolute similarity index (calculated as the Boltzmann-weighted sum of the inner product of the first ten PCA eigenvectors (24)) between the MC_DNA and the MD trajectory is 0.88, indistinguishable from the value obtained when the 1st and 2nd part of the MD trajectory are compared. (E) Results for the central 12-mer of several 18-mer duplexes (sequences available in Suppl. Table 1). Top: helical turns; middle: total bending, and bottom: RMSD of the atomic coordinates of the bases with respect to the average structure of the MD trajectory. (F) Values of Tilt, Roll and Twist (from left to right) along the sequence of the 36mer duplex (sequence available at http://mmb.irbbarcelona.org/BigNASim (10) ID 'NAFlex_36merTIP3P').

583x615mm (149 x 149 DPI)



Examples of the performance of the MC_DNA simulation engine for circular DNA and for DNA-protein complexes. (A) (Top) Average and standard deviation of Writhe and Twist of the simulations of the 260-mer mini-circle at different ΔLk by MD (red) and by MC_DNA (blue). Twist and Writhe values of the MD simulations are taken from (19), Twist in the MC_DNA simulations is the cumulative sum of the Twist value of all base-pair steps, the Writhe then is determined as ΔLk-Twist. Representative structures (bottom) of MD and MC_DNA runs for the different ΔLk show a very similar pattern. (B) Average and standard deviation of Radius of Gyration (top) and RMSd (bottom) for the four protein-DNA complexes studied (1TRO, 2DGC, 3JXC, 3TQ6) with both MC_DNA (blue) and Molecular Dynamics (red). MD simulations used correspond to 2 μs long trajectories containing both protein and DNA. MC_DNA simulations used contain 500 snapshots generated with the MC_DNA web server, restricting the movements for the base pair steps in contact with the protein. (C) RMSd of Roll (left) and Twist (right) profiles obtained from MD trajectories, MC_DNA ensembles, and the crystal structure. Note that the difference between MD and MC_DNA profiles is typically smaller than the difference between the MD simulation and the experiment.

556x474mm (149 x 149 DPI)

SUPPLEMENTARY MATERIAL

MC_DNA: A web server for the detailed study of the structure and dynamics of DNA and chromatin fibers

Jürgen Walther[1&], Adam Hospital[1&], Genís Bayarri[1], Felipe Cano[1], Marco Pasi[2], Victor López-Ferrando[3], J. Lluís Gelpí[3,4], Pablo D. Dans[1] and Modesto Orozco[1,4*]

juergen.walther@irbbarcelona.org, adam.hospital@irbbarcelona.org, genis.bayarri@irbbarcelona.org, felipe.cano@estudiant.upc.edu, marco.pasi@ens-cachan.fr, victor.lopez.ferrando@bsc.es, gelpi@ub.edu, pablo.dans@irbbarcelona.org, modesto.orozco@irbbarcelona.org

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology.
[2] École normale supérieure (ENS) Paris-Saclay. Paris. France
[3] Barcelona Supercomputing Center.
[4] Department of Biochemistry and Biomedicine. The University of Barcelona. Barcelona, Spain.

[&] Equally contributing authors.
* Correspondence to Prof. Modesto Orozco: modesto.orozco@irbbarcelona.org

SUPPLEMENTARY METHODS

Recursive Stochastic Closure Algorithm (RSC). In a usual Monte Carlo move in MC_DNA base-pair step (bps) coordinates of a bps are changed inducing a rigid body move of the DNA above the changed bp step. In the constrained circular model the rigid body move is done only for a segment of DNA of n base-pairs in length above the changed bps. The procedure of the segmental Monte Carlo move is explained in Supplementary Figure S7A. The initial Monte Carlo move is done on bps i. The orientations of bp i+2 until i+n are kept as before the Monte Carlo move. This is compensated by a change in rotational bps parameters of step i+1. We found this to be more efficient than distributing the distortion induced by the MC move in each bps. In the final step the recursive stochastic closure (RSC) algorithm adapted from ref. 16 in the manuscript is applied to close the circle. This is done by successive stages of stochastic partial closure that propagate the location of the chain break at i+n backwards along the chain. In our adapted version of the RSC algorithm, first the position of the base pair i+n is set to its position before the MC move (Suppl. Figure S7A). Then the stochastic partial closure (SPC) algorithm is applied. In the first SPC move a line is defined connecting bp i+n-2 and i+n (Suppl. Figure S7B). Then the intersection point of a sphere of radius r (r is the distance between bp i+n-1 and i+n before the chain break) around bp i+n with the line is determined. The new position of bp i+n-1 is then chosen by randomly selecting a point from a two-dimensional normal distribution around the intersection point on the surface of the sphere. In the original code the position of the new position of bp i+n-1 is chosen on a tangential plane of the surface, however we place the point directly on the surface of the circle since the point selected is usually very close to the anchor point on the surface of the sphere which makes our implementation of the code

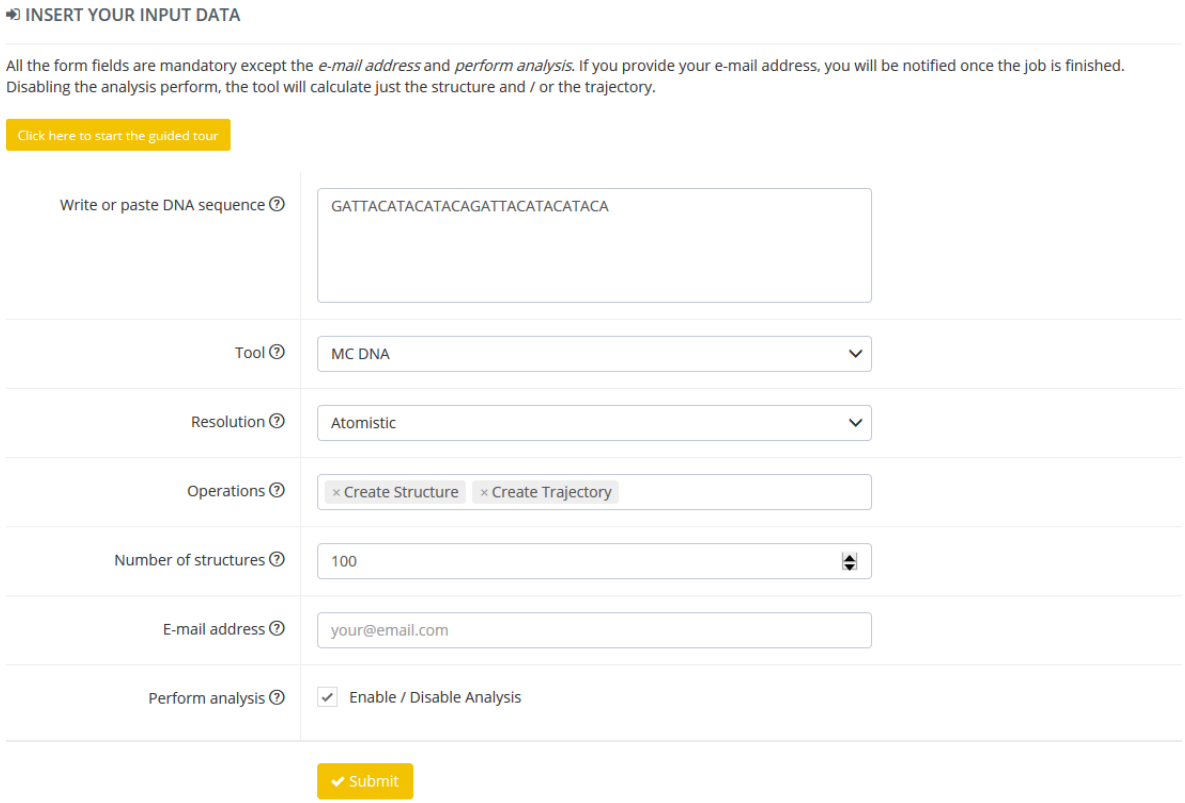
computationally more efficient. The SPC scheme is recursively executed until the position of bp i+2. Bp i+1 is obtained by the deterministic full closure (DFC) procedure (Suppl. Figure S7B, bottom). The position of bp i+1 is obtained on the intersecting circle between two spheres with radius d centered around bp i and i+2. Radius d is the distance between bp i and i+1 before the Monte Carlo move. Using this procedure there exists either no (in case the spheres do not overlap) or an infinite number of (in case of overlap) solutions. In case of overlap, the new position of bp i+1 is chosen by finding the point closest to the previous position of base pair i+1 on the intersecting circle.

SUPPLEMENTARY TABLES

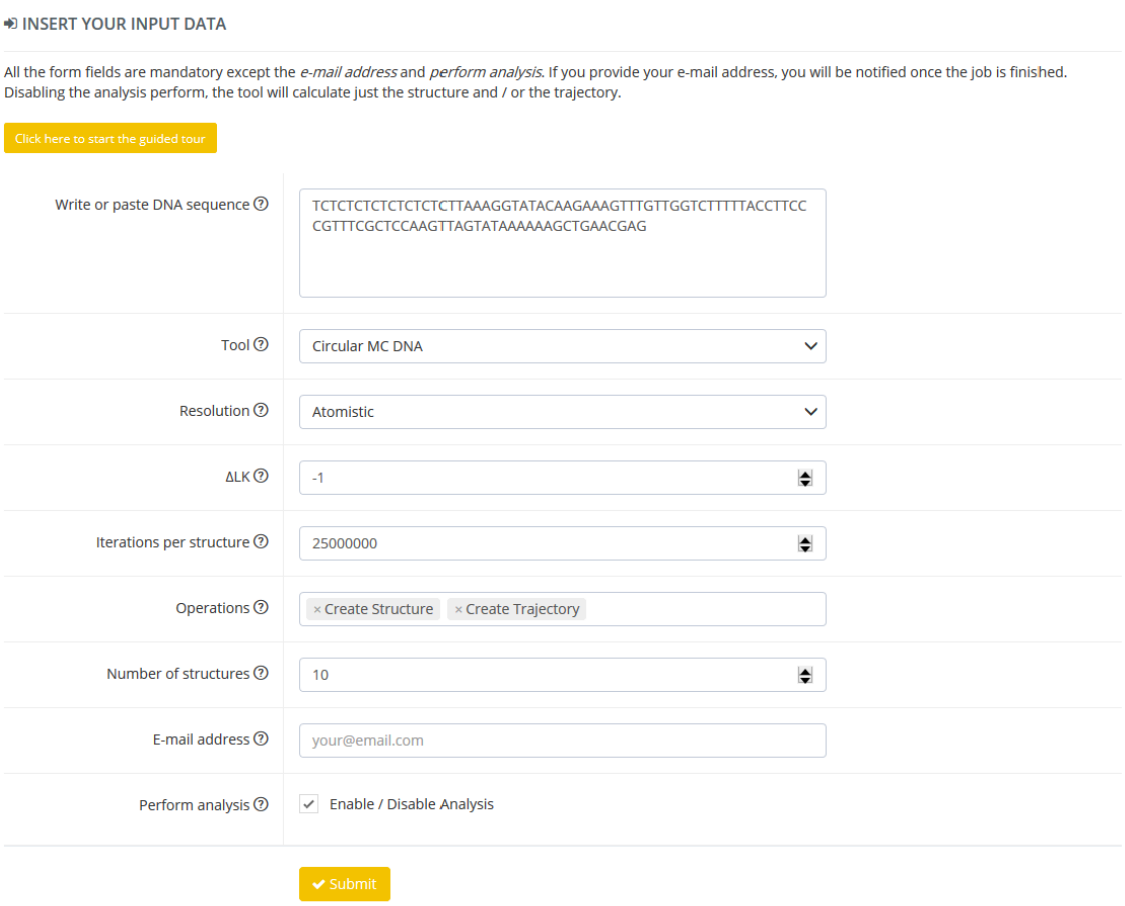
Supplementary Table 1: Set of sequences where the central 14 bp's contain all unique bps in all the different tetramer environments. This set of sequences was determined by the ABC consortium and it is known as the 'miniABC' library of sequences.

Seq. number	Watson strand (5'-3' direction)
1	GCAACGTGCTATGGAAGC
2	GCAATAAGTACCAGGAGC
3	GCAGAAACAGCTCTGCGC
4	GCAGGCGCAAGACTGAGC
5	GCATTGGGGACACTACGC
6	GCGAACTCAAAGGTTGGC
7	GCGACCGAATGTAATTGC
8	GCGGAGGGCCGGGTGGGC
9	GCGTTAGATTAAAATTGC
10	GCTACGCGGATCGAGAGC
11	GCTGATATACGATGCAGC
12	GCTGGCATGAAGCGACGC
13	GCTTGTGACGGCTAGGGC

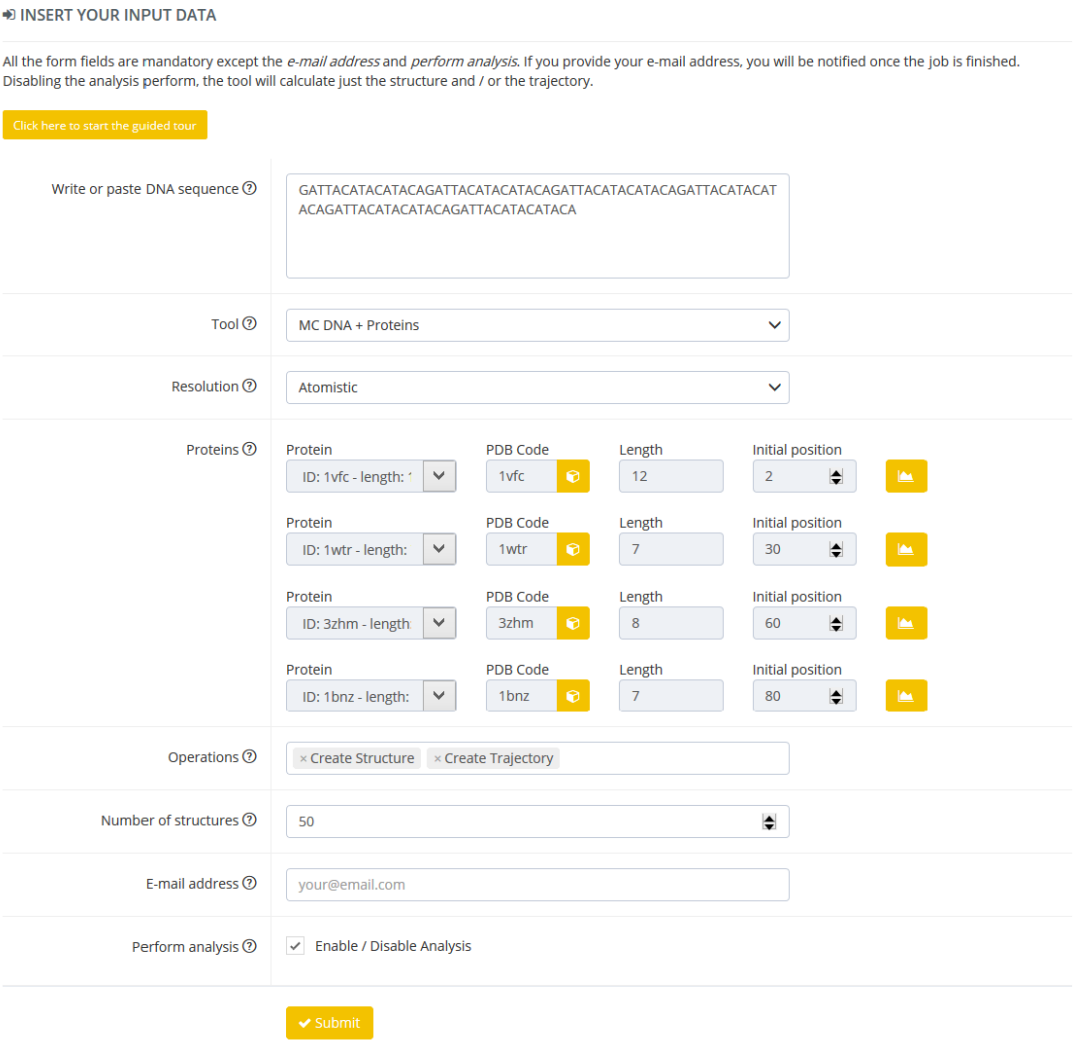
SUPPLEMENTARY FIGURES



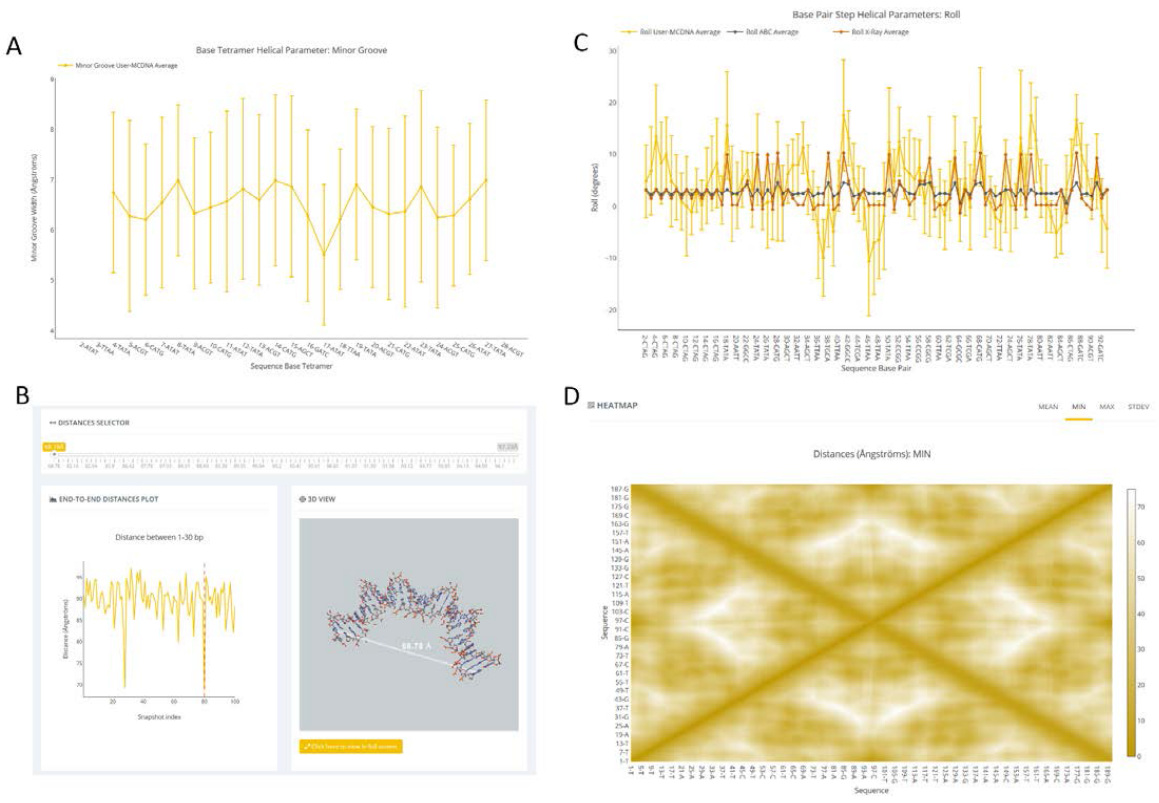
Supplementary Figure S1: In this Sample Input of the tool MC DNA of an example sequence the resolution is chosen to be atomistic and an equilibrated structure as well as 100 snapshots representative of a $1.5x10^6$ trajectory will be simulated.



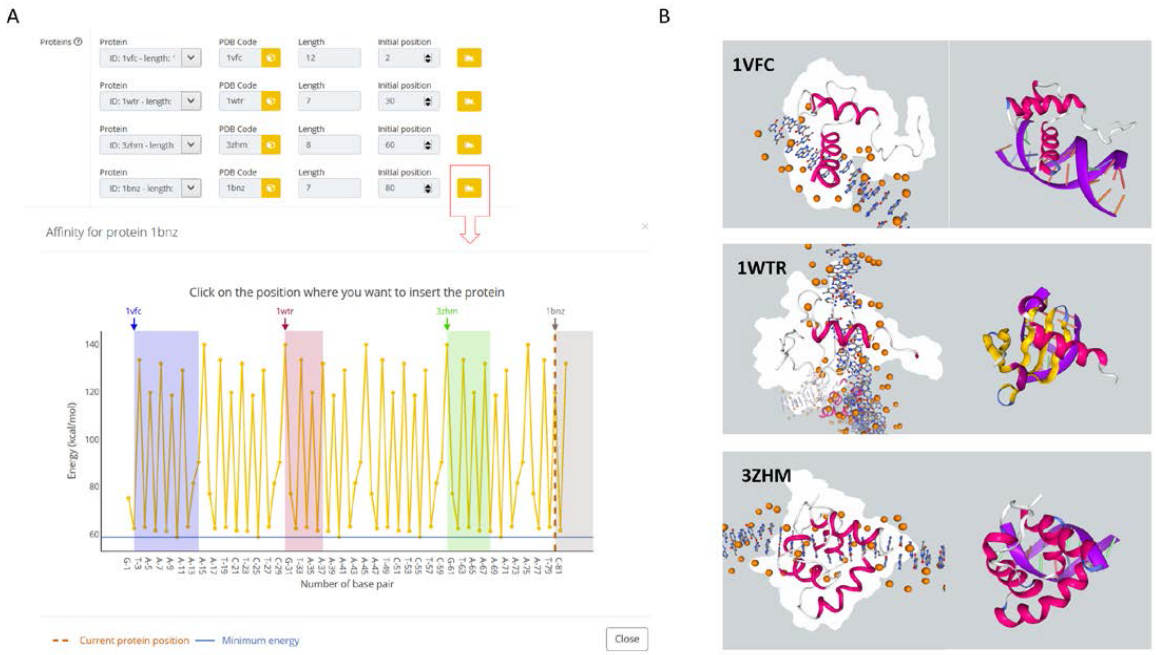
Supplementary Figure S2: In this Sample Input of the tool Circular MC DNA of a 94bp long example sequence the resolution is chosen to be atomistic, the torsional stress expressed in ΔLk is set to -1. With these settings a planar equilibrated circular structure as well as a trajectory of 10 snapshots with $25x10^6$ iterations per snapshot are chosen which enables to probe changes in the circular shape due to the induced torsional stress.



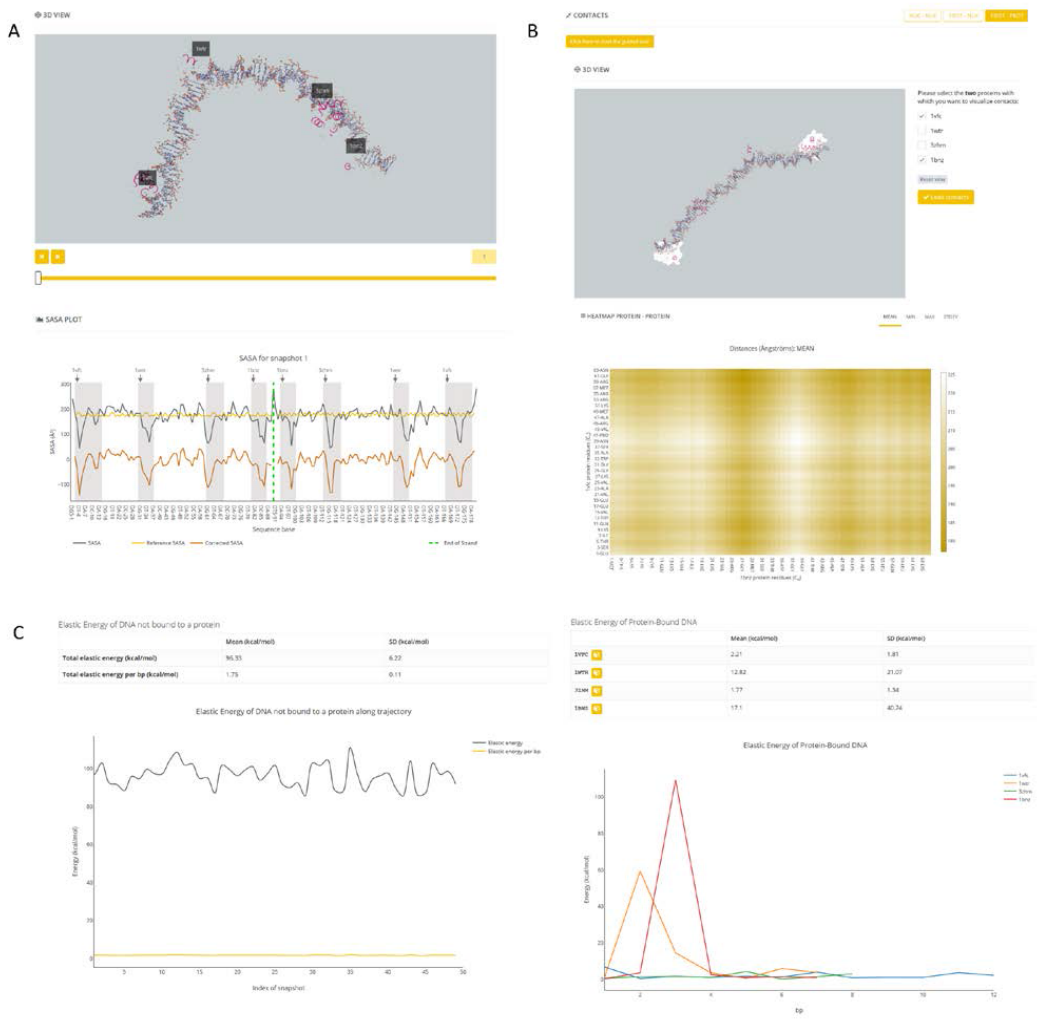
Supplementary Figure S3: In the Sample Input of MC DNA + Proteins the resolution is chosen to be atomistic and four proteins at specified initial positions are chosen for the chromatin fragment. The server will provide the equilibrated structure as well as 50 representative snapshots of the trajectory (around 4 million movements in this case).



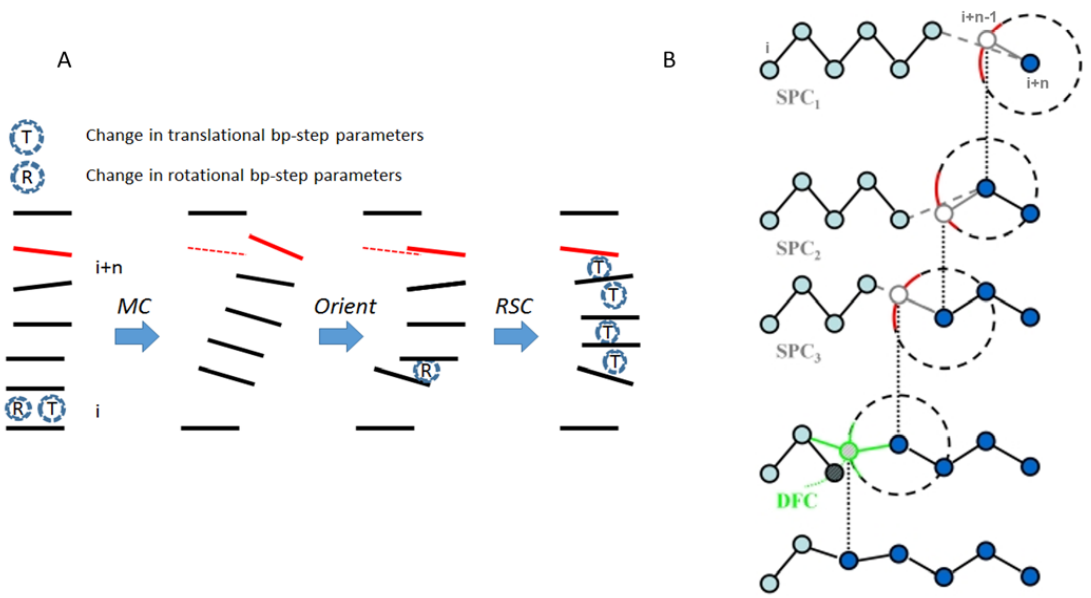
Supplementary Figure S4: Excerpt of Sample Outputs for MC_DNA and Circular MC_DNA. A: Minor groove width of trajectory of atomistic representation of sample input for MC DNA. B: End to End distance selector of trajectory of atomistic representation of sample input for MC DNA. C: Parameters of Roll of the trajectory analysis of atomistic representation of sample input for Circular MC DNA. D: Smallest distances between each base along the trajectory of atomistic representation of sample input for Circular MC_DNA. Due to the axis labels the plot shown here appears to have a resolution lower than one bps, however moving the cursor over the interactive plot in the server shows one base resolution



Supplementary Figure S5: Details on the placement of the proteins along the fiber. A: Fragment of the MC_DNA input form for the Protein-DNA method. A yellow box is attached to every input protein, offering the possibility to launch a protein affinity process to identify the most favorable regions of the sequence to position the protein structure. In order to avoid possible overlaps, the proteins already included in the fiber are highlighted in colored rectangles, taking into account the length of the sequence recognized by the protein. B: Examples of modeled protein-DNA complexes from MC_DNA. Modeled complexes (left), and especially the DNA fragment orientation, can be compared to the original PDB crystal (right).



Supplementary Figure S6: Excerpt of Sample Outputs for MC_DNA + Proteins. A: Solvent accessible surface area (SASA) of DNA along trajectory of atomistic representation of sample input. Top: Selection of snapshot; Bottom: SASA along the bases. B: DNA-mediated protein-protein contact analysis of trajectory of atomistic representation of sample input. Top: Selection of the two proteins to compare; Bottom: Mean distance between each amino acid of the proteins. C: Elastic Energy of DNA. Left: Table of mean and standard deviation of total elastic energy of DNA not bound to a protein along the trajectory, the graph underneath indicates the elastic energy of DNA not bound to a protein per snapshot; Right: Table of mean and standard deviation of elastic energy (per bp) of protein-bound DNA, the graph underneath shows the values of the elastic energy of the protein-bound DNA.



Supplementary Figure S7: Monte Carlo move corrections for the simulations of the tool Circular MC_DNA. A: Scheme of the segmental Monte Carlo move used in 'Circular MC_DNA'. The regular Monte Carlo move ('MC') is followed by a resetting step to keep the orientation of bps i+n until i+2 as before the MC move ('Orient'). In the end, the Recursive Stochastic Closure ('RSC') algorithm is applied to keep the circular structure intact. B: RSC scheme shows first several moves of the Stochastic partial closure (SPC, grey) until in the last step the deterministic full closure (DFC, green) algorithm completely repairs the chain break (adapted from ref 16 in the manuscript).

3. Development of a nucleosome fiber model (Publication 5)

The first compaction level of DNA in eukaryotic cell nuclei is the nucleosome with approximately 147 base pairs (bp) of DNA wrapped around an octamer of histones. The three dimensional arrangement of the nucleosome units is connected via a DNA linker, often referred to as 'beads-on-a-string' fiber, depicting the chromatin secondary structure (14). Advances in experimental techniques and in computational modeling in the last decade revealed that chromatin in-vivo adopts a dynamic and heterogeneous conformation (15) which depends on many different intra-cellular factors. Our goal was to determine, with the help of experimental data, three dimensional chromatin arrangements the way they can possibly occur inside the cell.

In yeast, for example, Micro-C (16) experiments revealed the formation of self-associating domains at the nucleosome level where domain boundaries are enriched in nucleosome depleted regions suggesting that the length of the DNA linker connecting two adjacent nucleosomes plays a decisive role in chromatin compaction (17). MNase-seq experiments can determine the nucleosome positions along the genomic sequence, however as is the case for Micro-C, those experiments are performed with many thousands to millions of cells, and the obtained results show population averages where no information for single cells can be obtained. To overcome this issue, we developed a machine learning method to deconvolute population based MNase-seq data to derive potential nucleosome positions in a single cell that lead to physically realistic chromatin conformations by probing clashes in 3D space via a simple nucleosome fiber model.

To obtain physically realistic conformations we developed a nucleosome fiber model using the bottom-up approach. In this model, linker DNA is modeled by the mesoscopic helical DNA model at base pair resolution as previously described and the nucleosome is assumed to be rigid and fixed in the sequence, with the nucleosomal DNA fixed to the 3D DNA path of the high-resolution X-ray structure (PDB 1KX5) and with a coarse grained description of charges and steric constraints of histone core and DNA to correctly account for electrostatics and excluded volume interactions. The energetic contributions from those long-range interactions (electrostatics and steric repulsion) combined with an elastic energy description of the DNA results in a Hamiltonian which is coupled with Metropolis Monte Carlo sampling algorithm (see Figure 1 in the following publication). There also exists the possibility to convert the coarse grain representation of the

chromatin fiber back to atomistic coordinates of its constituents. As validation, we found that the nucleosome fiber model reproduces well salt-dependent sedimentation coefficients in vitro (18) and shows correct compaction by evaluating the fiber volume of a 100 nucleosome fiber (see Figure 2 in the following publication).

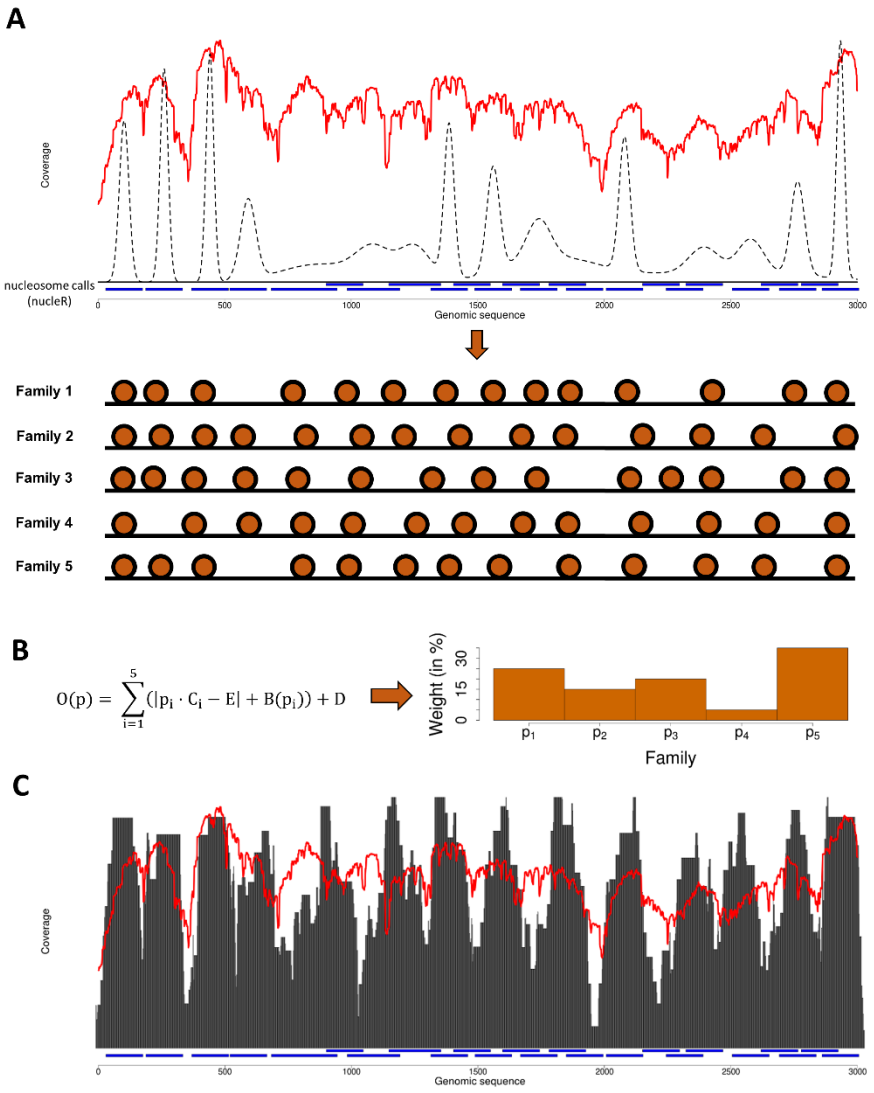


Figure 30. Procedure of deconvolution of the MNase data. A: The experimental coverage (red) of a genomic segment is approximated by a probability distribution (dashed black) based on the nucleosome calls by 'nucleR'. Possible physically realistic nucleosome configurations called families are sampled based on the probability distribution. B: Optimization to adjust the weights of each family. C: The combined artificial MNase signal of the families (black histogram) shows good agreement with the experimental coverage (red).

With this working coarse grain chromatin model, we can probe for physical clashes for the suggested conformations of single cells. To obtain those physically relevant configurations we cluster the MNase-seq data of a genomic segment into a set of families that each contain a number of nucleosomes similar to the sum the scores of the nucleosome calls detected by 'nucleR', a non-parametric method of detecting nucleosome dyads from MNase data. The families combined describe the nucleosome coverage of the whole cell population by optimizing the weights of each possible conformation (see Figure 30). Having incorporated nucleosome positioning restraints from MNase data into the nucleosome fiber model we can make use of 3C-based techniques (19) as a second filter to reach not only realistic nucleosome positions along the genomic sequence, but also realistic three dimensional conformations. Micro-C data (16) provides cell population data of chromatin compaction, with nucleosomal level of resolution. In our case, we used Micro-C data to refine the ensemble of structures emerging from each physically realistic nucleosome position configuration. The refinement procedure consists of smart structure filtering followed by reweighting of the kept configurations to correctly reproduce the Micro-C contact matrix (see Figure 4-5 in the following publication). This refinement procedure was applied to probe the compaction of single genes in yeast cells with and without oxidative stress (see Figure 6 in the following publication) with the result that in general genes under oxidative stress are less compact.

In summary, the pathway of introducing different experimental biases into in-silico base pair resolution nucleosome fiber conformation modeling such as 1D nucleosome position restraints based on MNase data and structure filtering based on Micro-C 2D contact matrices provides a step towards sampling kb-long chromatin fiber conformations possibly present inside the cell nucleus.

Publication:

Jürgen Walther, Pablo D. Dans, Manuel Sarmiento, Rafael Lema, Isabelle Brun-Heath and Modesto Orozco, A method to predict chromatin fiber conformations by deconvolution of nucleosome positioning data and Micro-C, (in preparation).

A METHOD TO PREDICT CHROMATIN FIBER CONFORMATIONS BY DECONVOLUTION OF NUCLEOSOME POSITIONING DATA AND MICRO-C

Jürgen Walther[1], Pablo D. Dans[1], Manuel Sarmiento[1] and Modesto Orozco[1,2]*

We present a method to obtain three dimensional ensembles of nucleosome fiber structures in genomic segment by coupling a physical model of chromatin with MNase-seq and Micro-C data obtained on a large population of cells. Machine learning methods deconvolute the MNase-seq signal of a genomic segment into a minimum number of families – non-overlapping in three dimensional space - that with appropriate weights reconstitute the experimental nucleosome coverage. Simulation of the 3D dynamics of the families using a simple chromatin fiber model yields an ensemble of structures which is then filtered to match the Micro-C contact matrix. The filtered ensemble represents the most probable three dimensional arrangements of the chromatin fiber in a given genomic segment.

---

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology.
[2] Department of Biochemistry and Biomedicine. The University of Barcelona. Barcelona, Spain.

* Correspondence to Prof. Modesto Orozco: modesto.orozco@irbbarcelona.org

INTRODUCTION

The nucleosome comprises the first compaction level of DNA in eukaryotic cell nuclei with approximately 147 base pairs (bp) of DNA wrapped around an octamer of histones. The chromatin secondary structure is then depicted as the three dimensional arrangement of the nucleosome units connected via a DNA linker, often referred to as 'beads-on-a-string' fiber. Several regular topologies of chromatin secondary structure have been proposed to exist based on *in-vitro* data (1), the two most accepted models being: the 'Solenoid' (2) and the 'zigzag arrangement' (3, 4). However, recent evidence has shown that chromatin *in-vivo* adopts more dynamic and heterogeneous conformations than those expected from in vitro biophysical studies (5), and current knowledge suggest the nucleosome string is defined as a quite plastic ensemble with different levels of structural organizations in a fast equilibrium that can be modified based on a variety of inter and intra-cellular factors.

Simple "chain-of-beads" models of chromatin representing the nucleosome fiber as a chain of beads fail when assuming that all the linker segments are equal (which will force regular secondary structures). Thus, recent Micro-C experiments (6) in *Saccharomyces cerevisiae* revealed that nucleosomes form self-associating domains of one to five genes in size (ca. 2-10kb) where boundaries are enriched in nucleosome free regions (NFR), something that will be never possible in a regularly distributed nucleosome fiber. Clearly, the length of the DNA linker connecting two adjacent nucleosomes plays probably a decisive role in chromatin compaction (7) as long NFR are acting as hinges making compaction of the nucleosome fiber possible.

Second generation "chain-of-beads" models implement the linker distributions (8) derived from MNase-seq experiments (9), and can incorporate Micro-C contact maps as 3D constraints that the folded fiber should fulfill. However, these models, which are a clear step forward in the field, suffer from the "ensemble" nature of the experimental data. That is, both MNase-seq and Micro-C are obtained from the analysis of millions of cells (to reduce noise), and the signal detected might not

correspond to a real chromatin fiber, as single cell experiments (10, 11) suggest cellular variability is huge. This is clearly shown in Micro-C maps, where incompatible contacts are found, and in MNase-seq experiments where protected reads are annotated to two "overlapping" nucleosomes.

We present here a method which employs machine learning procedures to deconvolute MNase-seq data of a genomic segment into a minimal set of families that combined describe well the experimental nucleosome coverage. Nucleosome positions are sampled from a probability distribution consisting of a set of Gaussians with their centers and standard deviations derived by the nucleosome calls of 'nucleR' (12), a non-parametric method of detecting nucleosome dyads from MNase-seq data. A short simulation using a simple nucleosome fiber model untangles possible physical clashes in Cartesian space and subsequent clustering of the physically realistic nucleosome conformations guarantees a minimal set of relevant configurations. An optimization procedure assigns a weight to each conformation to reconstitute the experimental coverage as accurate as possible. Simulation of this set of configurations using a physical model of chromatin accounting for sequence variability in the DNA allows to sample flexibility of long segments of chromatin. We develop strategies to bias such ensembles using Micro-C data, which is again considered as an "ensemble property" of a set of cells. We explore the behavior of the model in a study where the chromatin structure of yeast is investigated under oxidative stress (OS) and compared to the untreated cells (WT). We find that nucleosomes are more fuzzy in OS than in WT and as a consequence local chromatin compaction at gene level under OS is generally lower compared to the control no matter of the transcriptional state of the gene (upregulated, no change, downregulated, -1 nucleosome missing in OS) suggesting that oxidative stress influences the whole nucleus in an equal manner.

THE ALGORITHM

Chromatin fiber model. A simple nucleosome fiber model was developed to probe dynamics properties of chromatin at the kb-scale. The DNA is treated as a

freely moving unrestrained polymeric entity and is described by the recently developed extended nearest neighbor helical coarse grain DNA model (13). In short, an elastic potential in the inter base pair parameter space (shift, slide, rise, tilt, roll twist) describes the dynamics of a base pair step (bps). The flexibility parameters are derived from atomistic molecular dynamics (MD) simulations and depend on the tetranucleotide context of the central bps. For an accurate parametrization of the model the inter base pair parameter distributions for the central bps of each tetranucleotide are clustered into different helical states (see (13) for more details). The elastic energy of the DNA fiber is then calculated as the sum of the individual contributions of each bps

$$E^{DNA}(X) = -k_B T \sum_{j=1}^{N} \ln \sum_{i=1}^{n} e^{-\frac{1}{k_B T}\left(\frac{1}{2} K_{ij} \Delta X_{ij}^{\ 2} + E_{ij}\right)} \qquad (1)$$

where k_B is the Boltzmann constant, T is the temperature, N is the number of bps, n is the number of states in which the distribution of inter base pair parameters of a given bps (in its sequence environment) can be decomposed (see below), K is the stiffness matrix associated to the state i in step j; ΔX is the deformation vector (with equilibrium values dependent on step j and state i) and E_{ij} is the relative energy of state i at bps j (shifting values between states). Note that due to sequence end effects single state dimer stiffness parameters are used for the first and last bps.

The inter base pair coordinates can be used to derive Cartesian representations of the DNA. For a given set of inter base pair coordinates the positions of the phosphates are derived from the helical axis (see (13) for more details). Atomistic coordinates of the nucleobases can be obtained using the SCHNArP algorithm (14), however to increase computational efficiency in the Cartesian reconstruction process we restrict the algorithm only to reconstruct the center of each base pair.

When DNA is bound to histones we keep the DNA's geometry restrained to that in the experimentally resolved X-ray structure (PDB:1KX5). Nucleosome cores are approximated as spherical particles with the center being the center-of-mass of the

complex. The particles are anchored to the DNA bound to it and relative positions between bound DNA and protein remain unchanged.

To account for interactions between the constituents of the chromatin fiber excluded volume and electrostatic potentials are included in the model. The electrostatic contribution is calculated using a Debye-Hückel potential

$$E^{DH} = \sum_{ij} \frac{q_i q_j}{4\pi\varepsilon\varepsilon_0 r_{ij}} e^{-\kappa r_{ij}} \quad (2)$$

where summation goes over all charged constituents (all charged protein complexes, phosphates of DNA), ε_0 is the dielectric constant in vacuum, ε is the dielectric constant in the medium (set to 80), r_{ij} is the distance between charged constituent i and j and κ the Debye length due to ionic screening in solution. κ is adjusted to resemble the monovalent ion concentration inside the cell nucleus.
To prevent the fiber from physical overlap a Lenard-Jones potential is applied to all fiber constituents (DNA and proteins). It is calculated as

$$E^{LJ} = \sum_{i,j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \quad (3)$$

where ε_{ij} and σ_{ij} are strength and equilibrium distance parameters dependent on the fiber constituents i and j with distance r_{ij}. Values for parameters of equ. (2) and (3) are given in Supplementary Table 1.
The total energy of the chromatin fiber system is then calculated as

$$E^{tot} = E^{DNA} + E^{DH} + E^{LJ} \quad (4)$$

5

For example, if a DNA fiber is coated with only one type of protein complexes such as nucleosomes, the input of the model is the DNA sequence and the position of the nucleosomes along the genomic coordinates.

Simulation of the chromatin fiber model. The ensemble generation of the chromatin fiber model for given nucleosome positions and DNA sequence is done via a Metropolis Monte Carlo algorithm with pivot moves in the helical space of the unbound DNA. After each helical move the positions of the particles were reconstituted by an efficient coordinate transformation algorithm and the total energy of the system (equ. (4)) is evaluated based on the Metropolis criterion Configurations are drawn after a certain number of Monte Carlo moves to guarantee uncorrelated samples (for a more detailed description see Supplementary Information and Figure 1). For efficient sampling of the conformational space of the chromatin fiber, a short test run is performed and 1000 configurations are drawn equidistantly along the test run (~1M Monte Carlo steps) as new starting configurations for 10 independent simulations to obtain in total 10,000 conformations (~10M Monte Carlo steps). The ensemble is transferred to GROMACS (15) trajectory format to be able to use the GROMACS built-in suite of analysis tools.

Determining nucleosome positions from MNase-seq signal. To obtain nucleosome positions from the reads of MNase-seq experiment we used the freely available in-house package 'nucleR' (12). This algorithm employs a non-parametric method of detecting all nucleosome dyads and scoring of the calls. First, a Fourier analysis is applied to the raw coverage from NGS paired-end reads which results in a smoother signal and it cleans the distortions in the coverage peaks. Noise is removed from the coverage profile using FFT (16) resulting in a filtered trimmed coverage profile that is subject to nucleosome dyad detection by a simple local maxima search largely facilitated by the clarity of the filtered profile. The score of a nucleosome call (from 0 to 1) is determined by the height and width of the peak, giving a high score to well positioned nucleosomes with large and sharp peaks and a low score to fuzzy nucleosomes with small and wide peaks. The total number of

nucleosomes in a genomic segment can then be defined as the sum of the scores for each nucleosome detected within that segment

$$N = \sum_{i=1}^{L} s_i \qquad (5)$$

Where L is the total number of nucleosome calls detected in the considered genomic segment and s_i the score of nucleosome call i.

Sampling of physical relevant structures from determined nucleosome positions. The nucleosome coverage is approximated as a sum of normal distributions with the center, height and width of each Gaussian derived from nucleosome calls parameters obtained by nucleR. The total approximated coverage C of a genomic segment is then:

$$C(x) = \sum_{i=1}^{L} \frac{h_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \qquad (6)$$

With

$$\sigma_i = \frac{w_i}{10 * h_i^2} \qquad (7)$$

Where x is the genomic coordinate of the segment, the summation goes over all nucleosome calls of the genomic segment (total number of L nucleosome calls), μ_i the center of the Gaussian corresponding to the nucleosome dyad of the nucleosome call i, σ_i the standard deviation of the Gaussian and w_i and h_i the width and height of the nucleosome call i.

We then draw nucleosome configurations from C(x) which do not overlap in one dimensional sequence space. Following parsimony principle and available single cell data (10) we draw nucleosome configurations with a certain nucleosome occupancy defined by an interval around the expected number of nucleosomes N in the genomic segment: $N - m \leq n \leq N + m$, where n is the number of nucleosomes drawn and m is the margin (m = 2 in our study). For each of the different

nucleosome occupancies n, the drawn configuration is subject to a short simulation with the simple chromatin fiber model (straight linker DNA according to Arnott's B-DNA values is assumed as starting configuration of the model) to untangle possible physical clashes. Configurations without clashes are detected during the short simulation are kept for further processing until a total of M configurations are sampled. A Gaussian mixture clustering algorithm is applied to achieve a minimum number of conformations which together are able to describe the M configurations.

Reproducing experimental nucleosome occupancy data by reweighting realistic nucleosome conformations. The representatives of each cluster for all allowed nucleosome occupancy numbers n ($N - m \leq n \leq N + m$) are grouped together and followed by a reweighting procedure to reproduce the experimental coverage as accurate as possible. The objective function subject to minimization is

$$O(p) = \sum_{n=N-m}^{N+m} \sum_{k=1}^{P_n} [A(p_{kn}) + B(p_{kn})] + D \tag{8}$$

With p being a vector of all weights of each conformation k with nucleosome number n. The terms in the objective function are calculated as follows:

$$A(p_{kn}) = \left(\sum_{x=n1}^{n2} |p_{kn} C(x) - E(x)| \right) \tag{9}$$

$$B(p_{kn}) = |n - N|^2 \cdot p_{kn} \tag{10}$$

$$D = \left| \sum_{n=N-m}^{N+m} \sum_{k=1}^{P_n} n \cdot p_{kn} - N \right| \tag{11}$$

8

where $A(p_{kn})$ corresponds to the difference in coverage of the experimental coverage E and the reconstructed coverage C weighted by p_{kn}, n1 and n2 being the genomic coordinates of the start and end of the considered genomic segment. $B(p_{kn})$ penalizes nucleosome configurations with occupancy n which differ from the experimentally derived nucleosome occupancy N by a squared penalty while D assures that the nucleosome occupancy of the reweighted conformations stays close to N. The summation over k goes over all cluster representatives P_n for each allowed nucleosome occupancy n and the summation over n includes all allowed nucleosome occupancy value.

The optimization function is subject to a gradient descent optimization with carefully adjusted starting and boundary conditions. Conformations which contribute less than 1% to the experimental signal are neglected. The remaining nucleosome conformations (total number Q) are then referred to as 'families'.

Physical properties of a genomic segment. Each family is subject to a Monte Carlo simulation with the nucleosome fiber model (as described above) to capture ensemble properties. Physical properties of the nucleosome fiber such as radius of gyration, packing ratio and distance matrix can be calculated from the ensemble taking into account the weights of each configuration. A physical property L of a nucleosome fiber of a genomic segment is then:

$$L = \sum_{q=1}^{Q} p_q \cdot L_q \qquad (12)$$

where L_q is the dynamic property of family q weighted according to the contribution p_q of the nucleosome conformation to the experimental nucleosome positioning signal with Q being the number of families.

Structure filtering with Micro-C data. In case that additionally to MNase-seq data Micro-C data exists as is the case in yeast, the ensemble of structures of a family is filtered to match the Micro-C contact matrix of the genomic segment binned to the nucleosome positions of the family. For this purpose, a contact between two nucleosomes is annotated if the distance between the nucleosome

centers is less than 15 nm. In a first filtering process, all structures of the ensemble which experience at least one long-range contact are kept (a long-range contact is defined as a contact of nucleosome i with nucleosome i+0.8n where n is the total number of nucleosomes in the family). In a second filtering process, the complete ensemble of structures is clustered by the K-means method along the one dimensional representation of the contact matrix to obtain two equal-sized populations of structures, one involving structures with long-range contacts and a second one containing representatives of the whole ensemble. In a last refinement procedure the experimental contact matrix is intended to reproduce as accurate as possible by assigning a weight to each structure of the reduced ensemble. Self-interaction and contacts with the neighboring nucleosomes are neglected in the fitting procedure as neighboring contacts can be contaminated by di-nucleosome fragments which experienced insufficient MNase digestion. An objective function subject to minimization is defined to obtain appropriate weights

$$O(s) = \sum_{ij} \left| M_{ij} - E_{ij} \right| / A_{ij} \tag{13}$$

with

$$M = \sum_{k=1}^{N_q} s_k c_k \tag{14}$$

where M is the contact matrix derived from the filtered structures where c_k is the contact matrix of snapshot k with weight s_k, N_q is the ensemble size of family q, A is a matrix displaying average contact counts of a selected part of the genome (in the analysis of single genes A represents the average contact counts of all genes in the genome) and the sum over i and j extends over the upper triangle of the contact matrix without the diagonal and the +1 entries (see above). The weights s of the filtered structures are optimized by a steepest gradient descent algorithm.

To calculate physical properties such as radius of gyration or accessible surface area of the 3D conformation of a genomic segment the contribution of each

snapshot to the ensemble contact matrix is taken into account by extending formula (12) to:

$$L = \sum_{q=1}^{Q} p_q \sum_{k=1}^{N_q} s_k \cdot L_{kq} \qquad (15)$$

where L_{kq} is the physical property of snapshot k of family q weighted according to the contributions s_k and p_q to the contact matrix and experimental nucleosome positioning signal respectively with Q being the number of families.

Physical properties of chromatin fiber structures. With the above described method physical properties of nucleosome fibers can be calculated accounting for experimental nucleosome positioning (see equ. (12)) and contact matrices (see equ. (15)). In this work, we following parameters to probe chromatin compaction: The sedimentation coefficient S of a nucleosome chain is calculated as in (17)

$$S = S_1 \, (1 + \frac{2R}{N} \sum_{i}^{N} \sum_{j>i}^{N} \frac{1}{R_{ij}}) \qquad (16)$$

where S_1 is sedimentation coefficient of a mononucleosome taken as equal to 11.1 Svedberg, R the effective radius of the nucleosome (5.5 nm), N the number of nucleosomes in the fiber and R_{ij} the distance between the geometric center of two nucleosomes. Another parameter to probe the occupied space of a nucleosome fiber is the radius of gyration which is calculated according to

$$\mathrm{R_g} = \frac{1}{\mathrm{N}} \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathbf{r_i} - \mathbf{r_m})^2} \qquad (17)$$

where N is the number of nucleosomes, $\mathbf{r_i}$ the position of the geometric center of nucleosome i and $\mathbf{r_m}$ the center-of-mass of all nucleosomes. The solvent accessible surface area and the total volume of the fiber are calculated with the GROMACS suite of tools using a spherical probe of radius 3 nm, a size in the range of a transcription factor and 5.5/1.0 nm as excluded volume radii for DNA/nucleosome core.

RESULTS AND DISCUSSION

The nucleosome fiber model reproduces chromatin compaction. The three dimensional arrangement of the nucleosome fiber ensembles is a crucial indicator for the performance of the model. It is difficult, though, to obtain high quality in vivo data at bp resolution even though experimental resolution steadily increases (18). In some *in vitro* experiments chromatin fibers are artificially built to guarantee a robust comparison of the interplay of nucleosome positions and environmental changes. We compared our simulation results with sedimentation coefficient experiments at different salt concentrations of a fiber of 12 nucleosomes, each nucleosome separated by 62 bp from its neighbor (see Figure 2). The average ensemble properties of the simulation reveal a good agreement with the salt dependent sedimentation coefficient values bearing in mind that the only parameter subject to change resembling different salt concentration in the nucleosome fiber model is the Debye length for the electrostatic screening. We also compare the average volume experienced by a 100 nucleosome fiber with regularly spaced nucleosomes and a NRL of 162bp with experimental data. The average volume of our model deviates by less than 5% from the experimental data which shows that the nucleosome fiber model performs equally well for short and long fiber configurations suggesting its suitability to sample more in-vivo like chromatin conformations with non-regular nucleosome positioning.

Nucleosome positioning deconvolution recovers experimental data. As nucleosome positioning data arises from cell population data a single nucleosome position configuration is unable to capture the dynamics in nucleosome positioning of the whole population. A more biologically reasonable approach would be to represent the population signal by a minimum set of physically relevant nucleosome conformation. For this reason, we developed an algorithm based on machine learning approaches to represent the total experimental nucleosome coverage of a genomic segment by a weighted set of physically realistic families of different nucleosome positioning. First, the experimental coverage of a genomic segment is approximated by a probability distribution

which is based on the nucleosome calls detected by 'nucleR' (as an example we chose the PAM-1 gene in yeast with a total length from TSS to TTS of ~3 kb; see Figure 3A). Possible physically realistic nucleosome configurations (usually 1,000) are sampled based on the probability distribution. The number of nucleosomes in each conformation is kept close to the average nucleosome number in this genomic segment to satisfy principles of parsimony. A Gaussian mixture clustering algorithm on the 1D nucleosome positions reduces the sampled configurations into a set of families (see Figure 3A). The weights of the representatives of each family are carefully optimized by balancing the penalty of configurations with nucleosome occupancy differing from the experimentally derived nucleosome occupancy N while assuring that the nucleosome occupancy of the reweighted conformations stays close to N (see Figure 3B). The reconstructed coverage shows good agreement with the experimental coverage (see Figure 3C and Supplementary Figure S1 for more examples) suggesting that a small set of physically possible nucleosome fiber conformations where a single conformation could potentially represent a configuration in an individual cell are sufficient to describe population data.

Refinement by Micro-C delivers realistic 3D conformations. We used Micro-C data to refine the ensemble of structures sampled via the nucleosome fiber model (10,000 configurations for each family are sampled). For each family we binned the Micro-C contact matrix to the corresponding nucleosome positioning profile and we transformed each configuration of the nucleosome fiber model into a contact matrix containing 1 in case of a contact and 0 otherwise. A two-step filtering process ensures that only a relevant set of the sampled configurations are considered. In the first step, all structures which experience long range contacts are kept for further analysis. In the second step the complete ensemble of structures is clustered to obtain two equal-sized populations of structures, one involving structures with long-range contacts and a second one containing representatives of the whole ensemble (see Supplementary Figure S2). In a last refinement procedure, the experimental contact matrix is intended to reproduce as accurate as possible by assigning a weight to each structure of the reduced

ensemble (see Figure 4 and Supplementary Figure S3). Self-interaction and contacts with the neighboring nucleosomes are neglected in the contact matrices since neighboring contacts can be contaminated by di-nucleosome fragments which experienced insufficient MNase digestion. Formula (15) can then be used to obtain physical properties of the chromatin configuration of a genomic segment incorporating both weights arising from the deconvolution of the nucleosome positioning signal and structure refinement using Micro-C (see Figure 5).

Yeast chromatin experiencing oxidative stress is less compact. Yeast constitutes a suitable organism to probe local chromatin compaction by incorporating experimental restraints since because of its size MNase-seq and Micro-C experiments yield a good coverage along the whole genome. We analyzed yeast undergoing oxidative stress (OS) versus untreated cells. In general nucleosomes appear more fuzzy in OS and lead to a lower nucleosome occupancy according to formula (5) (Figure 6A). Furthermore, to detect difference in local chromatin compaction related to gene expression we analyzed the top five genes upregulated and downregulated in OS and a set of five key genes which do not change expression. Local compaction properties such as radius of gyration showed that incorporation of Micro-C contact matrices as a restraint in the model leads to more compact ensemble properties. This suggests that even though the nucleosome fiber model was parametrized according to in vitro data, the cellular environment drives the fiber probably to a more compact conformation (Figure 6A). We explored apart from the radius of gyration other physical properties such as volume and transcription factor accessible surface area to investigate in more detail the different compaction states (Figure 6B and 6C). Genes show higher volume and more accessible surface in OS in absolute (Figure 6B) values. To compare the compaction of different genes we calculated the volume or surface area per nucleosome (Figure 6C). We find no clear trend of any gene type being responsible for the compaction, instead a general trend can be deduced suggesting less compaction of yeast under oxidative stress.

CONCLUSIONS

We present a method to obtain physically realistic nucleosome fiber conformations with sequential nucleosome positions derived from in-vivo data where deconvoluted states can serve as possible single cell nucleosome arrangements. We extended this procedure to match the three dimensional conformation of the simulated fiber ensembles with Micro-C data. We find out with this method that the local compaction of yeast is lower in cells undergoing oxidative stress than in untreated cells, a general phenomenon independent of the transcription state of the genes. In summary, this method provides a large step towards sampling biologically and physically realistic chromatin fiber conformations possibly present in single cells.

ACKNOWLEDGEMENTS

We thank ... J.W. is a La Caixa PhD fellow (UB and IRB Barcelona, Spain). P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher. M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avançats) researcher.

AUTHOR CONTRIBUTIONS

J.W designed the nucleosome fiber model, developed and coded the method and performed all the analysis with support of P.D.D. and M.S.. M.S. developed the nucleosome deconvolution method. J.W., P.D.D., and M.O. discussed the analysis and wrote the manuscript with contributions from all the co-authors. M.O. directed the work.

REFERENCES

1. Tremethick,D.J. (2007) Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. *Cell*, **128**, 651–654.
https://doi.org/10.1016/j.cell.2007.02.008
http://www.ncbi.nlm.nih.gov/pubmed/17320503

2. Robinson,P.J.J., Fairall,L., Huynh,V.A.T. and Rhodes,D. (2006) EM measurements define the dimensions of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci.*, **103**, 6506–6511.
https://doi.org/10.1073/pnas.0601212103
http://www.ncbi.nlm.nih.gov/pubmed/16617109

3. Schalch,T., Duda,S., Sargent,D.F. and Richmond,T.J. (2005) X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, **436**, 138–141.
https://doi.org/10.1038/nature03686
http://www.ncbi.nlm.nih.gov/pubmed/16001076

4. Song,F., Chen,P., Sun,D., Wang,M., Dong,L., Liang,D., Xu,R.-M., Zhu,P. and Li,G. (2014) Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units. *Science (80-.).*, **344**, 376–380.
https://doi.org/10.1126/science.1251413
http://www.ncbi.nlm.nih.gov/pubmed/24763583

5. Grigoryev,S.A., Arya,G., Correll,S., Woodcock,C.L. and Schlick,T. (2009) Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci.*, **106**, 13317–13322.
https://doi.org/10.1073/pnas.0903280106

6. Hsieh,T.-H.S., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, **162**, 108–119.
https://doi.org/10.1016/j.cell.2015.05.048
http://www.ncbi.nlm.nih.gov/pubmed/26119342

7. Wiese,O., Marenduzzo,D. and Brackley,C.A. (2018) Nucleosome positions alone determine micro-domains in yeast chromosomes. *bioRxiv*, 10.1101/456202.
https://doi.org/10.1101/456202

8. Bascom,G.D., Myers,C.G. and Schlick,T. (2019) Mesoscale modeling reveals

formation of an epigenetically driven HOXC gene hub. *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 4955–4962.
https://doi.org/10.1073/pnas.1816424116
http://www.ncbi.nlm.nih.gov/pubmed/30718394

9. Teng,Y., Yu,S. and Waters,R. (2001) The mapping of nucleosomes and regulatory protein binding sites at the Saccharomyces cerevisiae MFA2 gene: a high resolution approach. *Nucleic Acids Res.*, **29**, 64e – 64.
https://doi.org/10.1093/nar/29.13.e64

10. Lai,B., Gao,W., Cui,K., Xie,W., Tang,Q., Jin,W., Hu,G., Ni,B. and Zhao,K. (2018) Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature*, **562**, 281–285.
https://doi.org/10.1038/s41586-018-0567-3
http://www.ncbi.nlm.nih.gov/pubmed/30258225

11. Stevens,T.J., Lando,D., Basu,S., Atkinson,L.P., Cao,Y., Lee,S.F., Leeb,M., Wohlfahrt,K.J., Boucher,W., O'Shaughnessy-Kirwan,A., *et al.* (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, **544**, 59–64.
https://doi.org/10.1038/nature21429
http://www.ncbi.nlm.nih.gov/pubmed/28289288

12. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.
https://doi.org/10.1093/bioinformatics/btr345
http://www.ncbi.nlm.nih.gov/pubmed/21653521

13. Walther,J., Dans,P.D., Balaceanu,A., Hospital,A., Bayarri,G. and Orozco,M. (2019) A MULTI-MODAL COARSE-GRAIN MODEL OF DNA FLEXIBILITY MAPPABLE TO THE ATOMISTIC LEVEL. *Submitted*.

14. Lu,X.-J., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP) 1 1Edited by K. Nagai. *J. Mol. Biol.*, **273**, 681–691.
https://doi.org/10.1006/jmbi.1997.1345
http://www.ncbi.nlm.nih.gov/pubmed/9356256

15. Abraham,M.J., Murtola,T., Schulz,R., Páll,S., Smith,J.C., Hess,B. and Lindahl,E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1–2**, 19–25.
https://doi.org/10.1016/J.SOFTX.2015.06.001

16. Smith,S.W. and W.,S. (1997) The scientist and engineer's guide to digital signal processing California Technical Pub.

17. Arya,G. and Schlick,T. (2009) A tale of tails: how histone tails mediate chromatin compaction in different salt and linker histone environments. *J. Phys. Chem. A*, **113**, 4045–59.
https://doi.org/10.1021/jp810375d
http://www.ncbi.nlm.nih.gov/pubmed/19298048

18. Ricci,M.A., Manzo,C., García-Parajo,M.F., Lakadamyali,M. and Cosma,M.P. (2015) Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*, **160**, 1145–58.
https://doi.org/10.1016/j.cell.2015.01.054
http://www.ncbi.nlm.nih.gov/pubmed/25768910

FIGURES

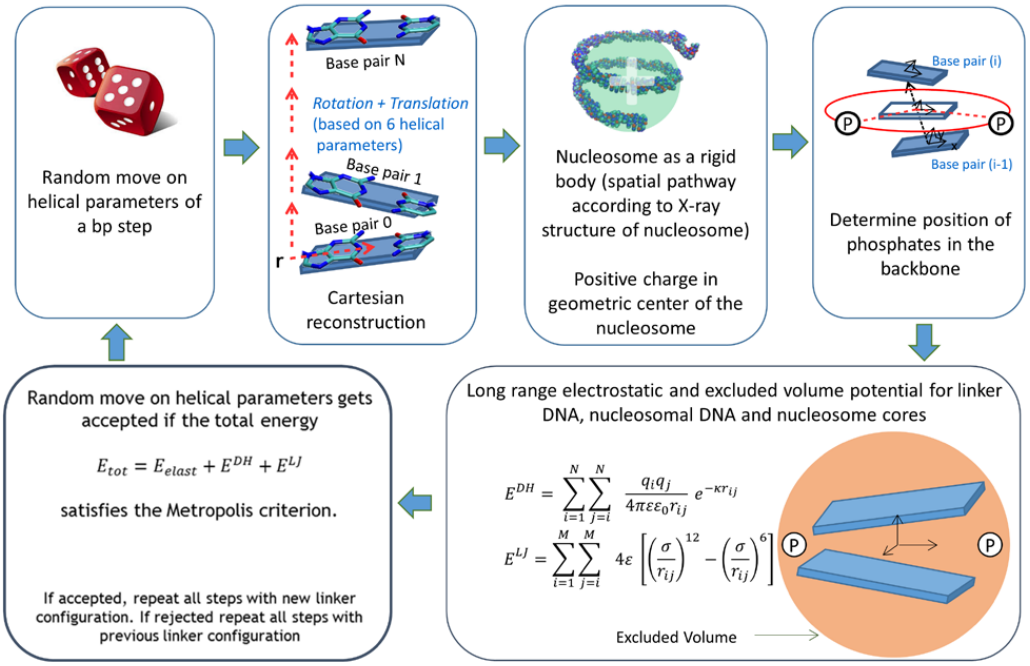


Figure 1. Workflow of the nucleosome fiber simulation.

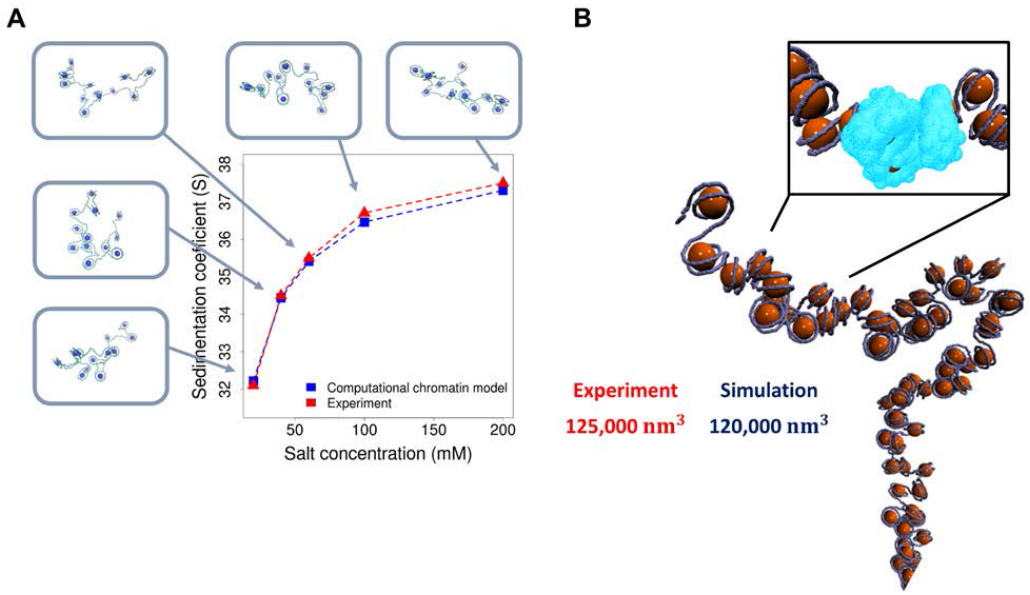


Figure 2. Comparison of compaction of nucleosome fiber model and experiment. A: Salt-dependent sedimentation coefficient experiments of a regular fiber of 12 nucleosomes with 62 bp of linker DNA. The insets show representative snapshots of the nucleosome fiber ensemble modeled at the specific salt concentration. B: Volume estimation of a regular fiber of 100 nucleosomes with 15 bp of linker DNA.

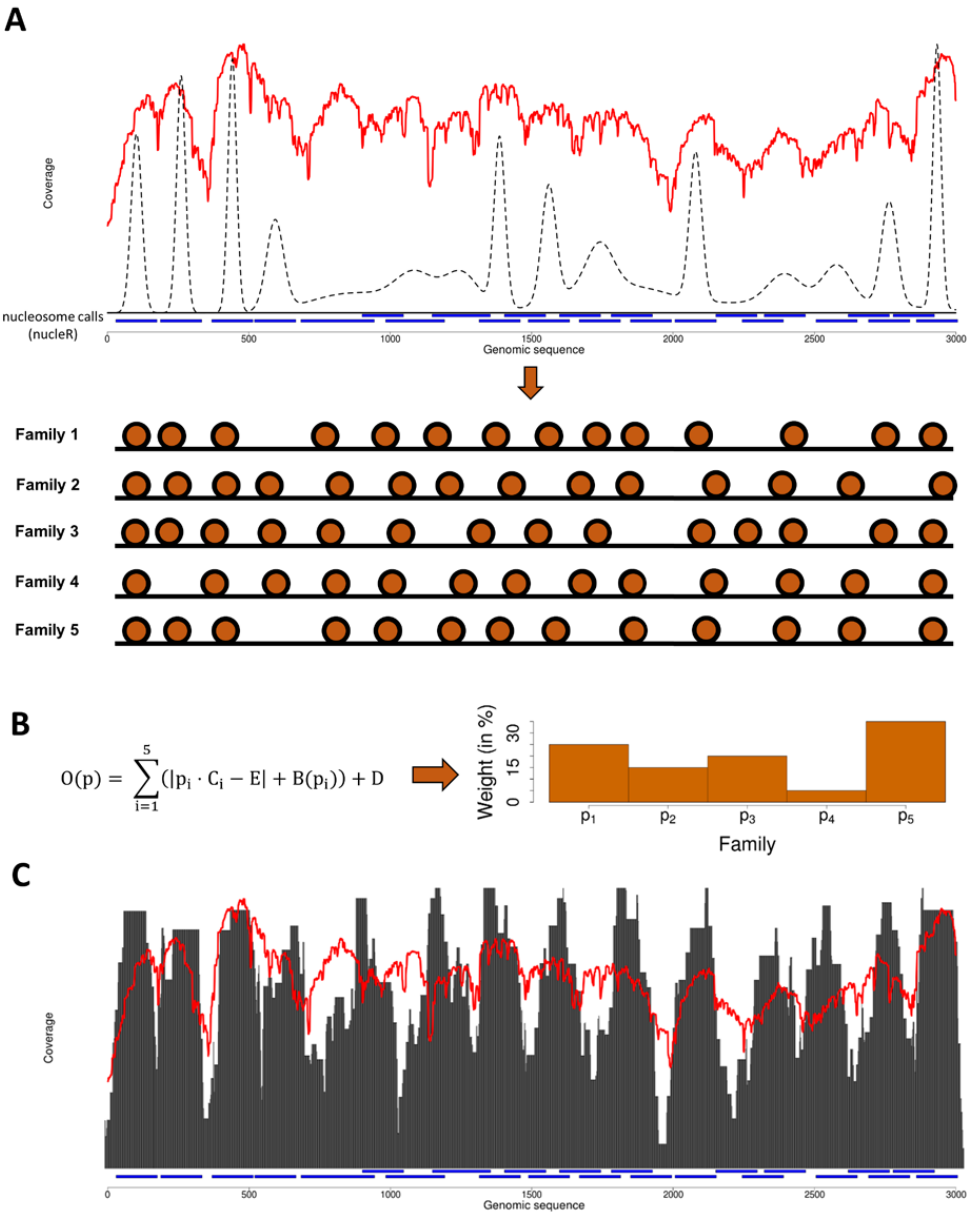

A
Coverage
nucleosome calls
(nucleR)
0
500
1000
1500
2000
2500
3000
Genomic sequence
Family 1
Family 2
Family 3
Family 4
Family 5
B
$O(p) = \sum_{i=1}^{5} (|p_i \cdot C_i - E| + B(p_i)) + D$
Weight (in %)
0
15
30
p_1
p_2
p_3
p_4
p_5
Family
C
Coverage
0
500
1000
1500
2000
2500
3000
Genomic sequence

Figure 3. Procedure of deconvolution of the MNase data. A: The experimental coverage (red) of a genomic segment is approximated by a probability distribution (dashed black) based on the nucleosome calls by 'nucleR'. Possible physically realistic nucleosome configurations called families are sampled based on the probability distribution. B: Optimization to adjust the weights of each family. C: The combined artificial MNase signal of the families (black histogram) shows good agreement with the experimental coverage (red).

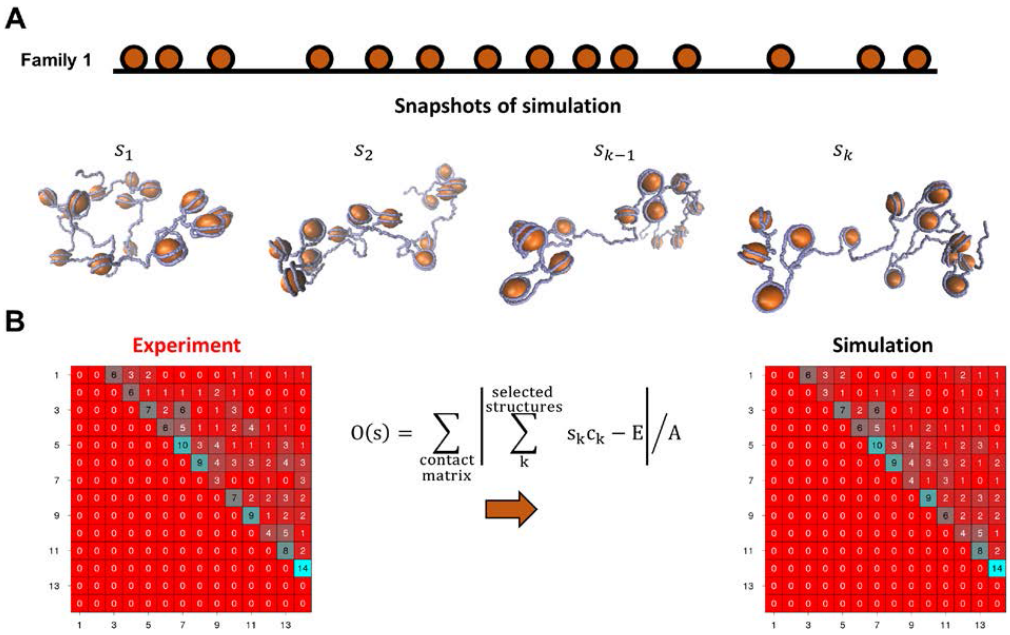


Figure 4. Reproducing experimental contact maps from Micro-C. A: Using the nucleosome positions of a given family a subset of the generated ensemble of structures (for more information see 'Materials and Methods') is selected for fitting. Representative structures of the ensemble subset are shown. B: An optimization procedure (center) assigns a weight to each structure of the ensemble subset to reproduce the Micro-C contact matrix binned to the nucleosome positions of the family (left). The resulting in-silico Micro-C matrix can be seen on the right.

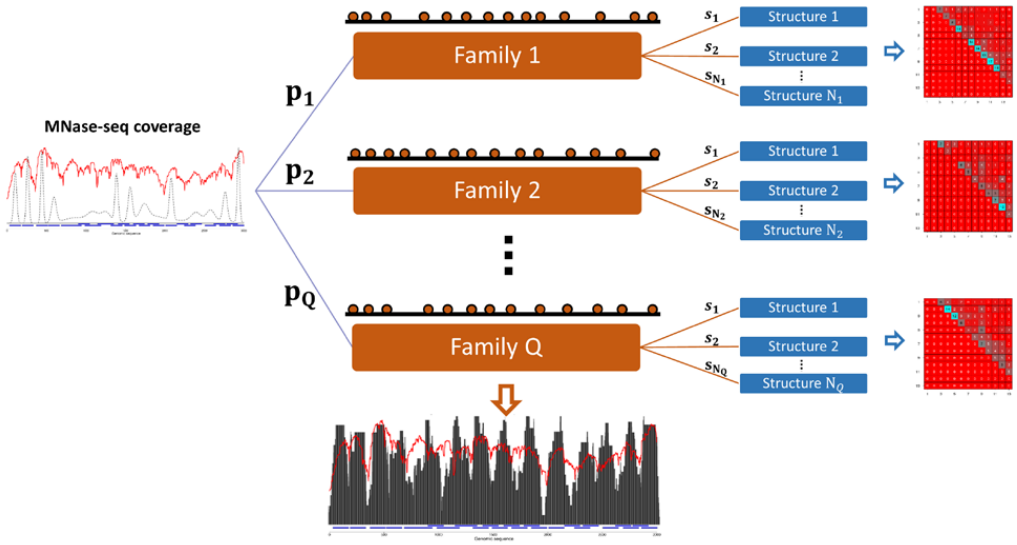


Figure 5. Workflow of Mnase-seq deconvolution and Micro-C fitting. The generated 1D families are assigned weights p to recover the experimental nucleosome signal. A simulation of each nucleosome family is performed to fit a set of 3D configurations to the experimental Micro-C matrix binned to the corresponding nucleosome positions using a reweighting procedure (of the weights s_i) of the relevant fiber conformations.

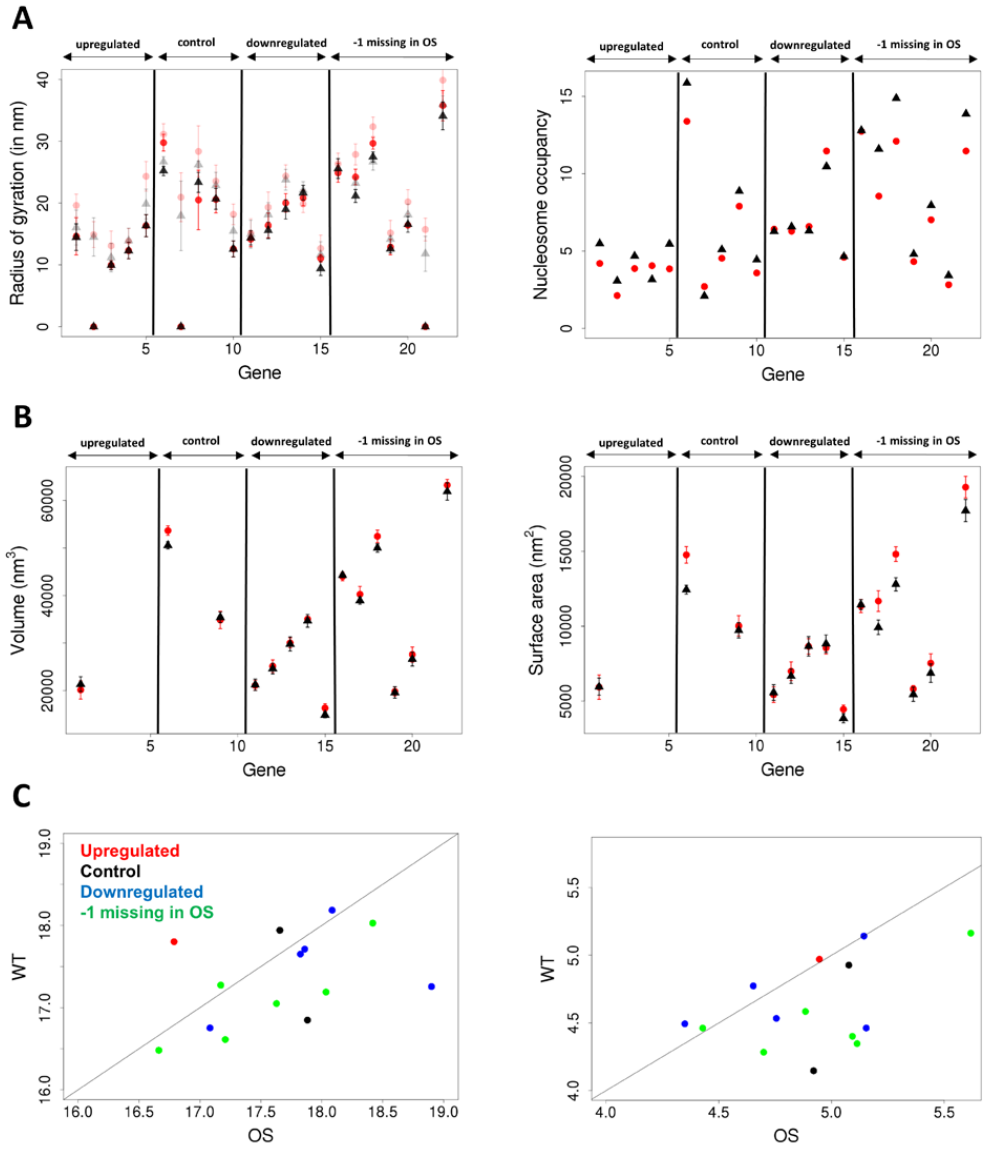


Figure 6. Physical properties of genes in oxidative stress (OS) compared to wildtype (WT). The five most upregulated genes (from WT to OS), five genes where expression remains unchanged, the five most downregulated genes and seven genes where the -1 nucleosome is missing in OS were selected to do an analysis of compaction (gene names are given in Supplementary Table S2). A: Radius of gyration (left) of the whole simulated ensemble of all families (light) and with

Micro-C fitting of the ensemble subset (bold) in OS (red) and WT (black) condition. In some cases, the number of nucleosomes in the gene was too low, so the fitting to Micro-C could not be performed (indicated as 0). Standard deviation was calculated as the variance of all families. Right: Nucleosome occupancy of all genes according equ. 5. B: Volume (left) and transcription factor accessible surface area (right) of the genes that could be fit to Micro-C in OS (red) and WT (black) condition. C: Normalized (by the number of nucleosomes) volume (left) and transcription factor accessible surface area (right) of OS versus WT. The black line represents the diagonal.

Supplementary Information

A METHOD TO PREDICT CHROMATIN FIBER CONFORMATIONS BY DECONVOLUTION OF NUCLEOSOME POSITIONING DATA AND MICRO-C

Jürgen Walther[1], Pablo D. Dans[1], Manuel Sarmiento[1], Rafael Lema[1], Isabelle Brun-Heath[1] and Modesto Orozco[1,2]*

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology.
[2] Department of Biochemistry and Biomedicine. The University of Barcelona. Barcelona, Spain.

* Correspondence to Prof. Modesto Orozco: modesto.orozco@irbbarcelona.org

Supplementary Methods

Simulation of the chromatin fiber model. In the following the ensemble generation of the chromatin fiber model for given nucleosome positions and DNA sequence is described (see Figure 1).

1) A Monte Carlo move is performed on randomly selected bps of the unbound DNA. For each MC move one to four inter base pair parameters are randomly selected to be modified. The strength of the change is determined by a scaling factor which is dependent on the diagonal entry of the stiffness matrix of the inter base pair parameter and which is scaled to guarantee ~40 % acceptance rate in the case of simulation of free DNA solely considering the elastic potential.
2) The inter base pair coordinates collected were transformed to derive Cartesian representations of the center of each base pair and the phosphates in the backbone of the DNA. The starting configuration was obtained using the SCHNArP algorithm (1) and an efficient coordinate transformation algorithm was used for the rigid body motion after each MC move which guarantees fast transformation between helical and Cartesian space.
3) The total energy of the system E^{tot} is evaluated. We calculate the long-range interactions E^{DH} and E^{LJ} of every i-th base pair with each other and with the nucleosomes to increase computational efficiency without influencing model accuracy (nucleosome core charges are then scaled accordingly; we chose i=7 to ensure that the fiber does not overlap sterically in between every i-th base pair, parameters for i=7 can be seen Supplementary Table 1). The MC move is then accepted or rejected based on the Metropolis algorithm.

Step 1) – 3) are repeated several million times until enough configurations are sampled. Configurations are drawn after a certain number of Monte Carlo moves defined by the run test method algorithm (2) (http://inka.mssm.edu/~mezei/mmc) to guarantee uncorrelated samples.

For efficient sampling of the conformational space of the chromatin fiber, a short test run is performed and 1000 configurations are drawn equidistantly along the test run (~1M Monte Carlo steps) as new starting configurations for 10 independent simulations to obtain in total 10,000 conformations (~10M Monte Carlo steps). The ensemble is transferred to GROMACS (3) trajectory format to be able to use the GROMACS built-in suite of analysis tools.

Supplementary Tables

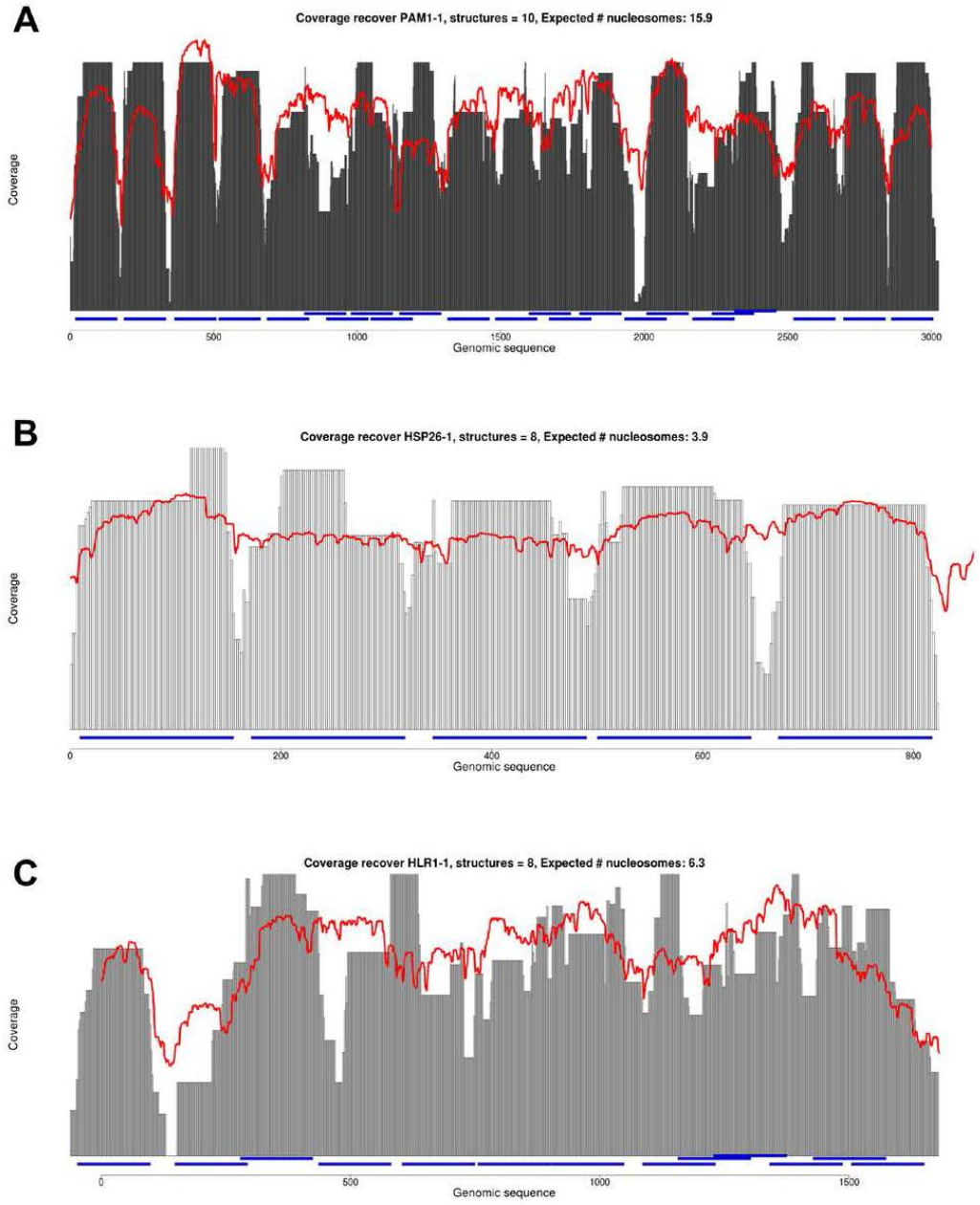
Parameter	Value
$q_{nucleff} = \frac{q_{nucl}}{i}$	$\frac{94e}{7} = 13.4\,e$
q_{bp}	- e
$4\pi\varepsilon_0$	$1.1126 \cdot 10^{-10}$ eV/Å
ε	80
κ	0.103 1/Å (for 100mM monovalent salt)
$\varepsilon_{nucl-nucl}$	1kT = 0.0257 eV
$\varepsilon_{nucl-bp}$	1kT = 0.0257 eV
ε_{bp-bp}	1kT = 0.0257 eV
$\sigma_{nucl-nucl}$	71.0 Å
$\sigma_{nucl-bp}$	24.3 Å
σ_{bp-bp}	12.9 Å

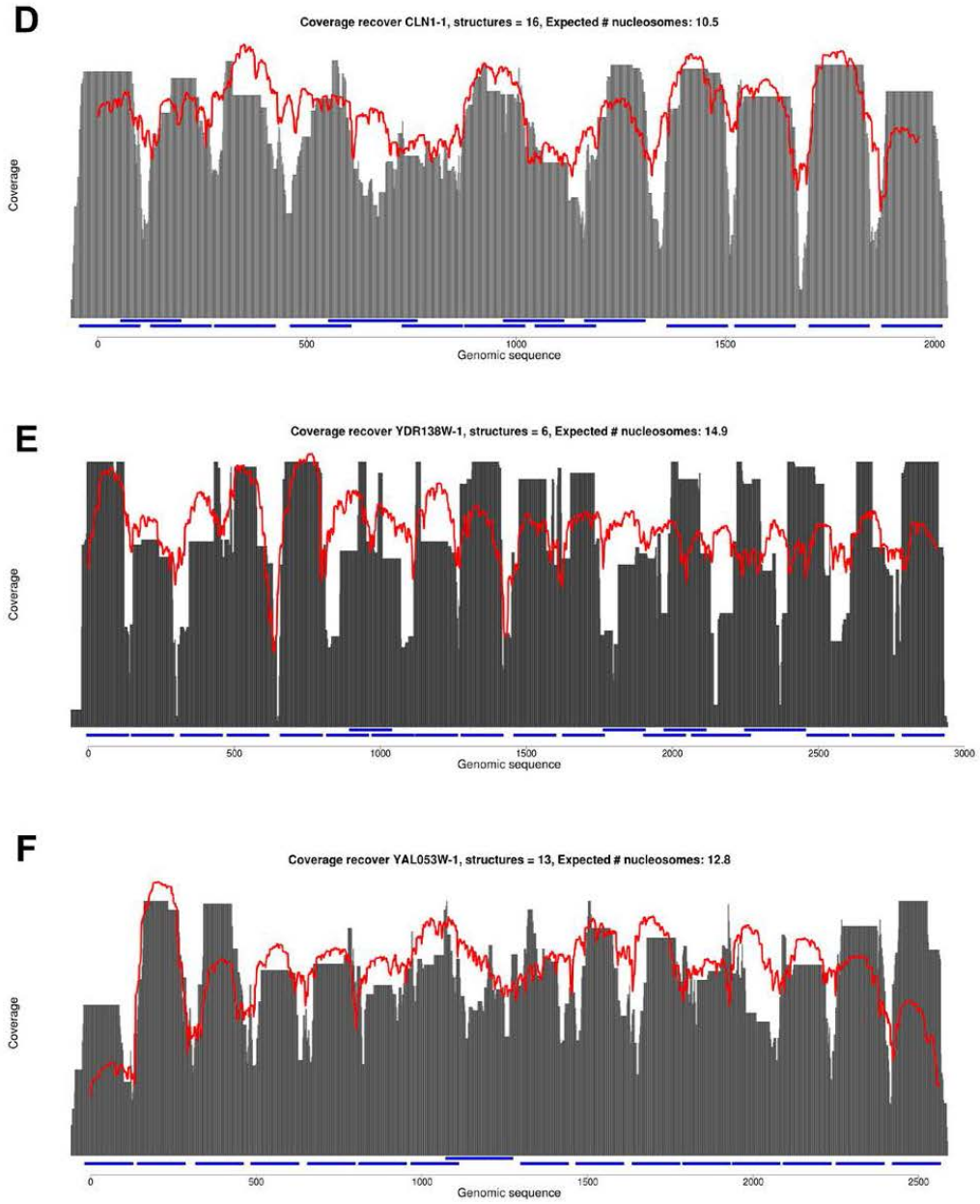
Supplementary Table 1. Parameter values for potentials of nucleosome fiber simulation. For $q_{nucleff}$ the division is by i and q_{bp} = -e because only the phosphate with the shortest distance to its interaction partner is evaluated for energy calculation.

Gene name	Index
HSP30	1
TDH3	2
ACT1	3
ALG9	4
HHO1	5
PAM1	6
HSP31	7
HSP26	8
GPX1	9
HSP42	10
PCL1	11
YOX1	12
HLR1	13
CLN1	14
HSH49	15
YAL053W	16
YBR169C	17
YDR138W	18
YGL096W	19
YLL011W	20
YLR350W	21
YOR098C	22

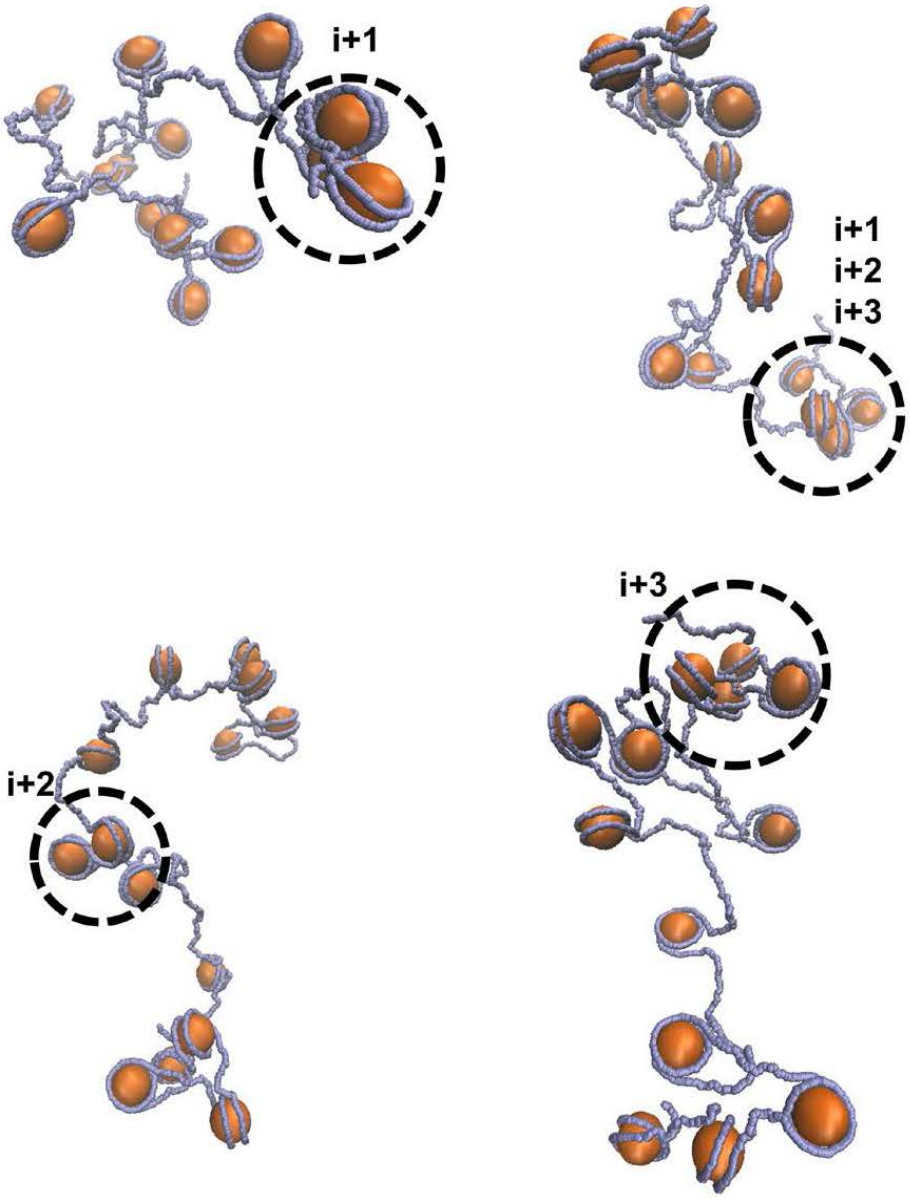
Supplementary Table 2. Gene names for the analyzed genes

Supplementary Figures

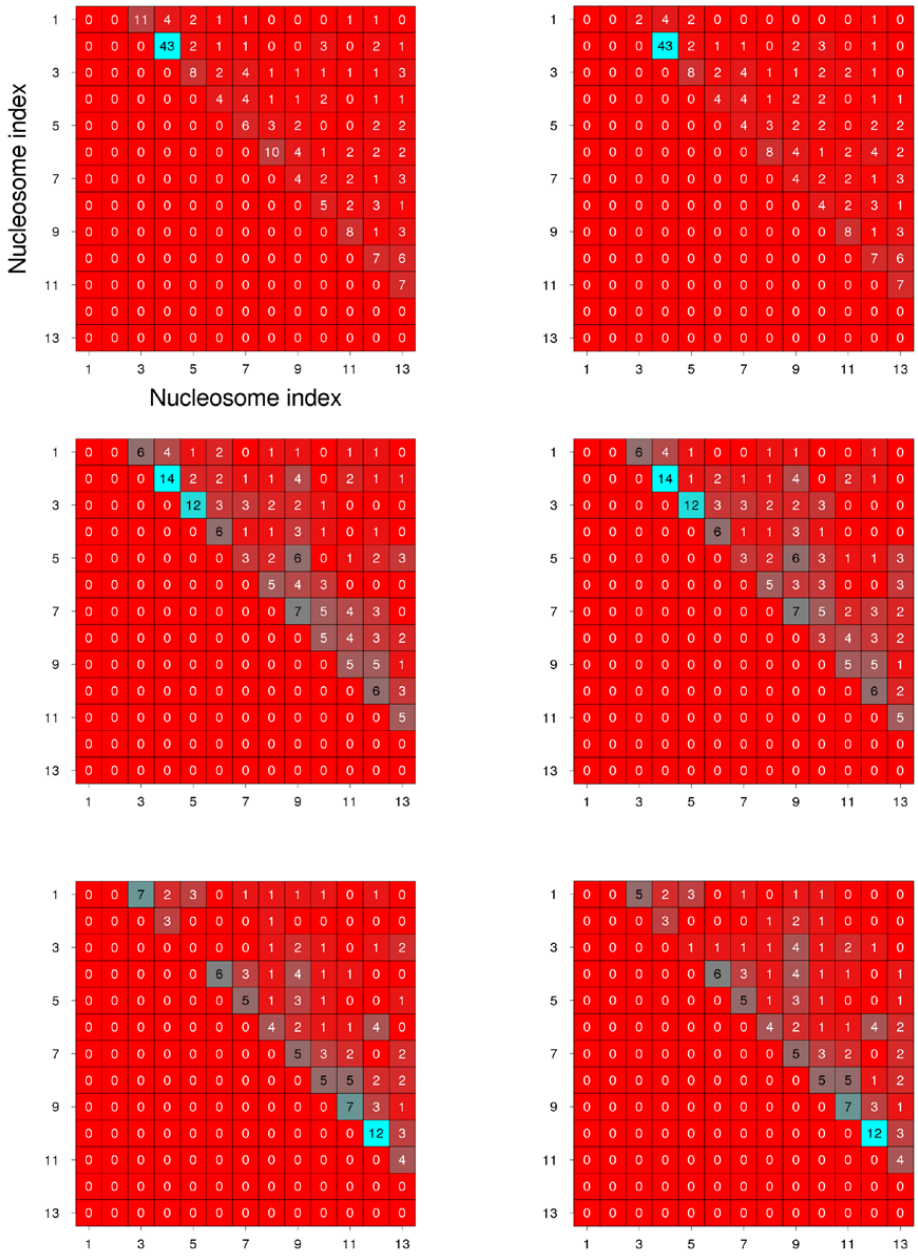

A
Coverage recover PAM1-1, structures = 10, Expected # nucleosomes: 15.9
Coverage
0
500
1000
1500
2000
2500
3000
Genomic sequence
B
Coverage recover HSP26-1, structures = 8, Expected # nucleosomes: 3.9
Coverage
0
200
400
600
800
Genomic sequence
C
Coverage recover HLR1-1, structures = 8, Expected # nucleosomes: 6.3
Coverage
0
500
1000
1500
Genomic sequence


D
Coverage recover CLN1-1, structures = 16, Expected # nucleosomes: 10.5
Coverage
0
500
1000
1500
2000
Genomic sequence
E
Coverage recover YDR138W-1, structures = 6, Expected # nucleosomes: 14.9
Coverage
0
500
1000
1500
2000
2500
3000
Genomic sequence
F
Coverage recover YAL053W-1, structures = 13, Expected # nucleosomes: 12.8
Coverage
0
500
1000
1500
2000
2500
Genomic sequence

Supplementary Figure S1. Reconstruction by families (black) of the experimental nucleosome coverage (red) of different genes. A: PAM-1. B: HSP26. C: HLR1. D: CLN1. E: YDR138W. F: YAL053W.


i+1
i+1
i+2
i+3
i+2
i+3

Supplementary Figure S2. Representative snapshots of the subset of the simulated ensemble of a family of the PAM-1 gene (gene number 6 in Figure 6). Contacts appear within the dashed circles.



Supplementary Figure S3. Reconstructed Micro-C matrices (left) compared to experiment (right) of the PAM-1 gene (gene number 6 in Figure 6). Each row corresponds to the fitting of a different family.

Supplementary References

1. Lu,X.-J., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP) 1 1Edited by K. Nagai. *J. Mol. Biol.*, **273**, 681–691. https://doi.org/10.1006/jmbi.1997.1345 http://www.ncbi.nlm.nih.gov/pubmed/9356256

2. Sun,J., Zhang,Q. and Schlick,T. (2005) Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *Proc. Natl. Acad. Sci.*, **102**, 8180–8185. https://doi.org/10.1073/pnas.0408867102 http://www.ncbi.nlm.nih.gov/pubmed/15919827

3. Abraham,M.J., Murtola,T., Schulz,R., Páll,S., Smith,J.C., Hess,B. and Lindahl,E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1–2**, 19–25. https://doi.org/10.1016/J.SOFTX.2015.06.001

Bibliography for Chapter III

1. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678. https://doi.org/10.1093/nar/gks884 http://www.ncbi.nlm.nih.gov/pubmed/23012264

2. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320. https://doi.org/10.1093/nar/gku809 http://www.ncbi.nlm.nih.gov/pubmed/25223784

3. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V., Šponer,J. and Lankaš,F. (2013) Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721. https://doi.org/10.1021/ct300671y

4. Pérez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.*, **45**, 196–205. https://doi.org/10.1021/ar2001217 http://www.ncbi.nlm.nih.gov/pubmed/21830782

5. Ben Imeddourene,A., Elbahnsi,A., Guéroult,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLOS Comput. Biol.*, **11**, e1004631. https://doi.org/10.1371/journal.pcbi.1004631

6. Parrinello,M. and Rahman,A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, **52**, 7182–7190.
https://doi.org/10.1063/1.328693

7. Whelan,D.R., Hiscox,T.J., Rood,J.I., Bambery,K.R., McNaughton,D. and Wood,B.R. (2014) Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation. *J. R. Soc. Interface*, **11**, 20140454–20140454.
https://doi.org/10.1098/rsif.2014.0454
http://www.ncbi.nlm.nih.gov/pubmed/24898023

8. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
https://doi.org/10.1093/nar/gku855
http://www.ncbi.nlm.nih.gov/pubmed/25260586

9. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M., *et al.* (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
https://doi.org/10.1093/nar/gkv1301

10. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 11163–8.
https://doi.org/10.1073/pnas.95.19.11163
http://www.ncbi.nlm.nih.gov/pubmed/9736707

11. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.*, **85**, 2872–2883.
https://doi.org/10.1016/S0006-3495(03)74710-9
http://www.ncbi.nlm.nih.gov/pubmed/14581192

12. Colizzi,F. and Bussi,G. (2012) RNA Unwinding from Reweighted Pulling Simulations. *J. Am. Chem. Soc.*, **134**, 5173–5179.
https://doi.org/10.1021/ja210531q

13. Minary,P. and Levitt,M. (2010) Conformational optimization with natural degrees of freedom: a novel stochastic chain closure algorithm. *J. Comput. Biol.*, **17**, 993–1010.
https://doi.org/10.1089/cmb.2010.0016
http://www.ncbi.nlm.nih.gov/pubmed/20726792

14. Luger,K., Dechassa,M.L. and Tremethick,D.J. (2012) New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.*, **13**, 436–47.
https://doi.org/10.1038/nrm3382
http://www.ncbi.nlm.nih.gov/pubmed/22722606

15. Grigoryev,S.A., Arya,G., Correll,S., Woodcock,C.L. and Schlick,T. (2009) Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci.*, **106**, 13317–13322.
https://doi.org/10.1073/pnas.0903280106

16. Hsieh,T.-H.S., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, **162**, 108–119.
https://doi.org/10.1016/j.cell.2015.05.048

http://www.ncbi.nlm.nih.gov/pubmed/26119342

17. Wiese,O., Marenduzzo,D. and Brackley,C.A. (2018) Nucleosome positions alone determine micro-domains in yeast chromosomes. *bioRxiv*, 10.1101/456202. https://doi.org/10.1101/456202

18. Hansen,J.C., Ausio,J., Stanik,V.H. and van Holde,K.E. (1989) Homogeneous reconstituted oligonucleosomes, evidence for salt-dependent folding in the absence of histone H1. *Biochemistry*, **28**, 9129–36. http://www.ncbi.nlm.nih.gov/pubmed/2605246

19. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing Chromosome Conformation. *Science (80-.).*, **295**, 1306–1311. https://doi.org/10.1126/science.1067799 http://www.ncbi.nlm.nih.gov/pubmed/11847345

CHAPTER IV - DISCUSSION

1. Sequence-dependent properties of B-DNA and structural polymorphisms

Experimental and computational data show that DNA dynamics is far from that of a homogeneous linear rod, and sequence plays a dramatic role in determining the physical properties of the duplex. An elegant way to determine flexibility properties of DNA is via the helical parameter space. In the inter base pair parameter space it can be clearly seen that each of the ten unique base pair steps prefers distinct internal geometries. While some conformations such as d(ApT) base pair step exhibit harmonic behavior in the inter base pair parameters, others can occupy two (or more) configurational sub-states in some inter base pair parameters such as d(CpG) in twist. Extensive simulations have shown us that the multimodality is not only a property of the base pair step, but it is also modulated by the neighbors, leading to a tetramer model of deformability. Extensive multi-μs MD simulations using parmbsc1 force field of all 136 unique tetranucleotides show that around 80% of the inter base pair parameter distributions of all tetranucleotides cannot be correctly described using a single normal distribution. For example, bimodality in shift is generally coupled to the appearance of high shift values (above 1 Å). These anharmonic deformations in helical geometries are particularly prevalent for certain tetranucleotide sequence contexts, and are always coupled to a complex network of coordinated changes in the backbone, with BI/BII equilibria being a major determinant. Some of these correlations can be summarized in an extended set of rules. For example RR backbones exhibit high BII levels contrary to YY which are biased towards the BI state, generating a strong asymmetry at RR:YY steps. In general for all tetramers there is a tendency of larger shift values with increased BII percentage of the central junction and a strong correlation between the backbone state transition BI -> BII of the central junction and the formation of an unconventional hydrogen bond of the type CH—O. In summary, the carried out analysis of this work leads to a detailed scheme with strong predictive power of DNA geometrical properties at the tetranucleotide level. The power for each tetranucleotide to predict internal geometries such as the multimodal nature of some inter base pair parameters

and complex couplings such as the connection between backbone sub-states and helical state can turn out to be a good starting point to refine standard helical coarse grain models of DNA flexibility based on a harmonic approximation by using a multimodal approach to display the internal conformation at higher detail.

A logical consequence after the study of all 136 unique tetranucleotides would be to study DNA in all its unique hexanucleotide environments, however the existence of more than 800 unique hexanucleotides would make a computational analysis very costly and probably too difficult to draw general conclusions compared to the tetranucleotide study. For this reason, we decided to focus on a single case of special complexity: the central d(TpA) step in the highly polymorphic CTAG tetranucleotide in different hexa- and octamer environments. We found that the conformational landscape in the distinct neighboring environments was formed by concerted and correlated movements of backbone and bases. More specifically, the previously shown correlation of BI/BII inter-conversion with the formation of specific hydrogen bond contacts between adjacent bases of type CH—O at the tetrameric level seems to be crucial in the propagation of structural information (especially due to frustration driven by the mechanical limitations imposed by DNA's crankshaft motions) at hexa- and octanucleotide level. Also the connection of backbone (BI/BII) and base (mostly shift and twist) polymorphisms can be spotted up to the octamer level. In summary, those long-range effects indeed modulate subtly the geometrical properties of the central d(TpA) step at both hexa- and octanucleotide level giving a possible explanation on how structural information can travel almost half a helical turn away from the central junction. The results also indicate a high detail description of DNA flexibility for the CTAG tetranucleotide that extends beyond the nearest neighbor model while in general nearest neighbor models which describe DNA as a chain of tetranucleotides can accurately explain the described remote effects in longer sequences.

2. A helical coarse grain model of B-DNA dynamics and its web implementation

Even though the study of the CTAG tetranucleotide suggests higher than nearest-neighbor effects of DNA flexibility for some tetranucleotides, a complete study of all unique sequence

environments is far from becoming reality due its immense computational cost. For this reason we decided to build a helical coarse grain model of DNA based on inter base pair coordinates exploiting the nearest neighbor sequence effect of DNA flexibility previously studied by atomistic MD simulations with parmbsc1 force-field. Using machine learning approaches – dimension reduction using PCA in helical space followed by a clustering algorithm - we could deconvolute the inter base pair parameter distributions of each tetranucleotide into several harmonic helical sub-states to correctly capture the conformational space sampled by MD which constitutes a significant improvement to the commonly used model with the use of only one Gaussian to describe the inter base pair parameter space, also referred to as the standard harmonic model. The energy function of the extended nearest neighbor model is motivated by valence bond theory and completely converges to the standard harmonic approach if only a single helical state is considered. The extended nearest neighbor model is coupled to a Metropolis Monte Carlo sampling algorithm in the inter base pair parameter space and the sampled configurations can be mapped from helical space to the fully atomistic level taking advantage of the correlation between helical states and backbone configuration. The resulting structures show very high similarity to global dynamics of atomistic MD simulations when comparing Boltzmann weighted absolute similarity indices. Several sequence-dependent properties such as backbone sub-state populations, groove widths and sugar puckering are reproduced with high fidelity and a comparison to experimental structures yields a good agreement in RMSd/bp and average inter base pair parameters. The high computational efficiency allows the treatment of DNA segments at time scales up to 10^5 times faster than conventional atomistic MD and offers simulations of long DNA stretches at unprecedented detail not reachable by atomistic MD. The implementation of the algorithm in a simple web interface and as a stand-alone package enables expert and non-expert users an easy access to this model.

A more elaborate web environment allows direct online simulation and analysis of trajectories simulated with our coarse grain model. In this implementation the user is given the option to simulate - apart from free linear B-DNA - DNA dynamics in a constrained environment such as supercoiled or protein-coated DNA. The analysis of the resulting trajectories is performed directly with the tool and can be viewed in interactive plots in the webserver. Analysis tools range from local helical analysis, distance contact maps, end-to-end distances and elastic energies to specific descriptors such as circular parameters or virtual DNA footprinting. To our knowledge, this

webserver constitutes the first of its kind to offer the simulation of trajectories at atomistic detail in-situ with an integrated analysis pipeline.

3. Development of a nucleosome fiber model

Due to recent improvements in experimental techniques there is increased knowledge of the general 3D genome organization at the nucleosome scale. However, the precise secondary structure of chromatin depends on the cell type and other internal and external factors, and is still controversial ranging from the detection of nucleosome clutches in human cells via STORM microscopy to the arrangement of a few genes into self-interaction domains detected in yeast by Micro-C suggesting that the nucleosome arrangement along the genomic sequence plays a crucial role in secondary chromatin structure. To probe chromatin dynamics at kb scale in different conditions, we designed a mesoscopic nucleosome fiber model at base pair resolution which is coupled with a Metropolis Monte Carlo sampling algorithm. The method is flexible enough to incorporate experimental data such as in-vivo nucleosome positions along the genomic sequence and three dimensional structural restraints derived from state-of-the-art experimental techniques such as STORM or Micro-C to refine the ensemble of simulated structures. This bottom-up model uses the mesoscopic helical DNA model for the description of linker DNA and assumes the nucleosome to be rigid and fixed in the sequence, with the nucleosomal DNA fixed to the DNA path of the high-resolution X-ray structure (PDB 1KX5) and with a simple coarse grained description of charges and steric constraints of the histone core. The nucleosome fiber model reproduces well in-vitro experiments of salt-dependent sedimentation coefficients and is able to show a fiber volume similar to that obtained experimentally.

Standard nucleosome fiber models assume a regularly spaced nucleosome distribution, which is not what is experimentally found. Thus we implement a method of nucleosome positioning based on the analysis of experimental MNase-seq maps, which provide averaged pictures of the preferred nucleosome positions in a pool of cells. To transform these average maps into individual nucleosome architectures which can be then used to bias the coarse grained simulations we developed a machine learning algorithm to deconvolute the MNase-seq signal of a genomic segment into a small number of physically realistic 3D fiber configurations representing individual

cells. The combination of the different configurations recovers the experimental signal, but avoiding the generation of physically unrealistic fiber arrangements.

With the emergence of 3C-based techniques the 3D arrangement of a population of cells can be represented via a contact matrix, with Hi-C providing information of genomic fragments of down to 1kb resolution and Micro-C providing even nucleosome level resolution. In terms of the nucleosome fiber model the Micro-C data of a genomic segment can be used to refine the ensemble of structures of each of the different physically realistic configurations arising from the deconvolution of the MNase-seq maps. A filtering procedure followed by structure reweighting ensures that a small set of relevant conformations describe the Micro-C contact matrix.

In summary, the nucleosome fiber model can accurately reproduce in-vitro chromatin compaction data. It is also used to provide physically realistic fiber conformations of genomic segments based on MNase-seq maps. The fitting of the ensembles of the simulated structures of the individual deconvoluted states is the last step of a rigorous description of second order chromatin organization by implementing experimental data at different dimensional level together with a bottom-up coarse grain model at bp resolution.

4. VRE implementation

The Virtual Research Environment (VRE) developed by the Multiscale Complex Genomics (MuG) consortium is a web environment in which tool developers can implement their programs related to genome structure at all resolution levels. A striking advantage consists in the fact that all those tools can be executed from a single user workspace where all the input and output data after tool execution can be viewed. This also enables the user to execute different tools and interconnect its output data. I integrated the helical DNA coarse grain model (MCDNA) as well the nucleosome fiber model (ChromatinDynamics) into the VRE to enable the user the simulation of unrestrained DNA and chromatin structure up to kb scale (see Figure 31).

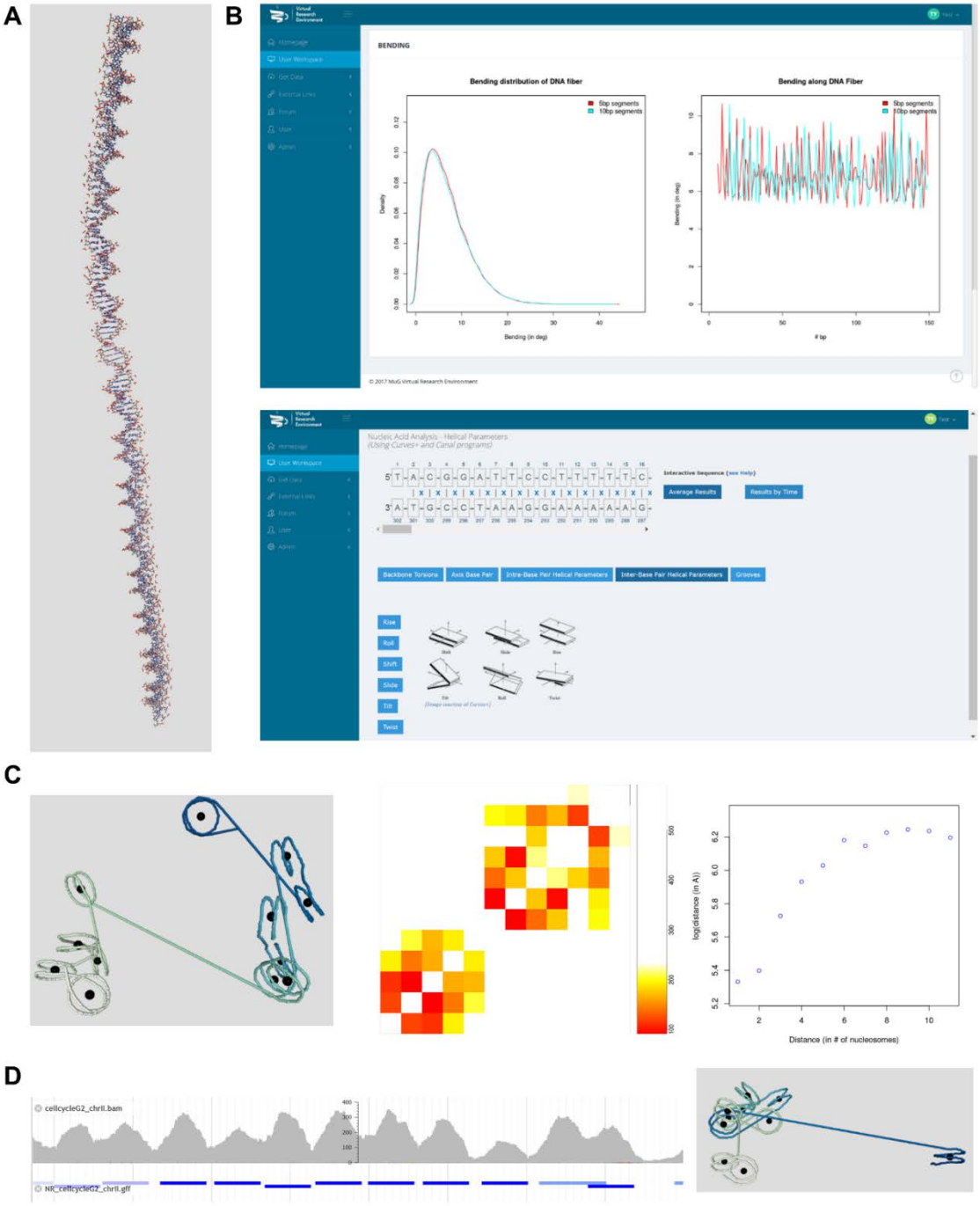


Figure 31. Snapshots of VRE output of MCDNA and ChromatinDynamics. A: Snapshot of a DNA fragment of 150bp in length simulated with MCDNA. B: Screenshots of bending analysis within MCDNA (top) and NAFlex (bottom). C: From left to right: Snapshot of a chromatin fiber created with ChromatinDynamics, its nucleosome distance matrix and internucleosomal distance graph. D: Nucleosome calls obtained with nucleR of a MNase-seq profile (left) is transformed into a 3D representation of the nucleosome fiber via ChromatinDynamics (right).

The tools make use of the interconnectivity, as for example the generated trajectory of the DNA coarse grain model at atomistic detail can be used as input for the NAFlex analysis suite where global and local DNA features can be analyzed in more detail. Another example is the direct connection between MNase-seq data and the nucleosome fiber model. The analysis of a MNase-seq experiment by the tool nucleR outputs nucleosome positions of a genomic fragment which can directly be used to construct a three-dimensional representation of said fragment with the nucleosome fiber model. This enables the user to instantly visualize genomic segments in 1D and 3D and avoids complex data format conversions between the output and input of different tools. In summary, the integration of the two models I developed into the VRE constitutes a step forward towards large-scale availability, usability and interconnectivity with other tools.

CONCLUSIONS

In this work scientific advances in every resolution level of the multi-scale simulation of DNA are achieved reaching from atomistic MD simulations to mesoscopic secondary chromatin structure modeling. We show that the theoretical description of DNA dynamics mesh together like cogs among different resolution levels. We developed a force-field for the accurate description of atomistic DNA dynamics based on quantum mechanical simulations. With the accuracy of parmbsc1, sequence-dependent effects of B-DNA beyond the base pair level were described and used as a starting point to parametrize a novel helical coarse grain model which shows similar accuracy to the DNA dynamics obtained by atomistic MD, but at much lower computational cost. In the nucleosome fiber model the coarse grain DNA algorithm is used for the linker DNA description and along with a simple mesoscopic characterization of the nucleosome chromatin dynamics can be probed at kb scale with a DNA model whose roots lie in the quantum mechanical regime.

The free availability of the developed helical DNA and nucleosome fiber model as stand-alone versions or integrated in a single webserver or large-scale online research environment platform correspond to the standards of today's research in terms accessibility and usability. In the following I will summarize the main conclusions of each topic in bullet points:

1. Parmbsc1 simulations of canonical B-DNA duplexes show that nearest neighbor effects strongly influence mechanical properties of the central base pair step with polymorphisms appearing in helical geometries for certain tetranucleotide sequence contexts, always coupled to coordinated changes in the backbone geometry.

2. The study of the highly polymorphic d(CpTpApG) tetranucleotide reveals that hexa- and octamer effects can influence helical dynamics at the central d(TpA) junction based on concerted and correlated movements of backbone and bases.

3. The newly developed coarse grain model for the simulation of duplex DNA accurately recapitulates the multi-harmonic nature of helical parameters in an extended nearest neighbor model, and with its ability of atomistic backmapping and smart backbone reconstruction it reproduces in high detail global and local dynamics simulated by MD or determined by experiments.

4. The multi-harmonic coarse grain model outperforms MD simulations in terms of simulation time by a factor of up to 10^5 and is available for free as a stand-alone version and in a simple web interface.

5. A more elaborate web environment called MCDNA allows direct online simulation and analysis of trajectories simulated with the coarse grain model of free duplex DNA and in a restrained environment such as supercoiled circular DNA or protein-bound DNA.

6. The new bottom-up nucleosome fiber model correctly reproduces chromatin compaction of short and long chromatin fibers and can be used to derive chromatin fibers with biologically and physically realistic nucleosome positioning based on MNase-seq and Micro-C maps.

7. The integration of the mesoscopic DNA model and the bottom-up nucleosome fiber model into the MuG-VRE constitutes a step forward towards large-scale availability, usability and interconnectivity with other tools.

RESUMEN EN ESPAÑOL

El estudio del ADN desde la escala atómica a la mesoscópica y la conexión entre dichos niveles de resolución constituye uno de los desafíos mayores del nuevo milenio. Desde el inicio del siglo XX, diversos experimentos han permitido elucidar la estructura del nucleosoma a escala atómica, y por otro lado capturar los contactos entre segmentos del genoma cuyas secuencias se encuentran muy alejadas. En paralelo, el desarrollo teórico de campos de fuerza para la simulación de sistemas atomísticos logró su primera madurez con la publicación de parmbsc0 en 2007, al tiempo que empezaron a salir publicados los primeros modelos de grano grueso para representar fibras de nucleosomas. La primera década del presente milenio termina con uno de los experimentos que considero personalmente de los más destacados a la hora de visualizar el genoma completo: Hi-C. Actualmente, a casi 10 años del advenimiento del Hi-C, la estructura del núcleo celular sigue siendo un campo muy activo. Es ahora el momento justo para cosechar de los frutos plantados por los pioneros una década atrás y trabajar en la conexión entre los diferentes niveles de resolución logrando una imagen completa y global del ADN en el núcleo celular desde los electrones hasta los cromosomas.

En este trabajo, usamos una aproximación computacional para integrar los diferentes niveles de resolución, desde simulaciones atomísticas de Dinámica Molecular hasta el modelado de fibras de cromatina. Desarrollamos un campo de fuerza atomístico que reproduce de forma exacta la dinámica del ADN, basado en cálculos de mecánica cuántica. Gracias a la exactitud de parmbsc1, los efectos estructurales secuencia-dependientes a nivel atómico fueron capturados y usados como parámetros para desarrollar un nuevo modelo helicoidal de grano grueso que ha mostrado una exactitud similar con un coste computacional mucho menor. En el modelo de fibra de cromatina, el modelo de grano grueso mencionado anteriormente es usado para simular el comportamiento del ADN “linker” (libre) entre los nucleosomas que son representados de forma simple pero que permiten estudiar fibras a la escala de kilobases (kb) con un modelo basado en la mecánica cuántica.

Sumado a lo anterior, y para hacer nuestros modelos y herramientas disponibles y accesibles de acuerdo a los estándares actuales, los modelos y métodos desarrollados en esta tesis se

distribuyen de forma libre como una versión “stand-alone” o integrado en una plataforma de investigación online.

Los capítulos de la presente tesis están organizados de la siguiente manera: En el Capítulo I se introduce de forma general los conceptos teóricos y experimentales comunes a los estudios realizados, desde la estructura del ADN hasta la organización de la cromatina en el núcleo celular. En el Capítulo II, se expanden y detallan algunos de los conceptos específicos necesarios para comprender e interpretar los resultados obtenidos en esta tesis. Los resultados se presentan en el Capítulo III en forma de compilado de siete artículos (publicados o en vía de publicar) que se organizan en tres sub-secciones: i) Un estudio de las propiedades físicas y dinámicas del ADN dependientes de secuencia (aplicando parmbsc1), y el análisis detallado de los polimorfismos estructurales a nivel de las nucleobases. ii) El desarrollo de un nuevo modelo helicoidal de grano grueso para simular y modelar secuencias de ADN en forma B basado en el conocimiento de la mecánica del ADN obtenido en (i), y su ejecución a través de un sitio web. iii) La implementación de un modelo de fibra de nucleosomas capaz de predecir conformaciones realistas compatibles con las que se pueden encontrar en núcleo celular. La discusión de todos los resultados generados es presentada en el Capítulo IV, junto a las conclusiones al final del presente manuscrito de tesis.

A continuación se encontra un resumen de los resultados más importantes obtenidos en esta tesis.

Efectos estructurales secuencia- dependientes del ADN en forma B más allá del par de bases

Estudios experimentales y teóricos han mostrado invariablemente que la dinámica del ADN no puede representarse adecuadamente usando un polímero lineal homogéneo. La secuencia del ADN tiene un efecto dramático a la hora de determinar las propiedades físicas de la doble hebra. Es posible estudiar la flexibilidad del ADN en un sistema de coordenadas especial, llamado espacio helicoidal. En ese espacio es claramente visible que las propiedades del ADN varían con la composición del par de bases (pasos). La mayoría de los pasos del ADN muestran un comportamiento armónico en los parámetros helicoidales, mientras que algunos pueden presentar distribuciones multimodales. A través de la simulación en los microsegundos de todos los posibles tetrámeros de ADN (136 combinaciones), hemos logrado desarrollar una serie de

reglas que permiten predecir el comportamiento promedio de cualquier ADN de doble hebra, y la aparición de polimorfismos estructurales. Por ejemplo, hemos mostrado como secuencias con dos purinas (RR) exhiben altas proporciones de la conformación BII, mientras que dos pirimidinas consecutivas (YY) favorecen el estado BI. El poder de las reglas derivadas en este estudio, es que permiten el desarrollo de modelos de grano grueso helicoidales que vayan más allá del modelo armónico y permitan reproducir estados multimodales describiendo la estructura interna con mayor detalle y exactitud.

Para estudiar el efecto de la secuencia más allá de los tetranucleótidos, centramos nuestros esfuerzos en un paso particular y complejo: d(TpA) embebido en el tetranucleótido altamente flexible y polimórfico CTAG y todos los posibles entornos de hexanucleótidos y octanucleótidos. Encontramos que el espacio conformacional en los distintos entornos podía comprenderse en términos de movimientos concertados y correlacionados entre las bases y la cadena de azúcar-fosfato. En concreto, se mostró cómo la correlación entre el sub-estado BI/BII y la formación de un enlace de hidrógeno del tipo CH-O a nivel del tetranucleótido era crucial para comprender la propagación de información estructural a través de la doble hebra de ADN. En resumen, mostramos cómo efectos de “largo alcance” eran capaces de modular sutilmente las propiedades estructurales del paso central d(TpA), dando una posible explicación a la manera en la que la información mecánica viaja por la hebra de ADN.

Un modelo mesoscópico de ADN y su implementación en un servidor web

Hemos desarrollado un modelo de grano grueso de ADN basado en las coordenadas de un par de bases de Watson-Crick, haciendo uso de los efectos que tienen los dos vecinos en la flexibilidad de los pasos que fueron estudiados a nivel atómico en la sección anterior. Usando una aproximación de “machine learning” (reducción del espacio helicoidal por técnicas de componentes principales seguido de métodos de agrupamiento “clustering”) hemos llevado a cabo una deconvolución de las distribuciones de cada parámetro helicoidal para cada tetranucleótido en varios sub-estados representados armónicamente (multi-modales). Esto representa una mejora significativa sobre los modelos armónicos tradicionales. El modelo desarrollado está acoplado a un algoritmo de muestreo Metropolis Monte Carlo en el espacio conformacional de los pares de bases, y las configuraciones obtenidas en el espacio helicoidal

pueden reconstruirse completamente a nivel atómico teniendo en cuenta el conocimiento generado sobre las correlaciones entre los sub-estados de las bases y la configuración de la cadena de azúcar-fosfato. Las estructuras resultantes muestran una alta similitud con las estructuras obtenidas en las dinámicas moleculares atomísticas. Varias propiedades dependientes de la secuencia como los sub-estados BI/BII, las dimensiones de los surcos, y la conformación de los azúcares se logran reproducir con alta fidelidad cuando se compara con estructuras experimentales. La gran eficiencia computacional del modelo de grano grueso desarrollado permite simular segmentos de ADN 10^5 veces más rápido que la dinámica atomística convencional. La implementación del modelo y algoritmos en una interfase web y como un paquete independiente permite el uso de nuestro desarrollo por la comunidad no-experta. Asimismo, se elaboró en base al mismo modelo un segundo sitio web más detallado que permite la simulación y el análisis de las trayectorias. En dicha implementación web avanzada, el usuario puede simular, además de ADN lineal, ADN en entornos constreñidos como en el caso del ADN circular o el ADN interactuando con proteínas. El análisis de las trayectorias se hace de manera interactiva, produciendo los gráficos de forma on-line.

Modelo de fibra de nucleosomas

Los recientes avances en las técnicas experimentales han permitido aumentar sensiblemente el conocimiento sobre la organización 3D del genoma a nivel de los nucleosomas. Sin embargo, la estructura secundaria precisa de la cromatina depende del tipo celular, y de otros factores externos, y sigue siendo muy controversial desde la detección de “nidos” de nucleosomas en células humanas usando microscopía STORM, hasta las conformaciones de unos pocos genes obtenidas a través de técnicas de Micro-C. Para poder estudiar la dinámica de fibras de nucleosomas en la escala de las kilobases (kb), diseñamos un modelo mesoscópico de fibra de cromatina acoplado a un algoritmo de muestreo basado en Metropolis Monte Carlo. El método desarrollado es lo suficientemente flexible como para incorporar datos experimentales como las posiciones de los nucleosomas obtenidas por técnicas in-vivo y restricciones geométricas experimentales derivadas de microscopia STORM o técnicas de Micro-C. Este modelo mesoscópico “bottom-up” usa el modelo de ADN presentado en la sección anterior para simular el comportamiento del ADN que conecta cada nucleosoma, manteniendo el ADN del nucleosoma fijo de acuerdo a una estructura experimental de rayos-X. El modelo reproduce correctamente

experimentos in-vitro como el coeficiente de sedimentación dependiente de la concentración de sales y se obtienen volúmenes de fibras en acuerdo con los datos experimentales.

Nuestro modelo se diferencia del resto al incorporar datos de MNase-seq para determinar la posición de los nucleosomas a lo largo de la secuencia. Hemos desarrollado un algoritmo de "machine learning" para deconvolucionar las señales que provienen de MNase-seq promedios - realizados sobre millones de células - en número bajo de configuraciones de fibras de cromatina físicamente realistas que vienen a representar las distribuciones de las células individuales. Esto permite obtener fibras de cromatina 3D con una distribución experimental de los nucleosomas a lo largo de la secuencia. Del mismo modo, con el advenimiento de la técnicas basadas en 3C, el arreglo 3D de una población de células puede representarse a través de una matriz de contactos que puede llegar a la resolución de un solo nucleosoma con técnicas como Micro-C. La matriz de contacto proveniente de Micro-C puede ser incorporada a nuestro modelo para refinar un conjunto de estructuras que contengan los contactos físicamente posibles y en acuerdo con los datos experimentales. Un procedimiento de filtrado y de peso de las mayores configuraciones 3D permite recuperar, con un conjunto relativamente pequeño de conformaciones, la matriz de Micro-C experimental.