Information Transfer and Dynamics of Nucleic Acids studied by Theoretical Approaches

Alexandra Balaceanu

UNIVERSITAT DE BARCELONA

Information Transfer and Dynamics of Nucleic Acids studied by Theoretical Approaches

Alexandra Balaceanu

UNIVERSITAT DE BARCELONA

FACULTAT DE QUIMICA

QUIMICA TEÒRICA I MODELIZATCIÓ COMPUTACIONAL

Information Transfer and Dynamics of Nucleic Acids studied by Theoretical Approaches

DIRECTOR	TUTOR	DOCTORAND
MODESTO OROZCO LOPEZ[1,2]	JUAN NOVOA VIDE[3]	ALEXANDRA BALACEANU

[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.
[3] Faculty of Quemistry, University of Barcelona, Spain.

UNIVERSITAT DE BARCELONA

FACULTAT DE QUIMICA

QUIMICA TEÒRICA I MODELIZATCIÓ COMPUTACIONAL

Information Transfer and Dynamics of Nucleic Acids studied by Theoretical Approaches

Alexandra Balaceanu

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

UNIVERSITAT DE BARCELONA

UNIVERSITAT DE BARCELONA

FACULTAT DE QUIMICA

QUIMICA TEÒRICA I MODELIZATCIÓ COMPUTACIONAL

Information Transfer and Dynamics of Nucleic Acids studied by Theoretical Approaches

Alexandra Balaceanu

DIRECTOR	TUTOR	SUPERVISOR
MODESTO OROZCO LOPEZ	JUAN NOVOA VIDE	PABLO DANS

Acknowledgements

Firstly, I am extremely grateful for the mentorship of my thesis director Prof. Modesto Orozoco, who made all my efforts take the shape of scientific research and from whom I have learnt so much in the course of my PhD.

A very special gratitude goes out to Pablo Dans, who was, as he himself put it, my Jedi master throughout this crazy journey: You were an inspiration and for that I thank you.

To Prof. Eric Westhof, who challenged my thinking and accepted nothing less than excellence from me.

To Prof. Juan Novoa, who always saw the best in me: I appreciate all the kind words and support. I also wish to thank Dr. Mercè Deumal, who was so very patient with me.

My eternal "cheerleader", Ricard Illa: I miss our interesting and long-lasting science chats. My great friend and genius, Diana Buitrago: You were always there to calm the waters when I was getting overwhelmed. To my very rebellious former labmate: You gave me space to vent and motivated me to see the bigger picture rather than the light at the end of the tunnel.

With a special mention to Xenia Villalobos, Leyre Caracuel and Clara Caminal, and all the administration team in general. It was fantastic to have the opportunity to work on outreach projects with you and see another side of scientific efort.

To my sister, who believes in me fiercely, but still affords to drag me down to common sense whenever I drift too far away. Daca nu sunt altceva, sunt cap de atom pentru tine! My forever interested, encouraging and always enthusiastic mother: you were always keen to know what I was doing and how I was proceeding. I thank you for that. I am also grateful to my other family members and friends who have supported me along the way.

And finally, last but by no means least, also to all the members past and present of the MMB group with which I had the fortune to cross paths: It was great sharing laboratory with all of you during the last five years.

Thanks for all your encouragement!

THESIS ORGANIZATION

1

"Everything that living things do can be understood in terms of the jiggling and wiggling of atoms." (Richard Feynman)

Life – what is biological life? Initially it was thought we had no chance of understanding life, that it contained something inexplicable and mysterious. We accepted this imposition, but set out to describe it in its most minute details and have now reached a point where we can define life without any need for a mysterious or miraculous force. Whether that is a good or a bad thing, and whether the lack of need implies a completely deterministic view of the world, now that is for philosophers to debate and probably they will never agree on one single answer. Having broken apart metabolism to its atomic structure does not mean that the whole is not bigger than the sum of its parts. Besides, the more we understand about the inner workings of biology, the more complex and fascinating it becomes. We are also ever so close to actually manipulating the genetic material at a large scale in different organisms and most controversially in human beings as well. Apart from the moral issues with such radical ideas that seem to be extracted from sci-fi literature (Brave New World), there is the question of whether we understand our own limits and are willing to acknowledge our blatant lack of foresight when it comes to the consequences. The fruit of knowledge is best served on a plate thoroughly dissected.

In 2003 we celebrated both the 50th anniversary of the discovery of the structure of the DNA double helix, and the announcement of the determination of the sequence of the human genome. Since then, we have come to understand the double helix to an unsurpassed level of detail, and in part this is due to the huge advances in computational simulation models. They have provided a vital tool with which to practically expose and look into the atomic underpinnings of molecular biology problems and envision their reasons and their implications. There is, of course, an intrinsic assumption I am working with throughout this thesis: I ask you to take the same leap of faith and trust in the ensemble results of computer simulations. I will make an attempt to convince you of their validity, accuracy and pertinence for usage in studies of DNA dynamics, but I will also not show reluctance to express my own doubts and point out some of the shortcomings that might arise from this mandatory assumption. It is, to begin with, a disheartening decision to give up the mathematical simplicity and elegance that went hand in hand with analytical sophistication for the apparently blunt, brute force of pumping computer power into the numerical simulation of the dynamics of a biological system/process composed of a huge number of particles. Miraculously, however, it gives extremely accurate results. It might not provide the philosophical answer of what to do with the knowledge we gain, but it sure does help along on out way to attain a deeper understanding of life, from the bottom up.

1 The History of DNA

The discovery of the DNA structure [1] was the catalyst for the development of a completely new field of science at the interface of chemistry and biology (Figure 1.1): Molecular Biology. After much speculation, DNA's role in heredity was confirmed in 1952 in the Hershey–Chase experiment [2]. Soon after that, Watson and Crick managed to construct the correct model of DNA double-helix based on X-ray diffraction images collected by Rosalind Franklin and Maurice Wilkins [3,4]. Their model not only satisfied all known facts, including previous experimental findings from Erwin Chargaff [5], who found 1:1 molar ratios of adenine:thymine and cytosine:guanine with DNA base composition varying between species, but the model also had profound biological implications.

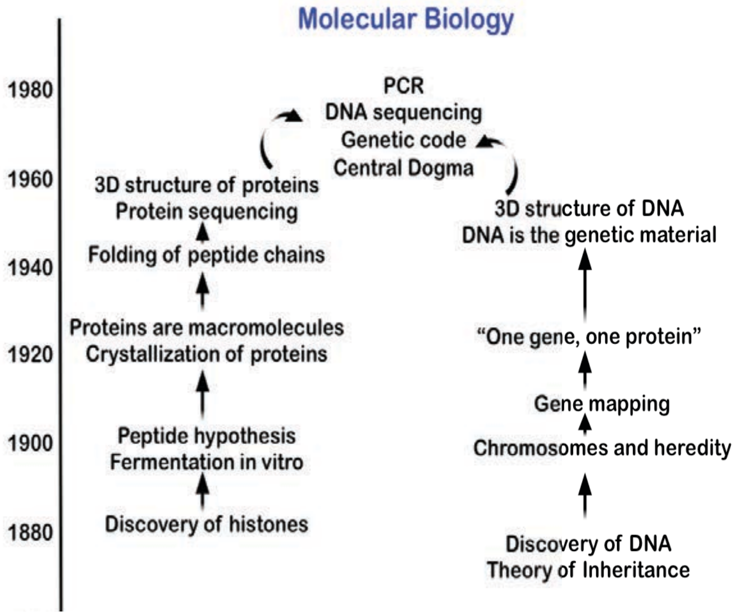


Figure 1 Scheme of the evolution of Molecular Biology

Locked into the versatile sequence of nucleotide bases in the DNA of a cell was all the information required to specify the diversity of biological molecules needed to carry out the functions of that cell. It did not take long for the double helical spiral of the DNA structure to become one of the most recognizable molecular shapes in the world and for the central dogma - DNA makes RNA makes protein - to be formulated. After that, advances had been made fast, if not effortlessly, with such marked discoveries as the DNA transcription into messenger RNA, RNA translation into protein sequences, DNA replication, responses to DNA damage, packaging of DNA into chromatin inside the cell, or regulating the expression of individual genes before reading/translating the encoded sequences, amongst other relevant processes.

Following Watson and Crick's great break-through, the structure of DNA has been intensively studied using a variety of methods including X-ray [6,7], NMR [8,9], and electron microscopy [10,11]. Additionally, computer-modeling simulations have shed even more details on the complexity of DNA structure and dynamics [12–14]. The "classical" and maybe even the "romantic" eras of nucleic acids study have long passed, but we are now in a time when incredibly powerful tools to manipulate life and to answer fundamental questions about chickens, eggs and what makes us tick are available for so many of us, giving us not only tantalizing promises of great discoveries, but also the burden of huge ethical issues related with the possibility to manipulate the essence of life.

2 Structure of DNA

Eukariotic DNA is a very long polymer divided into a few independent units named chromosomes, which in an organism like human can measure dm (if extended), but are highly packed into the µm- scale cellular nuclei. The long DNA fiber consists of two complementary biopolymer strands made of repeating units called nucleotides, coiled around each other to form a double helix. Each nucleotide is composed of one of four lipophilic nitrogenous bases (cytosine [C], guanine [G], adenine [A] or thymine [T]), a deoxyribose sugar, and a phosphate group fully ionized at the physiological pH. The nucleotides are held together in a chain by phosphodiester bonds, resulting in an alternating sugar-phosphate backbone. The phosphate groups are attached to the 5' carbon of one nucleotide and the 3′ carbon of the other, so that the full repeating unit in a nucleic acid is a 5′→3′-nucleotide. The nitrogenous bases are (mostly) planar aromatic rings that connect to the (deoxy)ribose ring by a glycosidic bond between the endocyclic base nitrogen and the sugar C1′ atom. The bases of two separate polynucleotide strands form hydrogen bonds to make double-stranded DNA, according to base pairing rules (A with T and C with G), and stack on top of each other to stabilize the double-helix (see Figure 1.2). I shall discuss each of these components and characteristics in detail below.

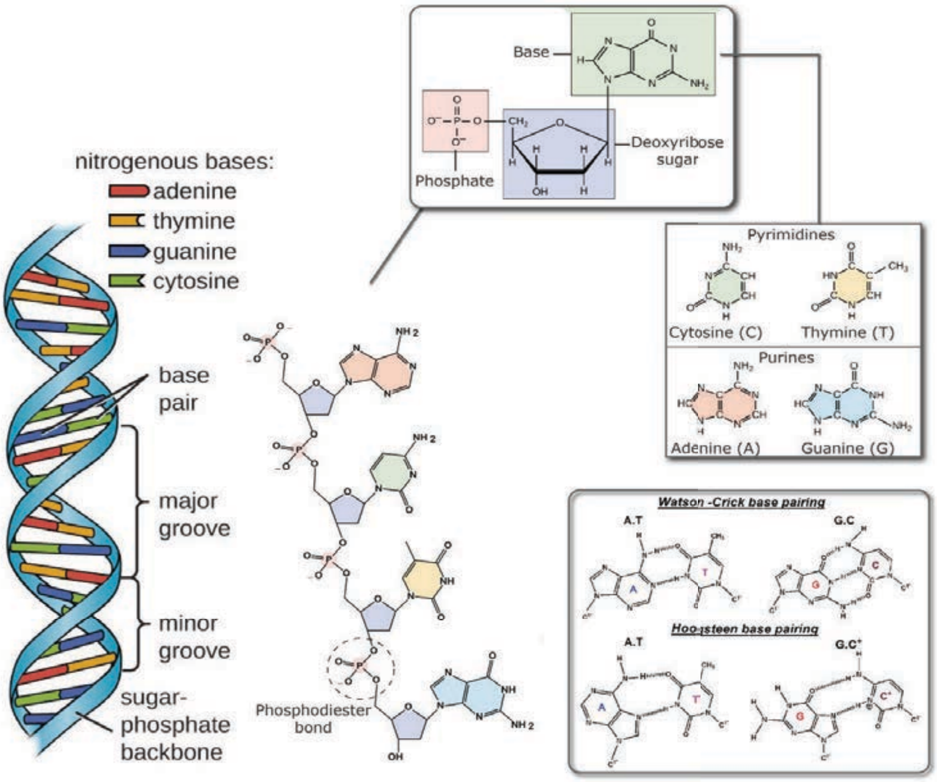


Figure 2 Structure of the DNA double helix, base composition and base pairing modes.

2.1 Central Base Pairing and the double helix

Since its discovery, DNA has been celebrated as a very elegant molecule, with a simplicity to its shape and composition that truly leaves the mind wandering how nature has written the entire language of life using just four chemical letters: guanine (G), cytosine (C), adenine (A) and thymine (T). Collectively denominated as "nucleobases", these chemical entities are planar aromatic heterocyclic molecules and are divided into two groups – the pyrimidine bases, C and T, and the purine bases, A and G.

Did our genetic code settle on this limited repertoire for a reason, or is it just lack of evolutionary imagination? As we understand better the precise molecular details of how DNA works, we uncover the high standard of prerequisites for a "good nucleotide" [15–17].

Isostericity of base pairing. The most obvious of these prerequisites is of course selective base pairing, by specific planar hydrogen bonds. In the standard base-pairing scheme, also called Watson-Crick pairing, adenine pairs with thymine and guanine pairs with cytosine. The A·T and G·C base pairs have different strengths (G·C base pair is around 1.5 kcal/mol more stable than A·T [18]), but they are isosteric and preserve the symmetry of the double helix and can stack in any combination without producing significant distortions in the structure. NMR experiments and theoretical studies [19–21] have shown that the double helix is able to accommodate for alternative base pairing, where purine bases could flip their normal conformations and form a new set of hydrogen bonds with their partners (see Figure 1.2).

Backbone accommodation of base pairing. Dispersion interactions and hydrophobic forces favor the stacking of the bases of DNA stabilizing regular polymers at physiological conditions. Stacking preferences and the physical properties of the sugar-phosphate backbone gives each step a slight twist (around 36°) with bases placed nearly parallel each other (inter-plane distance around 3.4 Å) and perpendicular to the helix axis.

Functional base derivatives. Chemical modifications of the nucleobases are possible, but happen always at the post-transcriptional level and are controlled by a complex network of regulatory enzymes. Modified nucleobases (especially cytosine derivatives) play a major role in DNA compaction and in the control of gene expression [22–25]. This fact further supports the idea that the chemistry of nucleobases has been fine-tuned by evolution to perfectly integrate with and respond to the different cellular metabolic processes.

2.2 Rigid base model: The Helical Parameters

The canonical model for the DNA structure implies a regular double helical arrangement of the two strands, each composed of a sequence of complementary nucleotide units (see Figure 1.2). However, in reality DNA is not a perfectly regular helix with identical steps. The single crystal x-ray analyses of B-DNA have revealed large variations in overall helix deformation [26,27] that have been partly assigned to sequence-dependent geometrical features. Such observations have raised the need for a consistent metric of nucleotide morphology that would help explain the sequence-dependent characteristics of the double helix. Particularly, descriptors of base and base pair configurations would be necessary, since these are the components that differ between the distinct nucleotides.

The elegant shape of the DNA allows the description of its defining structural elements in terms of a reduced set of helical internal coordinates. A number of rotational and translational parameters have been devised to describe the geometric relations between bases and base pairs (defined at EMBO meeting in Cambridge in 1988, also called "Cambridge Accord" and standardized at the Tsukuba Workshop on Nucleic Acid Structure and Interactions [28] so that a single reference frame would be used to calculate base morphology parameters and generate consistent values across studies). These parameters are defined either locally, with respect to a coordinate system attached to each individual base pair, or globally, with respect to a global curvilinear helical axis [29–31] (see Figure 1.3 for definition of the axes).

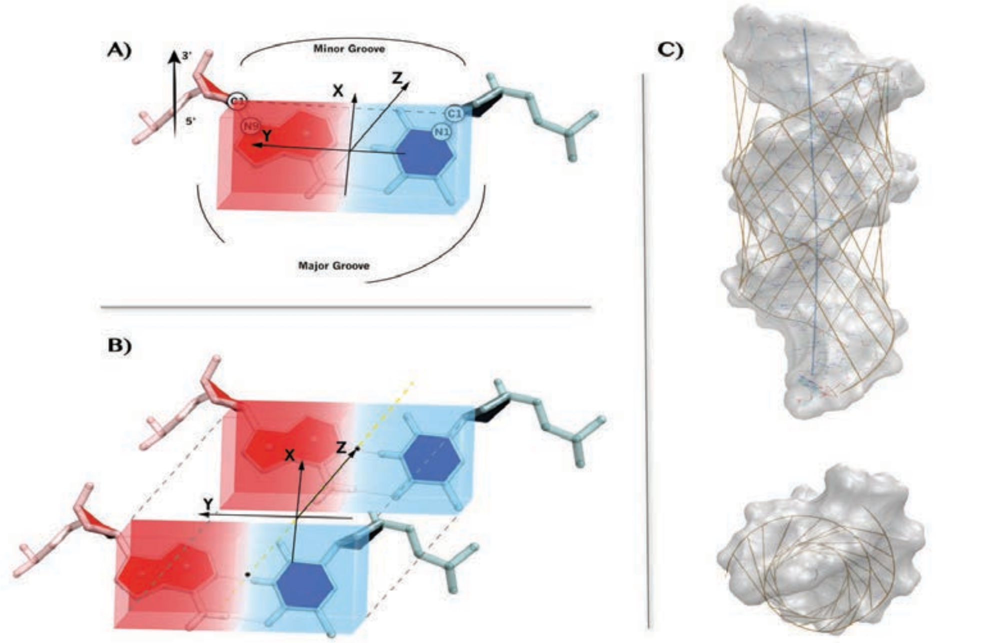


Figure 3 Definition of axes in helical space. A) base pair axes; B) base pair step axes; C) global helical axis

Accordingly, 3 translational and 3 rotational degrees of freedom can exhaustively define the relative geometry of an isolated pair of two **rigid-body bases**, together referred to as intra-base-pair parameters:

(a) **Buckle** is the relative torsion of base planes around their x-axis, **propeller twist** is the torsion between base planes around their y-axis, while **opening** is the torsion between bases around the helix, z-axis.

(b) **Shear**, **stretch** and **stagger** are relative displacements of bases along their x-axis, their y-axis and the helix axis, respectively.

The spatial orientation of a **base pair** modeled as a rigid body can be characterized with ten coordinates, six of which are relative to the previous base pair (3 rotational and 3 translational). They are defined in the dimer reference frame (see Figure 1.3-B) and are called inter-base-pair parameters:

(a) **Twist** is the angle between successive base pairs about the helix z-axis. More practically, it is measured as the change in orientation of the C1′-C1′ vectors on going from one base pair to the next. Similarly, **roll** is the dihedral angle for rotation of one base pair with respect to the other, about the y-axis of the base pair. Its positive value opens a base pair step towards the minor groove. **Tilt** is the corresponding dihedral angle along the x-axis of the base pair.

(b) **Rise** is the relative displacement of one base pair compared to another in the direction of the helix axis. Slide is the displacement of one base pair from its neighbor in the direction of the y-axis of the base pair, measured between the midpoints of each C6-C8 vector. Similarly, **shift** is defined as the relative displacement of a base pair from another in the direction of the base pair x-axis.

The remaining 4 coordinates describe an individual rigid base pair with respect to a local helical axis:

(a) **Inclination** is the angle between the y-axis of a base pair and a plane perpendicular to the helix axis and is defined as positive for a right-handed rotation about a vector from the helix axis towards the major groove. **Tip** is the angle between the x- axis of the base pair and a plane perpendicular to the helix axis and takes positive values for a right-handed rotation about the y-axis of the base pair.

(b) **X-displacement** and **y-displacement** define translations, along the x- or the y-axes, respectively, of a base pair mean plane from the helix axis. Positive X displacement is towards the major groove; positive Y displacement is towards the first nucleic acid strand of the duplex.

Together, the inter-base-pair parameters fully characterize the structure of the molecule as a ***stacked helix of rigid planar base pairs***. The six intra-base-pair degrees of freedom defined above must be additionally introduced only if the relative orientation of bases in each base pair is to be considered. Therefore, a dinucleotide unit (2 consecutive base pairs, or a base pair step) would be fully characterized by a set of 18 helical parameters in the reference coordinate systems attached to its constituent base pairs. (see Figure 1.4).

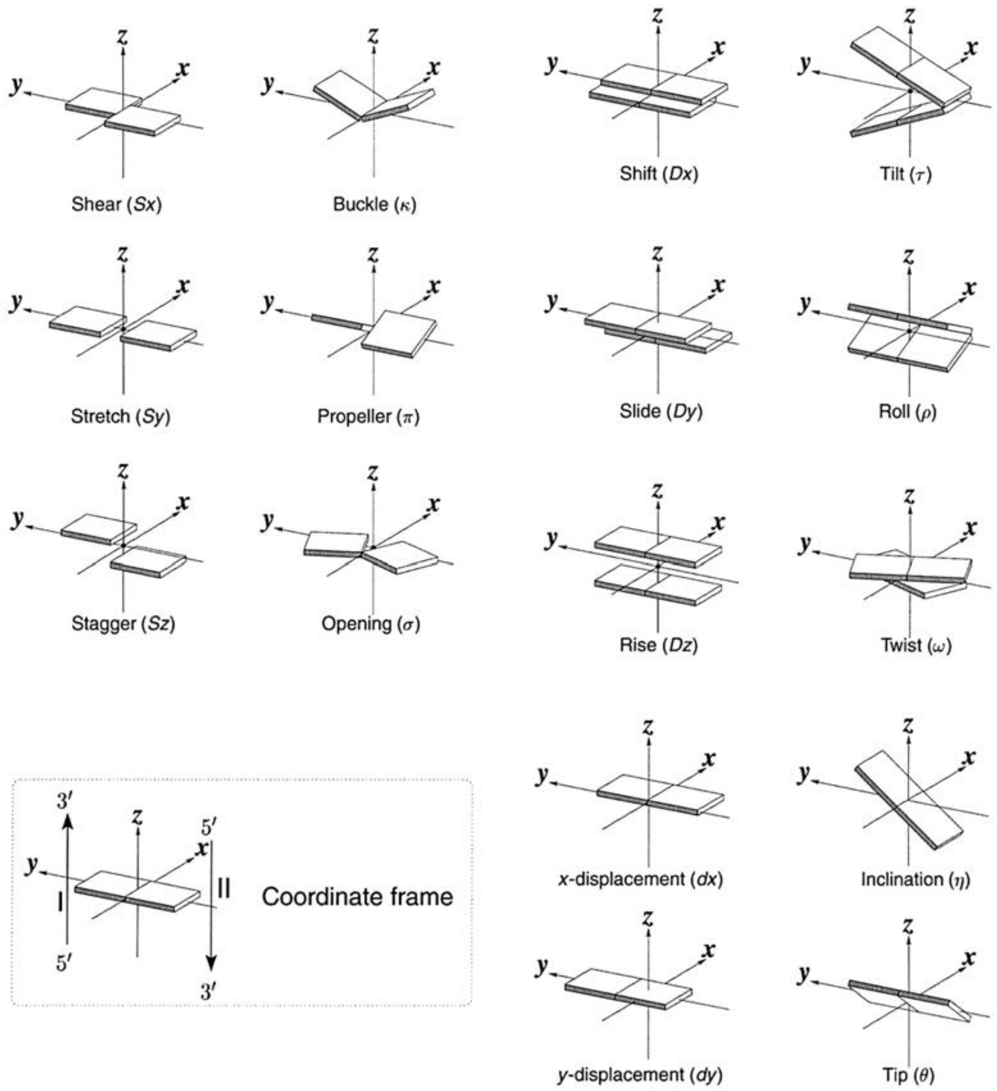


Figure 4 Rigid base and base pair model: helical parameters definition for base pair and base pair step. Taken from [32]

The helical parameters representation provides a good, complete and intuitive description of the helix at base resolution level, but completely ignores the backbone geometry, which plays a central role in defining, for example, the DNA recognition properties.

2.3 Backbone torsions.

The backbone of the DNA imposes a conformational landscape that can best be understood when considering its torsional degrees of freedom. The backbone geometry is thusly controlled by torsions around *six main chain, five sugar* and *one glycosidic* bonds (Figure 1.5).

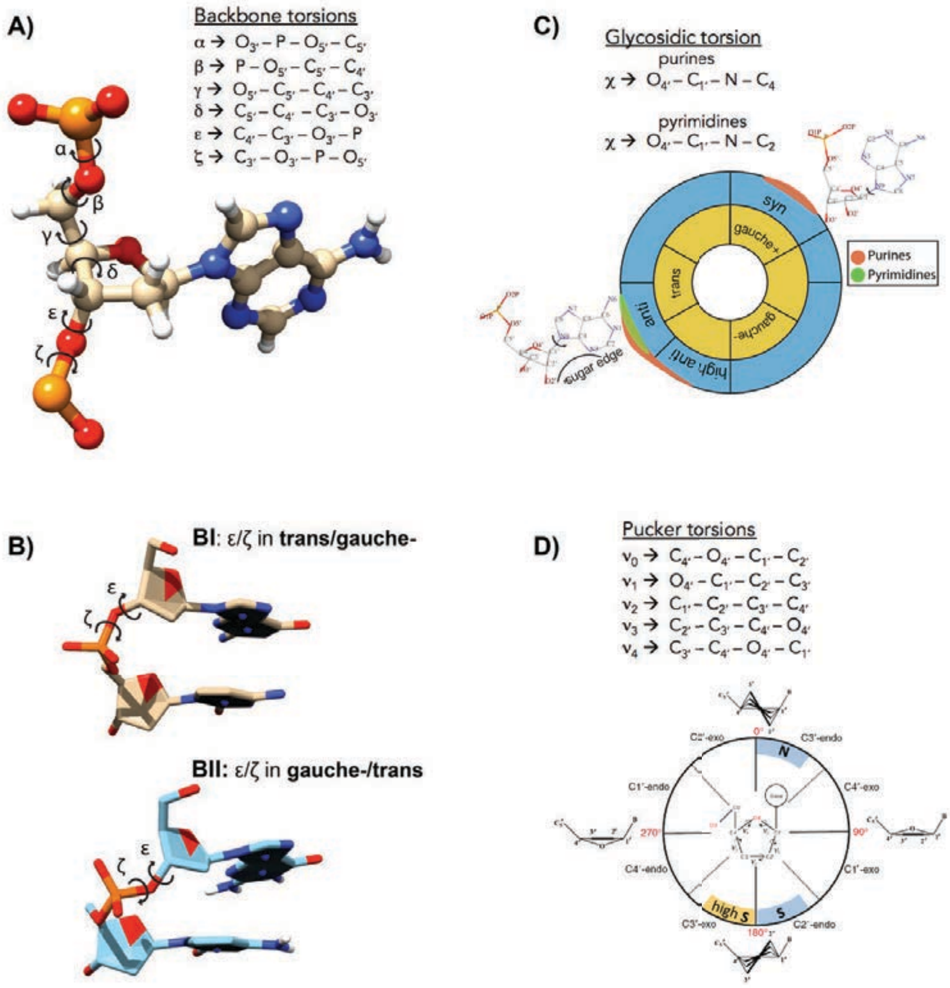

A)
Backbone torsions
α → $O_{3'} - P - O_{5'} - C_{5'}$
β → $P - O_{5'} - C_{5'} - C_{4'}$
γ → $O_{5'} - C_{5'} - C_{4'} - C_{3'}$
δ → $C_{5'} - C_{4'} - C_{3'} - O_{3'}$
ε → $C_{4'} - C_{3'} - O_{3'} - P$
ζ → $C_{3'} - O_{3'} - P - O_{5'}$
α
β
γ
δ
ε
ζ
C)
Glycosidic torsion
purines
χ → $O_{4'} - C_{1'} - N - C_4$
pyrimidines
χ → $O_{4'} - C_{1'} - N - C_2$
syn
gauche+
trans
gauche-
anti
high anti
sugar edge
Purines
Pyrimidines
B)
BI: ε/ζ in trans/gauche-
ε
ζ
BII: ε/ζ in gauche-/trans
ε
ζ
D)
Pucker torsions
ν_0 → $C_{4'} - O_{4'} - C_{1'} - C_{2'}$
ν_1 → $O_{4'} - C_{1'} - C_{2'} - C_{3'}$
ν_2 → $C_{1'} - C_{2'} - C_{3'} - C_{4'}$
ν_3 → $C_{2'} - C_{3'} - C_{4'} - O_{4'}$
ν_4 → $C_{3'} - C_{4'} - O_{4'} - C_{1'}$
0°
C2'-exo
C3'-endo
N
C1'-endo
C4'-exo
270°
90°
C4'-endo
C1'-exo
high S
S
C3'-exo
C2'-endo
180°

Figure 5 Definition of DNA backbone torsions. A) main chain torsions; B) BI/BII transitions in the main chain; C) Glycosidic torsion with ranges for pyrimidines (green) and purines (orange); D) Puckering types.

The *glycosidic torsion angle* represents the rotation around the bond joining the 1′-carbon of the deoxyribose sugar to the heterocyclic base and determines the orientation of the base with respect to the sugar-phosphate backbone. Rotation about this bond is restricted by steric and electrostatic factors. The precise conformation of a deoxyribose ring can be completely specified by the *five endocyclic torsion angles* within it, but the dependence between them imposed by covalent linkage allow pseudorotation models to compress this 5-torsions information into two parameters the *phase amplitude* (t_m) and *phase angle* (P). Under non-stress situations the *phase angle* provides all the required information on the sugar puckering [33,34]. The five sugar atoms can adopt a continuum of relative orientations, but here too there is a subset of dominant states that are observed, mainly due to nonbonded interactions between the four ring carbon atoms. The *main chain torsions* display highly correlated motions and tend to occupy one or at most two of three major discrete ranges. In addition, the restraints imposed by the Watson-Crick base pairing further reduce the number of possible low-energy conformations of the nucleotide unit.

The glycosidic torsion. Rotation of the base relative to the sugar is described by the torsion angle χ (O4′-C1′-N9-C4 in purines and O4′-C1′- N1-C2 in pyrimidines) around the glycosidic bond (in β-stereochemistry). This means that the base is always above the plane of the sugar in a transversal view, pointing towards the 5′ hydroxyl substituent and opposite the 3′ substituent. Two main ranges (anti and syn) and a minor one (high anti) are defined to describe the possible orientations adopted by the base, analogous to sugar moiety ranges. From sterical considerations alone, theory predicts the preferred domains of χ to be: *syn-* – between 30° and 90° and *anti-* – between 180° and 300°, with values around 270° ascribed to *high-anti*. The anti-conformation has the Watson-Crick hydrogen bonding groups directed away from the sugar ring, while in syn- conformations these groups are oriented towards the sugar. Pyrimidines occupy a narrow range of anti-conformations, whereas purines are found in a wider range of anti-conformations that can even extend into the high-anti range (Figure 1.5-C). The more compact syn- conformation is susceptible to steric clashes, which are rare in the extended anti- form. Although purine rings are generally larger, they have the smaller five-membered ring, as opposed to the six-membered ring of pyrimidines, attached directly to the sugar, so they will more readily adopt the compact syn- conformation than pyrimidines. The O5′ sugar atom has a marked contribution to this differential propensity, making the syn- orientation to be disfavored by pyrimidine bases, due to electrostatic repulsion towards the base oxygen O2 atom. On the contrary, purine bases have been found in a number of crystal structures to form a stabilizing hydrogen bond between the N3 base nitrogen atom and the O5′ of the sugar [35,36]. Nuclear magnetic resonance, CD and X-ray analyses all show that guanine prefers the syn- glycoside in mono-nucleotides and in some specific double helical structures [37]. Theoretical calculations suggest that this effect is due to favorable electrostatic interactions between the N2 amino group of guanine and the 5′ phosphate atom. For adenosine bases, the anti-conformation is still slightly preferred to the syn-. In double helical DNA structures, syn- conformations of the glycosidic torsion are almost never observed, since canonical Watson-Crick base pairing requires nucleotides to adopt an anti- orientation. There are cases however where an anti-to-syn transition can in fact promote alternative base-pairing modes (such as Hoogsteen base pairing). These conformations have mainly been observed as transient structures during certain DNA-repair mechanisms [38–41], but they also can play a significant role in some exotic DNA structures, such as quadruplexes or the triplex DNAs [42–45].

Sugar torsions. The five-membered furanose ring of the DNA is forced to deviate from planarity, leading to pucker conformations, which can be described by the five endocyclic torsion angles (ν0, ν1, ν2, ν3 and ν4), but can be qualitatively interpreted in terms of the atoms deviating from ring coplanarity. Several distinct deoxyribose ring pucker geometries have been observed experimentally [35]. The direction of atomic displacement from the plane is important. If the major displacement is on the same side as the base and C4′- C5′ bond, then the ring pucker involved is termed *endo*. If it is on the opposite side, it is called *exo*. If only one atom deviates from the ring plane the pucker is referred to as *envelope* and if there are equal displacements of either side, the pucker is termed *twist* conformations.

As noted above, he five-membered ring conformational can be described reduced to an elegant representation of two parameters: the pseudorotation phase angle P, and the puckering amplitude t_m [33,34]. The value of P, the phase angle of pseudorotation, indicates the type of pucker and is calculated as:

$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2 \cdot \nu_2 \cdot (\sin 36° + \sin 72°)} \tag{1.1}$$

The standard conformation for nucleic acids is by definition characterized by a maximally positive C1′–C2′–C3′–C4′ (ν_2) torsion angle, where P is equal to 0 by convention. The puckering amplitude t_m describes the maximum out-of-plane pucker and is given by:

$$\tau_m = \frac{\nu_2}{\cos P} \tag{1.2}$$

The C2′-endo family of puckers have P values in the range 140° to 185°. The C3′-endo domain has P values in the range –10° to +40°. Sugar pucker states are named after cardinal directions, where C3′-endo is called *North*, O4′-endo is called *East*, C2′-endo is called *South* and O4′-exo is called *West* (see Figure 1.5 D). High-level quantum mechanical two-dimensional energy scans in gas phase have shown energy profiles of the DNA nucleosides to contain two minima: a global minimum C2′-endo conformation and a C3′-endo local minimum, which is on average 3.00 ± 0.18 kcal/mol higher in energy [46]. The transition pathway for the DNA is usually through the East conformation as the barrier between the two (East conformers can be partially populated in some DNA structures), while West is the energetically highly unfavorable. South conformations have been prevalently found also in DNA structures, while North conformations are more common in RNA structures [47–50].

Sugar rings are the flexible link between the nucleic acid nucleobase and phosphate backbone, with different puckering modes influencing their relative orientation. On one side, there are pronounced correlations between sugar pucker and glycosidic angle, which reflect the changes in non-bonded clashes produced by C2′-endo versus C3′-endo puckers. Thus, syn- glycosidic angles are not found with C3′-endo puckers due to steric clashes between the base and the H3′ atom, which points toward the base in this pucker mode. On the other side, the sugar pucker is bound to influence or at least show strong correlations to the main chain torsions, which will be presented below. Each major sugar pucker, C2′-endo and C3′-endo, leads to very different relative positions of phosphate groups at the C5′ and O3′ ends of the sugar ring (Figure 1.5 C), with consequences for the overall architecture of the resulting helical conformation.

Main chain torsions. The phosphodiester backbone of a nucleotide has six variable torsion angles (see Figure 1.5 A), designated α, β, γ, δ, ε, ζ. These torsions each have precise ranges for the values they can adopt, on the basis of steric constraints, correlated motions due to energy minima distribution of nonbonded interactions, as well as restraints imposed by the Watson-Crick base-pairing in the double helix.

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°. The *β torsion* displays a rather wide but unimodal distribution in the trans region. Therefore, the orientation of the phosphate group to the furanose in the same nucleotide unit is controlled mainly by *angles α* (about the P–O5′ bond) and *γ* (the exocyclic angle about the C4′–C5′ bond), which are strongly correlated. Both torsions can adopt any of the g+, g−, or t conformations, but for example, in B-DNA double helix the canonical conformation of the *α*/*γ* torsional couple is g-/g+ [51–53]. In the B-form the α/γ rotamers may be flipped from their canonical g-/g+ values to the g+/g- values or some other rare conformation, which appear almost exclusively due to interactions with proteins [54,55].

The *torsion angle δ* around the C4′–C3′ bond adopts values that relate to the pucker of the sugar ring, since the internal ring torsion angle ν_3 is defined around the same bond. It has a narrow peak in its distribution around 135° (specific for B-DNA and correlated to C2′-endo) and a sparsely populated minimum at lower values of around 80° (observed on A-DNA and correlated with C3′-endo puckers) [53].

Another torsional couple is ε/ζ, torsions around oxygen O3′. Their concerted rotation gives rise to two major domains of backbone conformation and one of the best studied backbone transitions, namely the BI and BII states (see Figure 1.5 B). The two states are characterized by the ranges of ε and ξ or by the torsion difference (ε - ξ). The local changes in the two dihedrals are coupled with the motion of O3′ atom and the phosphate group from the following nucleotide unit (on the 3′ side). In the double helical B-DNA the canonical state is BI corresponding to a combination of ε and ξ torsions of 120°-210° (trans) and 235°-295° (gauche-), respectively. Transitions to the BII state push the phosphate group towards the minor groove, narrowing it and correspond to ε = 210°-300° (gauche-), and ξ = 150°-210° (trans) [56,57]. The angle difference (ε - ξ) is close to -90° for BI and +90° for BII phosphates. Moreover, the BI and BII states are suggested to be functionally relevant, and are in fact visible in some high-resolution crystal structures [58] and in 31P NMR experiments [59,60].

Overall, the local DNA structure is the result of the interplay between optimal base pair helical parameters, sugar conformations and preferred backbone dihedrals.

2.4 Helices.

The local nucleotide, base pair (bp) and base pair step (bps) degrees of freedom can be complemented with a number of parameters that define the helix as a whole [61,62]. For example, *helix sense* refers to the helical rotation of the double helix. *Residues per turn* refers to the number of base pairs in one helical turn of DNA, that is, the number of bases needed to complete one 360° rotation. The ideal structure described by Watson and Crick, "textbook B-form" DNA, contains 10 bp per turn. *Helix pitch* is the length of one complete helical turn of DNA. In textbook B-form DNA, one helical turn of 10 bp is completed in 34 Å. Diameter of the helix refers to the width across the helix. B-DNA has a diameter of 20 Å. DNA *curvature* is a measure on of the three dimensional bending of the helix and can be computed globally, per helical turn or more locally at a base pair level to describe acute kinks in the DNA structure. It can be described mathematically in terms of roll, tilt and twist helical parameters [63–65].

A dominant feature of a DNA helix is that the sugar groups attach to the same side of a base pair and define two types of indentations: a *major groove*, delineated by N7 of the purine and the C6 of the pyrimidine, and a *minor groove* with purine N3 and pyrimidine O2 (Figure 1.6). The nomenclature is made in reference to the more common B-DNA, where the empty volume provided by the major groove is larger than that of the minor, but it is kept for other DNA forms for consistency. Grooves are characterized by two parameters, groove width, defined as the perpendicular distance between phosphate groups on opposite strands with respect to the helix axis, and groove depth, defined as the difference in polar radii between phosphorus and N2 guanine or N6 adenine atoms, for minor and major grooves respectively. The grooves provide access to the nucleobase surfaces and can serve as a binding site for different molecules such as proteins, in case of major groove, or smaller ligands, in case of minor groove. Different functional groups on the purine and pyrimidine bases are accessible from the major or the minor groove. The major groove is richer in groups with H-bonding capabilities – O6, N6 of purines and N4, O4 of pyrimidines – than the minor one. This, together with the steric differences between the two, has important consequences for interaction with other molecules.

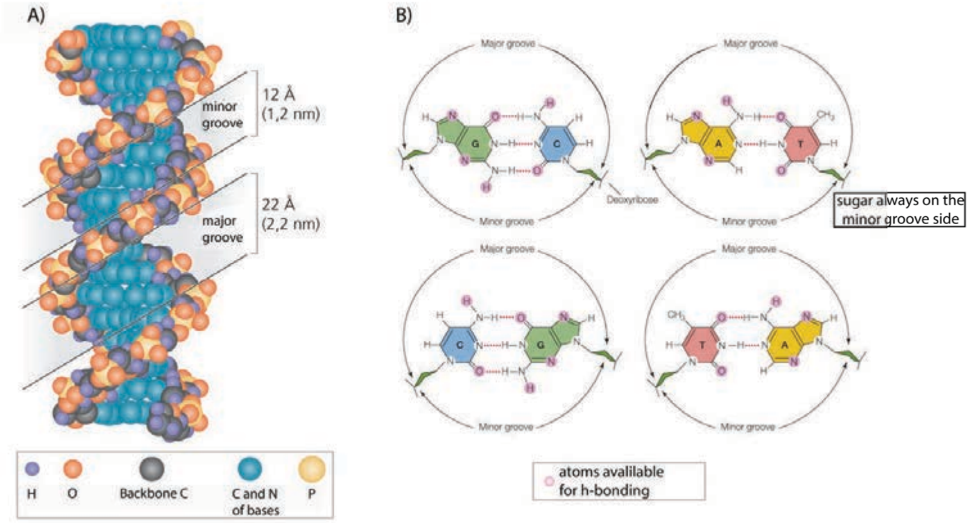


Figure 6 A) Definition of major and minor groove; B) Atoms exposed for intermolecular interactions in all base pairing schemes.

2.5 Major DNA structural families.

Three major families of DNA helices have been described under laboratory conditions: **B-DNA**, which is the most common and specific for mixed sequences (although the exact conformation varies between different sequence combinations); **A-DNA**, which can form within certain stretches of purines (e.g. GAGGGA), under non-physiological conditions (low humidity) or in DNA-RNA hybrids; and **Z-DNA**, which is favored by alternating pyrimidine–purine steps (e.g. CGCGCG) at high ionic strength (above 4M NaCl). The A-and B-DNA families are *right-handed* helices, while the Z-DNA family has a *left-handed* orientation of the helix [61,62] (see Figure 1.7).

The most ubiquitous form, present in most DNA at neutral pH, high relative humidity and physiological salt concentrations, is B-form. B-DNA is characterized by a right-handed helix formed by two anti-parallel polynucleotide chains that complete a full helical turn over 10 equal-width base-pairs. The B-form DNA has a solid central core, with deep grooves, of which the major groove is wide and the minor groove is narrow. It has the sugar pucker in the C2′ form and an anti- conformation of the glycosidic torsion. That is the classic, *right-handed* double helical structure we have been discussing up to this point.

A shorter and more compact right-handed duplex (the A-form) can form from B-DNA under dehydrating conditions, obtained for example by adding alcohol to the buffer. The A-form has also been described for RNA-DNA duplexes and is the major form of RNA-RNA duplexes [66,67]. The compaction of A-DNA means that each helical turn contains over 11 base pairs, with smaller distance between them. A-DNA has an axial hole at its center and the base pairs are inclined relative to the helical axis. It has a deep, narrow major groove and a wide, shallow minor groove, with a C3′ sugar pucker, but maintaining an anti- glycoside.

A more unfamiliar form of the double helix is the Z-DNA, which is a *left-handed* double helical nucleic acid conformation, in which purines are disposed in the syn conformation, resulting in a "zig-zag" arrangement of phosphate groups. The pyrimidines are in the anti- glycoside and thus in order to preserve base pairing, the helix in Z-DNA has to accommodate the distortion of its purines in the syn conformation. As noted above, Z-DNA conformation can be formed by sequences of alternating purines and pyrimidines at high salt concentration [68–70]. It has a more or less flat major groove and a deep narrow minor groove and has the sugar in the C3' endo conformation for purines and C2′ endo for pyrimidines. The main structural characteristics of these three most common families of DNA are summarized in Table 1.1 and their ideal conformations are depicted in Figure 1.8.

Table 1 Geometrical features of the 3 major DNA helix families (Neidle 2008)

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

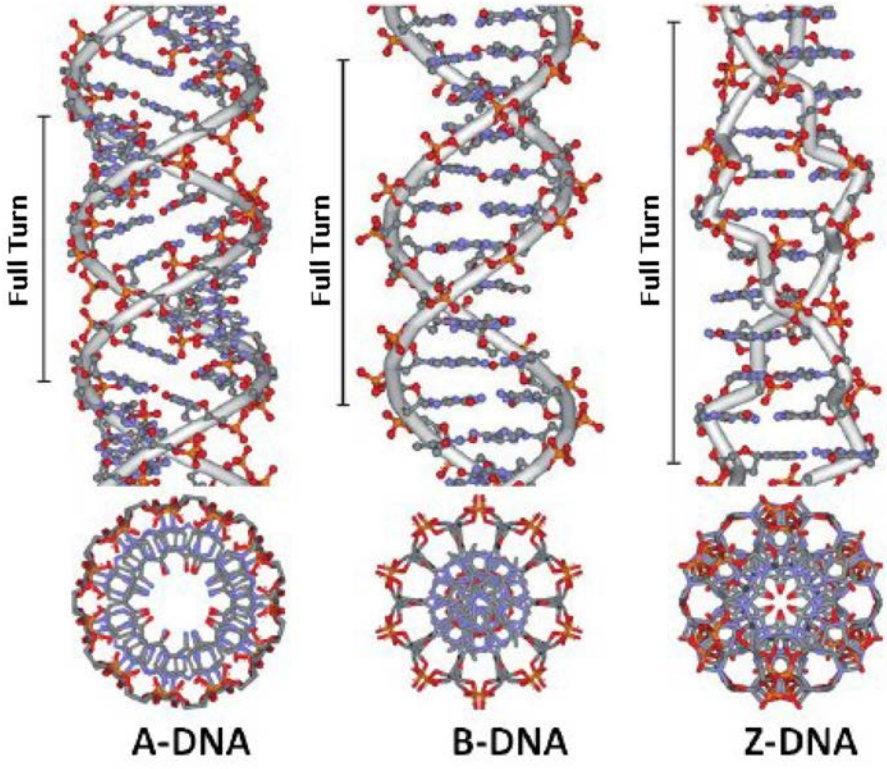


Figure 7 The three major forms of DNA double helix.

On larger scale, DNA, similarly to proteins, has a quaternary structure, referring to higher-level of organization of nucleic acids that define the chromatin. DNA compaction into chromatin is achieved through recruitment of ions and proteins, despite the DNA being one of the stiffest natural polymers. DNA interaction with the small proteins, called histones, leads to formation of nucleosomes, which are compacted further forming nucleosome clusters (perhaps superhelices), topological associated domains (TADs), globular territories and at the mitotic phase the packed chromosome [71] (see Figure 1.8). Many features of the DNA double helix contribute to its high stiffness, including the mechanical properties of the sugar-phosphate backbone, electrostatic repulsion between phosphates (DNA bears on average one elementary negative charge per each 0.17 nm of the double helix), stacking interactions between the bases of each individual strand, and inter-strand interactions.

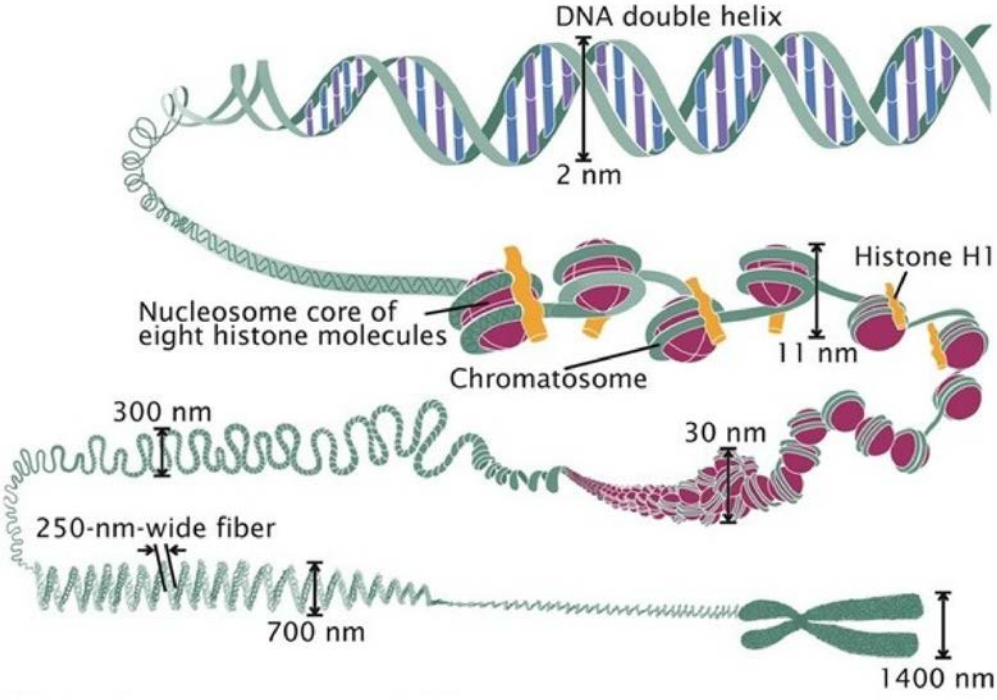


Figure 8 DNA compaction in eukariotic cells.

3 Short Incursions into the RNA world

This section attempts to give a simple account of what RNA is made of. It introduces the names and nomenclature commonly used to describe RNA molecules and provides a small sample of the complexity of this fascinating biopolymer.

3.1 DNA to RNA to Protein

DNA stores biological information, meaning that within its sequence it contains all necessary instructions for dictating the metabolic processes in the cell. However, in order for this information to be interpreted and put to its proper use, there is need for a new class of "adaptor" molecules, which have been hypothesized by Francis Crick before any shred of experimental evidence for their existence [72]. Once the existence of RNA had been demonstrated, Crick formulated *the central dogma of molecular biology* [73] (Figure 1.9), which describes the normal flow of biological information. Thus, information transfer is possible from DNA to DNA (*DNA replication*), from DNA to RNA (*transcription*), from RNA to DNA (*inverse transcription*), from RNA to RNA (*RNA replication*) and from RNA to protein (*translation*), but never from protein to nucleic acid.

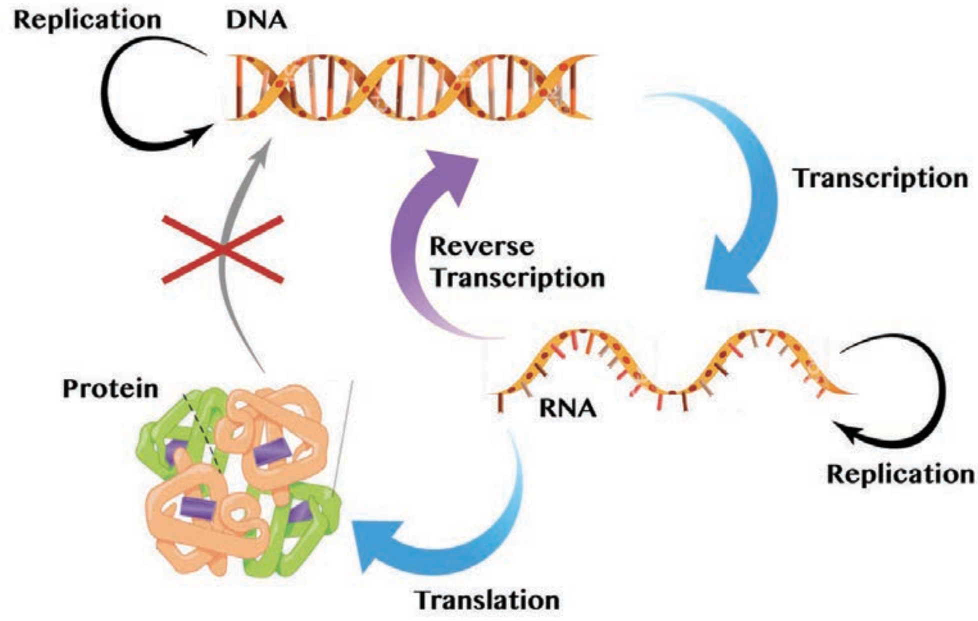


Figure 9 Central Dogma of Molecular Biology

The major implication of the central dogma is that genetic information is transferred from the DNA into protein, but not back out. Proteins do not change genetics, at least not directly, a conclusion that has held up remarkably well over the years. This guideline is not based on any physical law (in principle, all reactions involved in translation are reversible) but rather on a fundamental “biological law” that probably stems from the design of the translation system and is deeply rooted in the molecular setup of the information flow in all cells.

Even more interesting, the RNA, first hypothesized for the single specific task of transcription, is actually an incredibly versatile molecule, with a large number of types and functions [74]. The RNAs involved in protein synthesis include the abundant ribosomal RNA (rRNA), messenger RNA (mRNA), transfer RNA (tRNA) and signal recognition particle RNA (7SL RNA); however, there are many other types of RNAs with various roles, from enzymatic activity, to regulatory or post-transcriptional modification functions. The ability of RNA to both store genetic information and perform different regulatory and enzymatic functions has made very attractive the idea of the so-called RNA world, where life would have its origins in pre-cellular self-replicating RNAs. However, the difficulty of synthesizing RNA from abiotic molecules and other unanswered questions leave the dispute open. In any case, RNA is certainly a crucial part of biological life as we know it, and its bewildering variety and multitude of roles are only now starting to be uncovered and understood.

3.2 RNA fundamentals

Similar to the DNA, RNA is an oligopolymer with a phosphate backbone, its basic units (called a ribonucleotides) are also composed of a sugar and a nitrogenous base, but differ from the DNA nucleotides in two key chemical aspects.

Firstly, the ribose sugar groups of RNA have a hydroxyl group attached at the 2′ position, which is not present in the DNA deoxyribose sugars. The effect of the extra hydroxyl group is profound, making RNA sugars are much more rigid than DNA ones, and sterically forcing them into a C3′-endo pucker (Figure 1.10). The 2′ hydroxyl group also contributes to the capacity of the RNA to interact in versatile ways with itself and a variety of ligands, or to be readily hydrolyzed and cleaved.

The second difference is that the DNA thymine bases are replaced by uracils (also depicted in Figure 1.10) that employ the same base-pairing mode, but are missing the methyl group at the 5′ position. Since uracil is one product of hydroxylation of cytosine, its presence makes RNA more susceptible to mutations than DNA.

The overall double-edged consequence of this chemical change in the ribose is that RNA molecules are in the same time more versatile, being able to adopt a large variety of conformations and perform different functions, but also highly unstable [75,76].

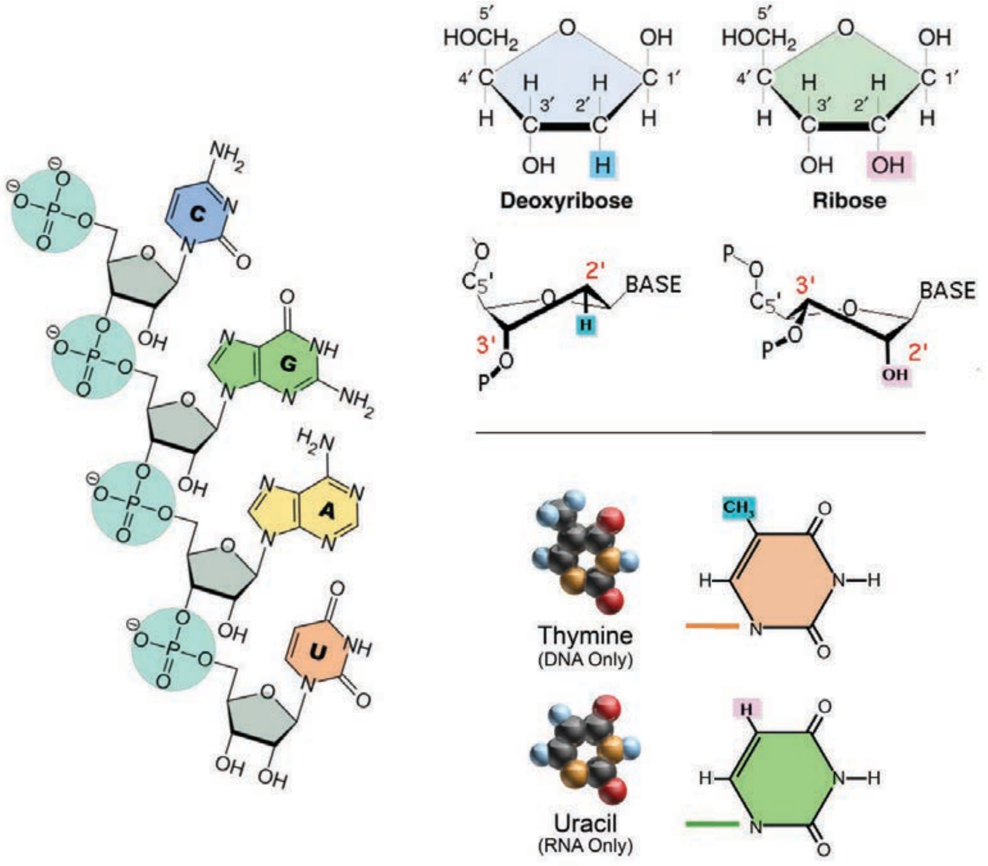


Figure 10 Left: Structure and base composition of an RNA strand; Right: Main differences between DNA and RNA molecular composition

3.3 RNA local structure

Typically, RNA is a single stranded polymer, but its structure is by no means linear, with the ribonucleotides being prone to interact between themselves in two principal ways, either through stacking on their flat faces or by hydrogen bonding across their edges. They can form isosteric base pairs that are analogous to the canonical DNA Watson-Crick ones, where A pairs with U and G with C; but there is one more isosteric pairing mode that is available to the RNA bases, which is the G·U wobble pair. The G·U pair has a similar stability to A·U [77,78] and is commonly observed in RNA structures. These three most common base pairing modes in RNA preserve therefore the overall dimensions of consecutive base pairs and also keep the ribose group of both bases on the same side of the pair, allowing for the definition of a major and minor groove. As a consequence, an arrangement of two or more of such pairs in a RNA sequence, either in single- or double-stranded RNA, will form a helix (or a stem, as it is sometimes referred to). The C3’-endo pucker of the ribose means the RNA helix will assemble into the A-form with 11 bp/turn, a deep narrow major groove, and a relatively shallow minor groove. At an intermediate level of analysis, denominated its secondary structure (SS), the fundamental structural element of sequence of RNA is the double helix. Once helices are specified, the unpaired regions between them can be classified into several types of structural elements, collectively termed loops. Loops can be internal, between two helix stems (a one-sided internal loop is called a bulge), or hairpin loops, consisting of several unpaired bases that are bounded on each side by the same helix, or multi-branched loops at the intersection of three or more helix stems (see Figure 1.11).

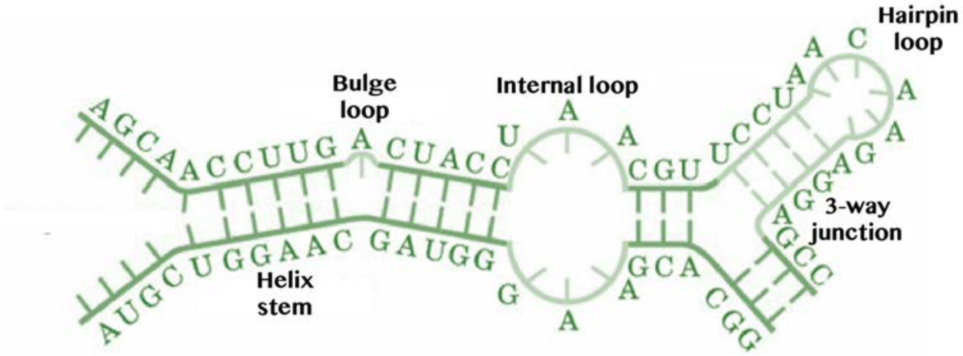

Hairpin loop
Bulge loop
Internal loop
3-way junction
Helix stem
AGCAACCUUG A CUACC U A A CGU UCCUA A C A A G A G G A
AUGCUGGAAC GAUGG G A A GCA CGG CC G

Figure 11 Representation of the most common secondary structure elements in RNA.

These nondescriptive secondary structure elements (SSE) can and do in fact come together and form tertiary interactions between separate regions of the SS. This is possible mainly because in RNA base interactions can occur regularly through the Hoogsteen edge or the sugar edge (see Figure 1.12). Taking into account that the glycosidic bonds can be oriented either in cis- or trans-, 12 principal geometric types are possible with at least two hydrogen bonds connecting the bases (Figure 1.12) [79]. This ability of the RNA bases to form hydrogen bonds in a multitude of combinations, sometimes involving more than two bases, is largely responsible for its formidable structural variability.

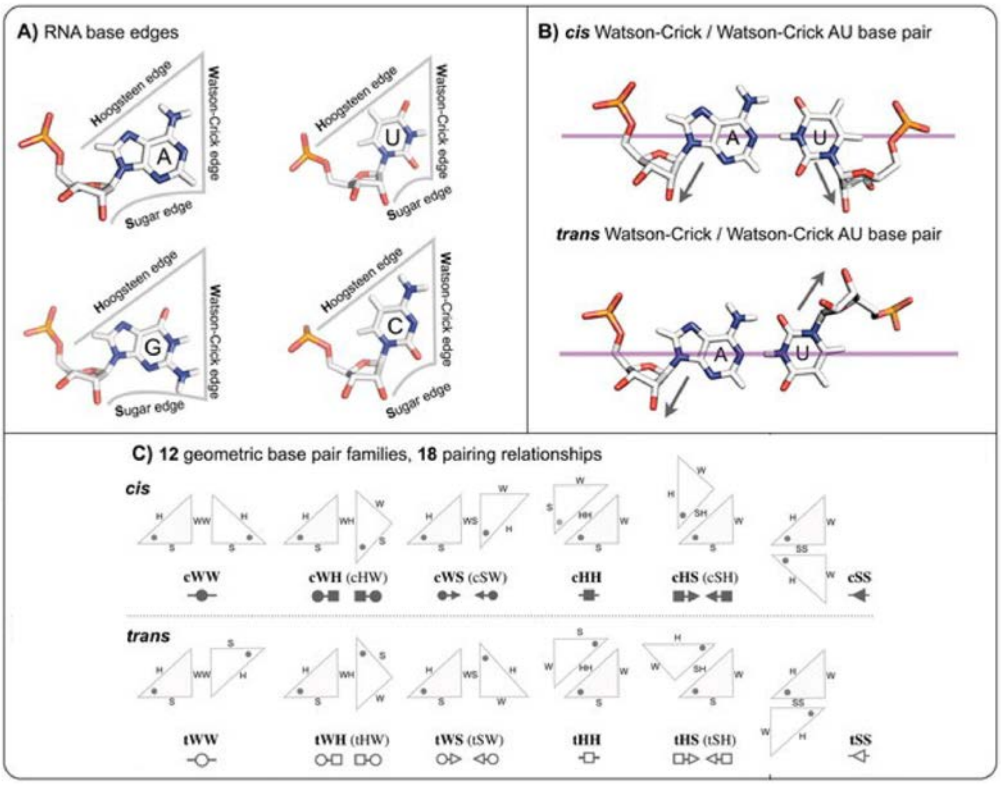

A) RNA base edges
Hoogsteen edge
Watson-Crick edge
Sugar edge
A
U
G
C
B) cis Watson-Crick / Watson-Crick AU base pair
trans Watson-Crick / Watson-Crick AU base pair
C) 12 geometric base pair families, 18 pairing relationships
cis
cWW
cWH (cHW)
cWS (cSW)
cHH
cHS (cSH)
cSS
trans
tWW
tWH (tHW)
tWS (tSW)
tHH
tHS (tSH)
tSS

Figure 12 Summary of Leontis/Westhof base pairing classification in RNA (taken from [79])

3.4 RNA motifs

Interestingly, even with access to a large number of possibilities in which the SSEs can interact with each other, RNAs were shown to be able to adopt complicated and yet precise structures [80–82], not unlike proteins. The commonly reoccurring types of small tertiary structure entities, which are frequently used in different combinations (like building blocks) to generate a rich variety of molecular shapes, are collectively names *RNA motifs*.

Some of the recurrent RNA motifs, formed through interactions between secondary structural motifs are depicted in Figure 1.13 [83,84]. For example, he K-turn introduces a tight kink into the helical axis, bringing together the minor groove sides of its two supporting helices. Another common RNA motif is the pseudoknot, where a single stranded region of a hairpin loop base pairs with an upstream or a downstream sequences within the same RNA strand. Technically, the pseudoknot is a SSE, but according to some definitions (and typically employed in SS prediction algorithms), SS does not include intercalated helices. The kissing loops motif forms when the single-stranded loop regions of two hairpins base pair with eachother. Finally, a very important small fold RNA motif is the three-way junction (the most ubiquitous type of multi-branched loop), a structural scaffold in which three helical stems, linked by at most three single-stranded segments, converge. This element can also be represented in the secondary structure, where it appears as open and unpaired, but in reality it is the tertiary structure of this motif, established by non-canonical base interactions, that determines its characteristic topology [85–87]. RNA junctions serve an essential role due to their ability to orient and bring together different segments of RNA. They also play crucial functional roles in a wide variety of biochemical processes.

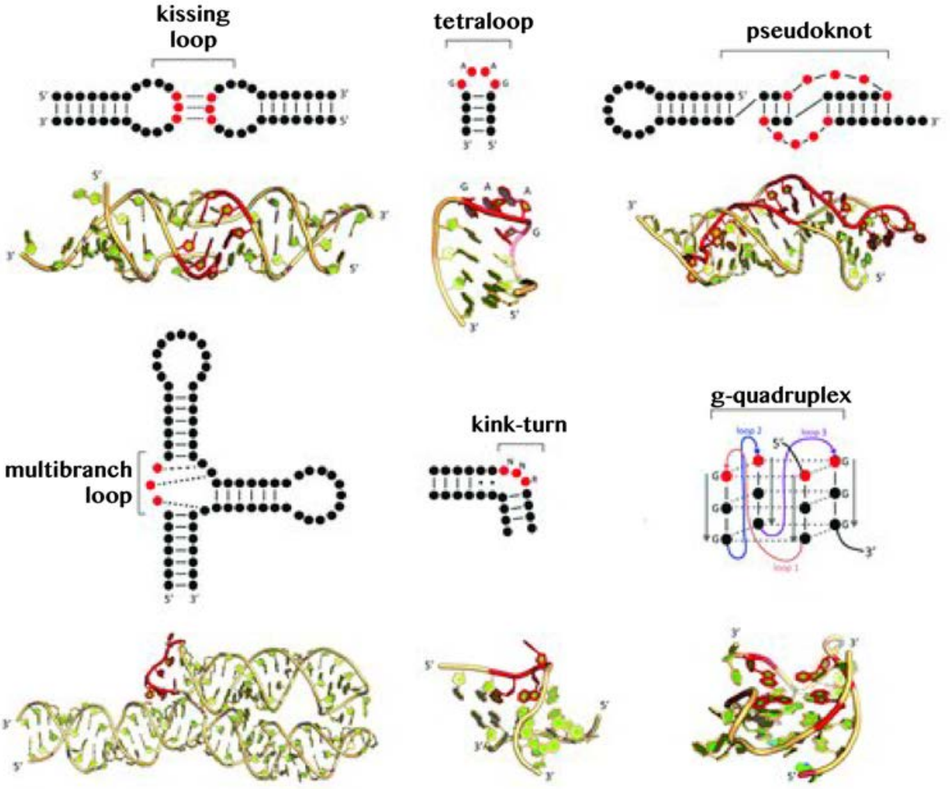


Figure 13 Main tertiary structure motifs in RNA.

3.5 RNA architecture

RNA is very similar to DNA in the chemical. However, the small differences that do exist have a very profound effect, giving the RNA a conformational richness similar to that of proteins. The size of RNA molecules can vary significantly, from a few tens of ribonucleotides as in the case of most tRNAs to several thousands in ribosomal RNAs and other types. However, RNA does not accomplish lengths similar to the chromatin DNA, because of its higher instability, but adopts specific (although flexible) 3-dimensional architectures, which are crucial to understand the moonlighting properties of RNA (carrier of information, catalyst or even controller of gene expression [76,88].

Even though direct experimental observation of global structural properties at atomic level is still not even close to offering a comprehensive view (and will most probably forever lag behind sequence determination), some general observations have been made over the years that can serve to bring some order to this complex fresco. First there was the realization that RNA structural features are much better conserved than sequence during evolution [89–91], similar to the protein case. Then a second principle developed that the complex three-dimensional architecture of an RNA molecule can be understood as hierarchically determined from stable SSEs that come together to form different RNA motifs through tertiary contacts, which in turn assemble in different combinations that impose the overall fold [92–95].

These two facts have implications that significantly ease the efforts of an exhaustive RNA structural characterization, by allowing the use of high-resolution 3D information about one molecule in analyzing the sequences of another molecule, whether or not the molecules are homologous.

4 Sequence-dependent landscape of physiological DNA.

The existing library of solved DNA crystal structures [96] reveals sequence-dependent irregularities in the conformational space of different double helices. Evidence of sequence-dependent variation of helical parameters within the B- family came with the original Dickerson–Drew structure [27], and since then there have been a large number of studies that have examined the underlying structural basis of local conformational heterogeneity. Taking the base pair step as the structural unit of a DNA sequence (the Nearest-Neighbour (NN) model), some significant observations have been made [97,98]:

- The local helical twist varies by up to 15° (the mean twist angle in the crystal structures is 36°,), with pyrimidine-purine (YpR) steps having lower than average values and purine-pyrimidine (RpY) steps having higher than average ones.
- Propeller twists are significantly greater for A•T base pairs than for G•C ones, by an average of 5–7° (The average value of Propeller in A- and B-DNA crystal structures is around –11º)
- Roll angles for YpR steps tend have positive values and open up toward the minor groove, whereas RpY steps have negative roll angles with major groove opening
- Backbone torsions also show considerable variability with respect to sequence (values spread over >45°)
- The RpY and AA•TT steps bend predominantly into the minor groove, whereas the YpR and GG•CC base steps bend more frequently toward the major groove

The sequence-dependent changes in helical and propeller twist, roll, and slide were initially rationalized on the basis of steric clashes between substituent atoms on individual bases – the Calladine Rules [99]. Essentially, Calladine proposes a model where the avoidance of steric clashes is accomplished through changes in base orientation, leading to changes in twist, roll, and slide:

- Minor-groove clashes are avoided by a *decrease* in local *twist*,
- The *roll* angle *increases* towards the groove that might have clashes,
- The *slide* of successive purines is *increased* to increase the spatial separation between them,
- A *decrease* in *propeller twist* will allow for higher base pair overlap.

Additional studies analysing increasingly larger databases of experimental structures [100] have shown that changes in slide result in alterations in the backbone angle δ, which thus has a higher value for purine compared to pyrimidine nucleosides. High propeller twist has been related to high twist as a way to reduce water exposure of the hydrophobic bases when they are forced to twist. Other correlations have been established between twist and roll, slide and roll, between slide, shift and backbone state and between propeller twist and slide [101].

As the amount of data increases more evidence exist that the helix parameters associated with a particular dinucleotide may vary depending on flanking base sequence. As a consequence, the nearest-neighbour model, which was the dominant one for decades, should be abandoned in favour of a tetramer level description (there are 136 unique 4-bp sequences) of structural DNA features (see Results section).

5 DNA dynamics and polymorphism: It's about TIME

Understanding the basic structural features of DNA and putting forth a set of rules that govern them has been a significant achievement since the discovery of the double helix over half a century ago. However, with simultaneous advances in experimental [7,9] and theoretical methods [12–14], as well as a growing understanding of molecular cell biology beyond the central dogma [102], it has become imperative to additionally take into account the dynamics and flexibility of DNA. The dynamic changes of DNA happen in a huge range of time scales, form the yearly time-scale (10^{8}–10^{10} s) of aging-related metamorphosis, to the daily time-scale of chromatin reorganization along cell cycle, onto the millisecond scale (10^{-3} s) of local nucleobase breathing, all the way to the sub-femtosecond time-scale ($<10^{-15}$ s) of electronic rearrangements. I shall focus here on up-to-the-millisecond time scale, where local structural anharmonicity can be captured, but also other significant events such as small ligand and protein binding, ion equilibrium and even large-scale conformational transitions.

5.1 DNA Conformational Transitions.

Ever since the fifties there had been clues about the highly heterogenic nature of DNA structure, as researchers realized that changes in the solvent composition could result in very different X-ray diffraction patterns, implying medium-dependent conformational transitions in DNA [3,4]. Thus, depending on the fluidity of the environment, the pressure, the temperature or the salinity, thermal fluctuations will allow the DNA to visit and populate several different combinations of substates, giving rise to the major conformations (A, B, Z, etc) and several sub-variants (A, A', C, D, T, BI/BII, ZI/ZII, etc). Furthermore, other oligomeric states of DNA, such as the triplex or quadruplex can be populated [42,67,70] (see Figure 1.14). However, it is only in the last decade that the availability of high resolution X-ray and accurate NMR data have revealed DNA polymorphism at the molecular level going well beyond the usual image of DNA architecture and strongly dependent on sequence contexts. Polymorphic behaviors at the *backbone level* [103,104], but also at the *base level* [105] were observed through both experimental and theoretical techniques. Thus, for B-DNA a specific combination of ε and ζ backbone torsions gives raise to a fundamental structural polymorphism at the junction of a given base pair step (bps) level called BI and BII substates. In general, under physiological conditions the BI substate is the thermodynamically preferred one, and was labeled by the community as the "canonical" substate present in B-DNA, while the existence of the BII conformation depends on the sequence context and perhaps on the distortion induced by an external molecule after binding to DNA. BI/BII transitions have been connected with changes in the grooves width and depth [103], and more recently to twist bimodality [106,107] (the dynamic equilibrium between two separate twist states), which has implications in Protein-DNA binding through the so-called indirect readout mechanism [108]. For free B-DNA structures in physiological conditions, the α/γ conformational landscape is dominated by the g-/g+ canonical conformation, with theoretical studies indicating that spontaneous flipping of these torsions from their canonical distribution is highly improbable [51]. However, in crystal of protein complexes with B-DNA unusual α/γ states are often encountered. Studies of the structural consequences of α/γ transitions show that the non-canonical backbone geometry has a significant impact on the roll and twist values and reduces the equilibrium dispersion of other structural parameters [109].

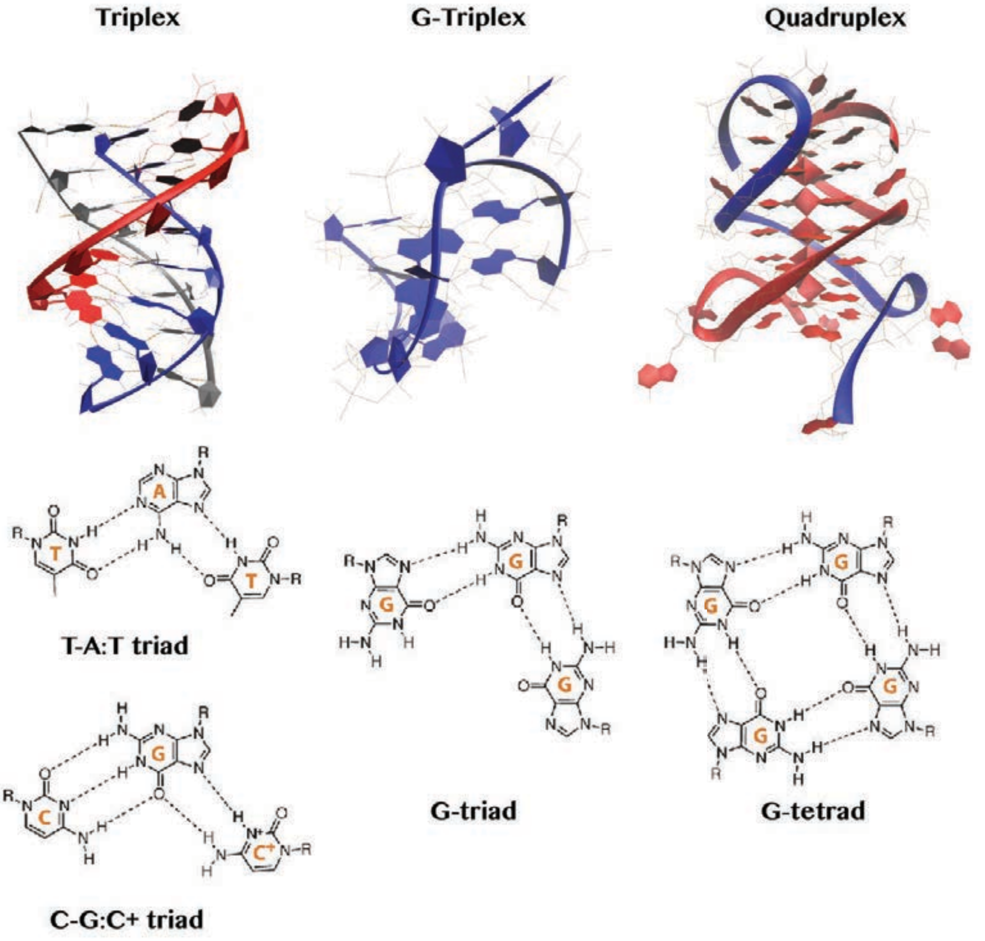


Figure 14 Illustration of Triplex, G-Triplex and Quadruplex DNA motifs with examples of base pairing modes indicated for each (taken from [79]**)**

5.2 Importance of the Solvent Environment.

The solvent hull around DNA has long been recognized as an integral part of the double helix with sequence-specific and conformation-specific interactions that are correlated with changes in the helix structure. The shift from the initial view of a sequence-independent delocalized cloud of counter ions surrounding the DNA [110] has been triggered by the pioneering work of [110] that gave insights from molecular dynamics (MD) simulations. Thus, MD simulations have been extensively employed over the years [31,111,112] in the study of ion coordination in the grooves or along the backbone of DNA. With the increase of high performance computer resources, storage capacity and the development of more efficient force fields, simulations achieved gradually increasing timescales and provided more complex understanding of salt-dependant DNA properties [113,114]. Sequence specificity of ion binding appears to go beyond base pair or base pair step levels and to contribute significantly to the heterogeneity of the DNA structure (see Figure 1.15).

Many of the structural features of the double helix are sensitive to sequence specific interactions with cations. Ion penetration in the minor groove was shown to modulate the groove width [115]. Dynamic correlations between bending events and ion proximity were found [111]. It was proposed that conformational transitions are driven by ionic distribution [101] and clear evidence of temporal correlation unraveled the key role of cation binding in triggering transitions of the backbone structure [31,106].

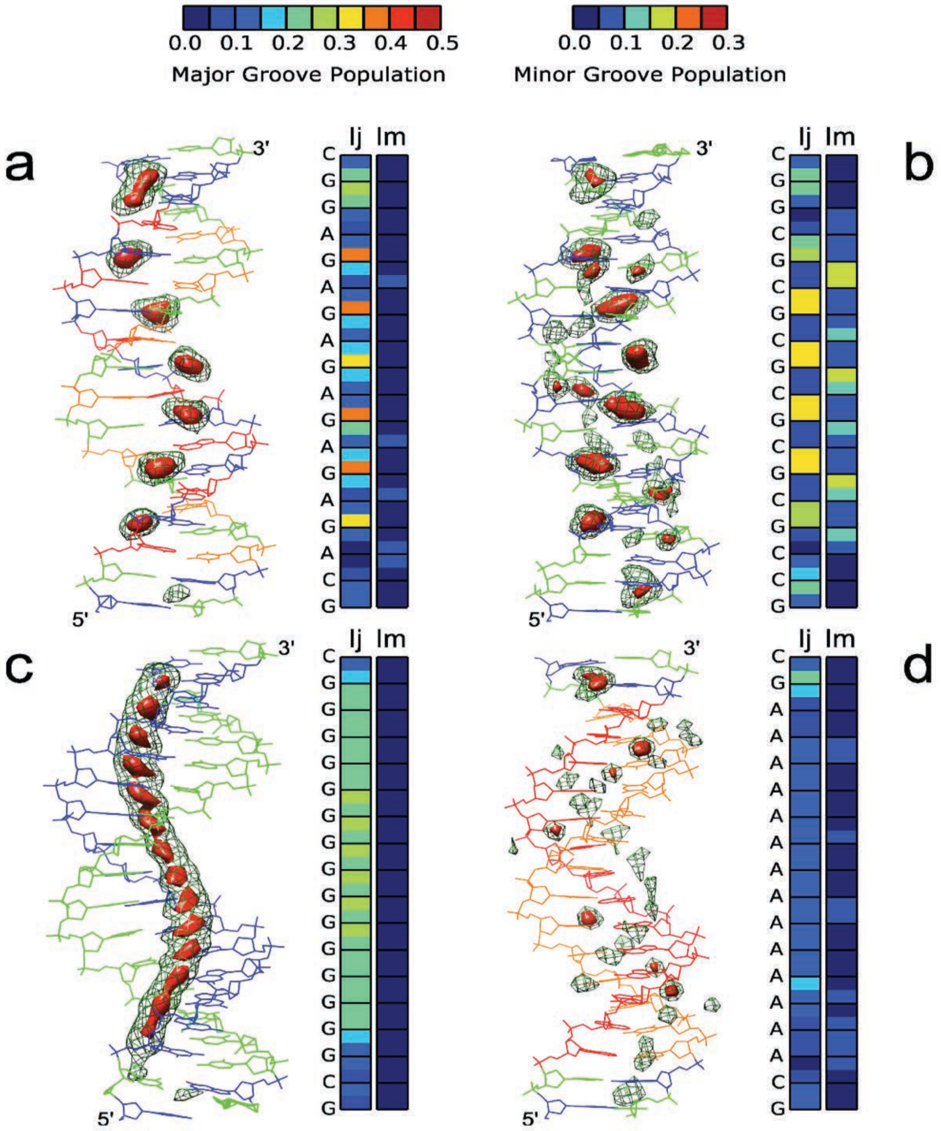

0.0 0.1 0.2 0.3 0.4 0.5
Major Groove Population
0.0 0.1 0.2 0.3
Minor Groove Population
a
b
c
d
3'
5'
Ij
Im
C G G A G A G A G A G A G A G A C G
C G G C G C G C G C G C G C G C C G
C G G G G G G G G G G G G G G G C G
C G A A A A A A A A A A A A A A C G

Figure 15 K+ distributions along the helix. Two representations of the K+ atmosphere around the AGAG (a), CGCG (b), GGGG (c) and AAAA (d) oligomers (taken from [31])

5.3 Flexibility Properties.

There is an increasing amount of experimental evidence accumulating that, on top of the one-dimensional sequence information and also the sequence dependent DNA structure, the deformability (or flexibility) of DNA represents an additional level of complexity to such a code (refs). Without question, DNA is a very flexible polymer and the differential propensity for deformation of different DNA sequences is a property that has a very significant impact on gene regulation mechanism – for example, flexibility changes between promoter and non-promoter sequences [116]. However, sequence-dependent, atomic resolution flexibility properties beyond the nearest neighbor model and the harmonic regime have yet to be documented in a comprehensive way.

Experimental methods – such as circularization experiments, atomic force microscopy, optical or magnetic tweezers, and permeation in nanopores – generally face a large number of limitations [117] in the determination of flexibility and can only obtain low resolutions information.

An alternative approach to evaluate flexibility is to generate an ensemble of structures and use the inverse of the covariance matrix (either in Cartesian or helical space) to calculated force constants associated with the elastic deformation modes [100]. Data inferred from ensembles of X-Ray or NMR determined structures, sorted by the different base pair steps of DNA, makes the assumption that variations of the helical properties in the crystals correspond to the amplitudes of thermal fluctuations in solution. It is hard to justify this assumption. Furthermore, it also relies on the assumption of the normal distribution of these properties and that known structures provide a dense enough sampling of the accessible conformational space. As discussed below, none of these two additional requirements is fulfilled. In this context, molecular dynamics simulation can become a source of flexibility parameters [118].

In particular, the efforts of the Ascona B-DNA (ABC) consortium have been very useful in providing information on the conformational properties of the 136 unique tetranucleotide sequences [119,120]. What emerges unquestionably from systematic database analysis and state-of-the-art molecular simulations is that sequence strongly influences the equilibrium conformation of DNA through a complex choreography of structural and energetic factors, involving often water and counterions (Figure 1.15). Average information obtained for the different tetramers can help to project the linear sequence of genetic information into a spatial code that governs the global organization of the double helix and unravels the functional implications of sequence-dependent conformational variability.

MD simulations have been able to address properly the sequence coverage problem, but with results showing that equilibrium distributions of helical parameters have severe deviations from normality, it is clear that the harmonic approximation implicit to elastic models is inaccurate. Several groups have already made significant progress in addressing these problems. For instance, Maddocks's group [121] has suggested the use of an alternative coordinate system consisting of rigid bases instead of rigid base pairs. Moreover, they have reported a new, elegant method to compute local stiffness parameters based on fitting the global (rather than local) flexibility.

6 Protein-DNA interactions.

The interaction of regulatory proteins with DNA is essential for the faithful completion of a large number of biological transactions ranging from gene expression regulation, to DNA replication, repair and packaging. For example, certain genes must be expressed at a precise time during development in a particular type of cell, or perhaps at one particular time during the cell cycle. Other genes must be expressed continually. Regulatory proteins must have the ability to bind a short unique base sequence out of the maybe billions of bases in the genome. Other proteins involved in major reorganization of DNA structure, such as histones or polymerases, must bind DNA in a sequence-independent fashion. The recognition of the DNA sequence is a very complex process, and it is not possible to articulate a simple code to define DNA sequence recognition [108], but there are some patterns that have been observed and some general conclusions that have been drawn based on both structure-determination experiments and MD studies.

Although still far from complete, the database of protein-DNA structures is growing considerably, and it is strongly backed-up by increasingly reliable theoretical models (As of March 2018, the NDB had 3434 protein-DNA structures, whereas only ~20 years before that, in 1997, there were just 241 structures). As a first division, the recognition of a DNA molecule by a small molecule or protein can be either highly specific, solely recognizing a defined sequence within a gene or even a genome, or nonspecific, without significant preferential binding to a particular nucleotide sequence.

6.1 Protein-DNA recognition mechanisms

Nonspecific protein-DNA binding has been shown to be widespread across genomes of different organisms [122]. The notion of nonspecific protein-DNA binding can be broadly described by two key, related mechanisms [123]: (i) the overall electrostatic attraction between protein binders and DNA, and (ii) the overall geometry of DNA . The first mechanism implies the binding of cationic residues of the protein to either the phosphates or the broad negative electrostatic potential inside the grooves, while the second implies that the overall helical arrangement of DNA generates a template for unspecific recognition of proteins displaying a complementary helical arrangement.

Proteins recognize **specific** DNA sequences by two strategies commonly referred to as a ''direct'' and ''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.

Direct read-out. Direct reading of a DNA sequence generally occurs via the hydrogen-bonding edges of the bases [124], either by small molecules that can insert themselves in the accessible volume or by proteins presenting features that enable them to form contacts through either major or minor grooves. The unique arrangement of hydrogen bond donor and acceptor sites for each dinucleotide within the grooves provide the specificity utilized by proteins to discriminate specific DNA sequences and are inherently directional. There are two hydrogen bond acceptors and one donor groups on the *major groove* surface of all four dinucleotide pairs (A·T, T·A, G·C, C·G; Figure 1.16). In addition, there is a methyl group at the C5 position of thymine that can participate in van der Waals interactions. These features of the B-DNA major groove make it both richer in potential contacts and in its ability to facilitate discrimination between different DNA sequences, which is essential for specific protein binding. Discrimination is achieved on the basis of differences in hydrogen bonding pattern between base pairs. C·G Watson-Crick base pair has a distinct pattern from that for the reversed G·C base pair. The A·T and T·A base pairs have identical major groove donor/acceptor patterns, but the presence of the methyl group on thymine introduces an asymmetry in the groove that can enable effective discrimination between these two base-pair sequences. Thus the major groove is generally the preferred site of direct information readout. On the contrary, patterns of hydrogen bonding in the minor grooves of each pair of sequences are symmetric (the N2 atom of G provides a hydrogen bond donor at the center of the minor groove for both C·G and G·C dinucleotide pairs), so discrimination on this basis alone by molecules entering the minor groove is not straightforward. Nonetheless, the minor groove is an important target for some regulatory and structural proteins, especially those that are able to deform DNA, expanding the minor groove.

Sequence selectivity can also occur at the dinucleotide level, by the so-called *bidentate* hydrogen-bonding pattern recognition [124] of two consecutive bases on the same strand. This feature is important, since it surpasses a one-to-one recognition code for protein-DNA direct readout interactions. Finally, it is worth to note that recognition is not restricted to the Watson-Crick base edges. There are several structures of complexes [20,40] where protein binding occurs with the transition of bases to a syn-conformation, and therefore a Hoogsteen base-pairing mode, exposing new functional groups to protein interaction.

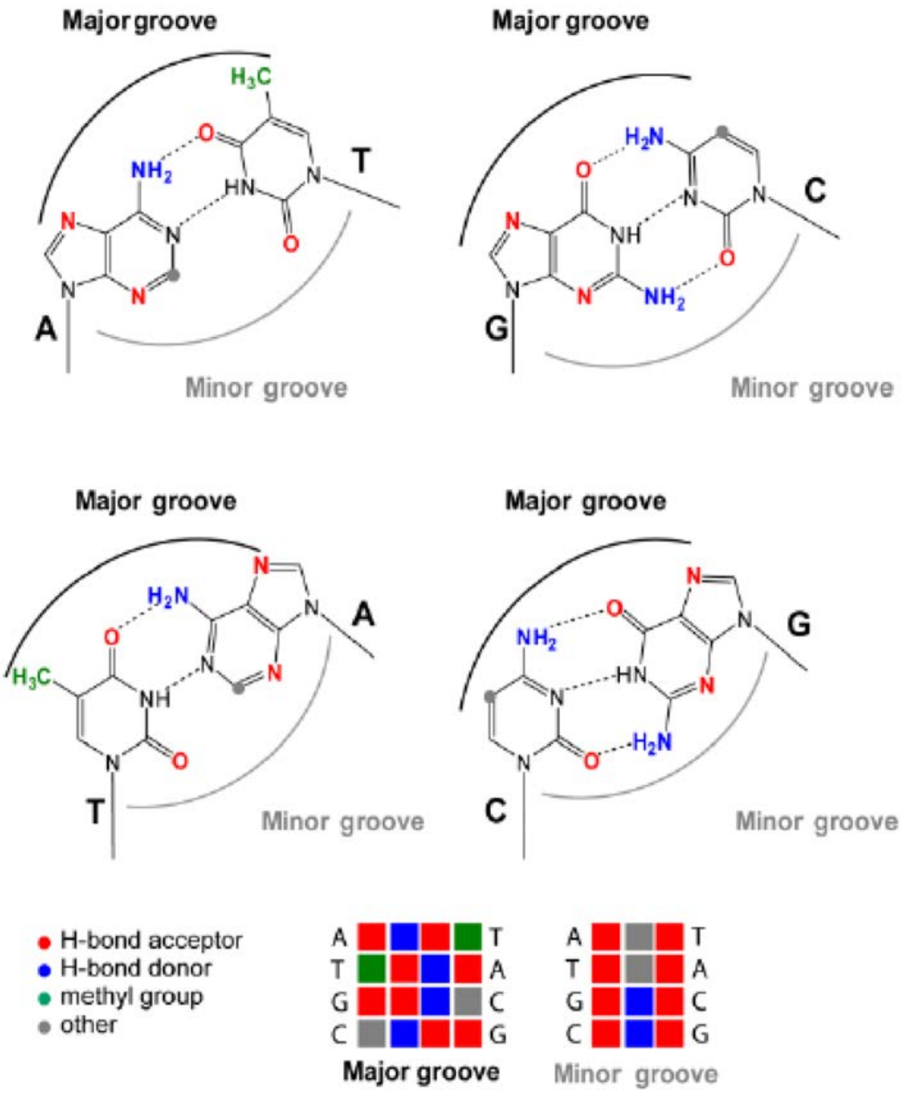


Figure 16 Base readout in the major and minor groove: Functional groups of the DNA base pairs in the major and minor DNA groove. . Hydrogen bond donors in blue, acceptors in red and thymine methyl group in green (taken from [125])

Indirect read-out. Indirect sequence readout by definition involves interaction with structural elements of DNA other than the base-pair hydrogen bonds, and is related to the ability of the DNA to adopt the "bioactive conformation" required to interact with the target protein. One major feature that proteins are able to recognize, and thus display sequence selectivity towards, is the size of the DNA grooves. So in this case accessibility is a consequence of groove-width variations, which are themselves sequence-dependent in both static and dynamic ways (see Figure 1.17-A). It has been established from the large number of oligonucleotide crystal structures that

A/T-rich minor grooves tend to be narrower than average; their widths depend both on the length and the nature of the A/T sequence (ApT versus TpA, for example). However, these trends are by no means absolute. A more realistic view is that A/T regions are more flexible than G/C ones, and that particular features such as a highly ordered spine of hydration and preferential ion binding, often result in narrowed minor grooves in A-tracts [108,114,115,124]. Several sequence-dependent structural features of the grooves are sensed by proteins through a negative indirect readout process. These include the shallow minor groove in a G/C rich region, caused by the bulky exocyclic amino group at the 2-position of guanine or the methyl group of thymine in the major groove, both of which destabilize protein binding by steric clashes. As a result, proteins will then bind more readily to other regions.

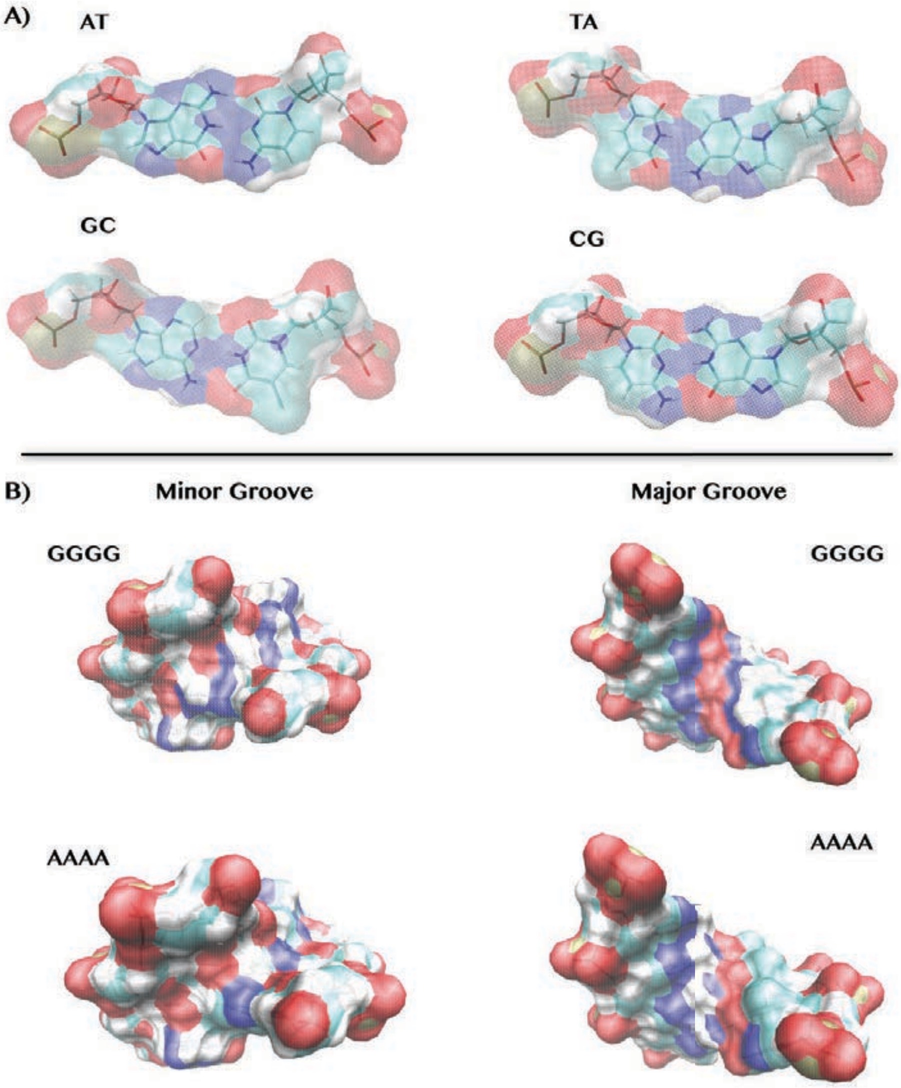


Figure 17 Electrostatic Potential surfaces for (A) each base pair type and (B) in the minor and major grooves of A- and G-tracts.

The backbone provides several mechanisms for indirect readout, involving very often a fine-tuning of the groove dimensions that alter general protein-DNA binding landscape [103,126]. There are also significant differences in electronic character between A·T and G·C base pairs, with the latter being more electron-rich (Figure 1.17-B). On one hand, this causes that the various dinucleotides have different stacking strengths (with a general tendency of less stacking between YpR steps than RpY ones [99,127]), which can modulate the accessibility and affinity of planar intercalator molecular groups. On the other hand, different sequences produce a characteristic electric field and potential in their vicinity [128], which impact mainly the groove generating a structure-dependent interaction fingerprint that is recognized by proteins (Figure 1.17).

Despite the existence of cases where binding can be unequivocally assigned to the direct or indirect paradigms, in the majority of cases both direct and indirect readouts work in a complementary way for specific protein binding [129]. Most DNA binding proteins are designed to recognize a particular shape or flexibility of the double helix in addition to a direct readout of individual bases in the recognition site.

6.2 Dynamic aspects of DNA-protein binding.

Dynamic aspects of protein-DNA interactions can be broadly separated in two types: the dynamics of the binding process itself and the dynamics of the complex after its formation.

The binding process. Experimental and theoretical evidences support the same model for the DNA recognition process, where the protein first binds non-specifically to DNA, and then it diffuses along the double helix, rapidly searching its sequence for the presence of binding sites [130]. The search process has been mainly ascribed to the formation of transient salt bridges between charged functional groups of amino acids and backbone atoms in the DNA. The protein typically first recognizes the local DNA shape and flexibility of its binding site, and only afterwards forms direct stable contacts to base or backbone atoms. Among factors governing conformational recognition in the first step, there have been reported differences in the propensity of the DNA structure to occupy BI or BII substates, specific groove widths, bending of the helical axis and other sequence-dependent helical parameters [23,100,103,108]. These effects can take place in a sequential way, with faster time scales for recognition of backbone states, followed by changes in helical parameters, then groove widening and bending.

Once it reaches its actual binding site, the protein has to be stabilized into the complex form with significantly higher affinity than at similar sequences. One hypothesis is that proteins preferentially bind to their recognition sites because of the specific deformability properties that make accessible the structure of the protein-bound form [131,132]. The concept of proteins taking advantage of the sequence-specific flexibility pattern of their DNA binding motifs is called *structural adaption*, which explains the large conformational strains that some proteins have been experimentally observed to impose on their DNA binding partners [100].

Complex Dynamics. After binding, it turns out that the protein-DNA complex is also not a collapsed configurational ensemble. As opposed to crystallographic studies, novel solution state NMR experiments can capture fast motions of flexible residues and several studies have shown that within specific complexes direct interactions are regularly broken and remade [133,134]. Several novel MD simulation studies on the microsecond timescale have confirmed this result, depicting a protein-DNA interface that undergoes significant dynamics [135–137]. Amino acid side chains often flip between hydrogen bonding partners, generating a number of different conformational substates that until now have probably been disregarded because of being averaged out in experimental data.

6.3 Energetics of protein-DNA binding

Separating the free energy into its enthalpic and entropic terms and comparing between a number of theoretical studies [138,139], it becomes clear that their relationship is compensatory in nature, and both proportion and sign can vary among complexes. Breaking down further the components of the free energy of protein binding, it might be useful to look separately at relative contributions of electrostatics (intramolecular and intermolecular), energetics of shape complementarity (packing) as reflected in van der Waals energies, solvent release, and reorganization on complex formation including the hydrophobic effect, deformation expense and internal entropies.

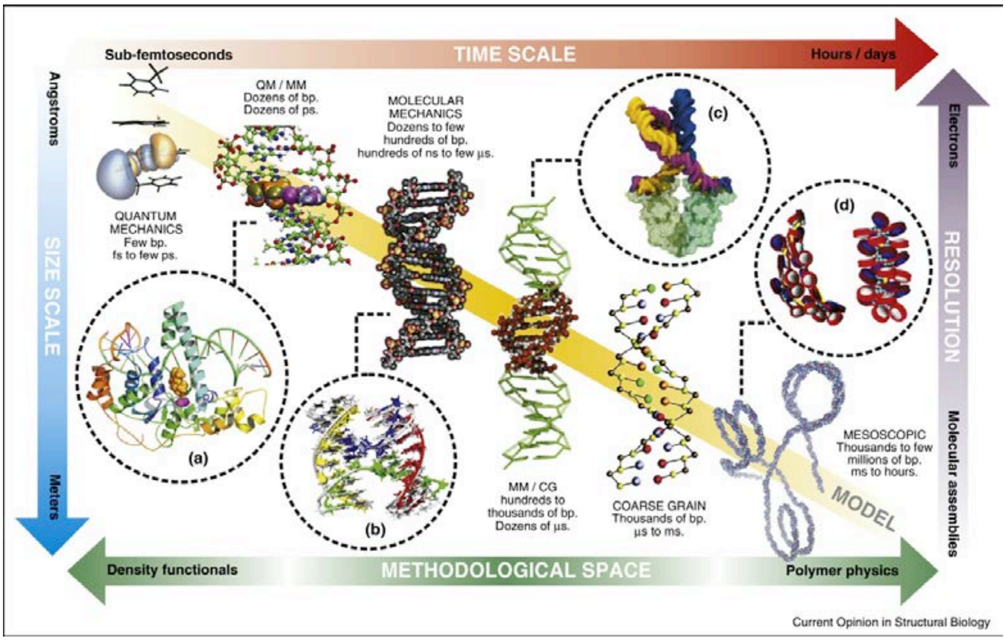
Electrostatics has a direct contribution due to the protein interacting with DNA, but an indirect contribution due to the relative effects of solvent interaction in the initial and final-state species. The net electrostatic contribution is therefore case-specific, as compensations between direct electrostatic interactions (typically favorable) and desolvation electrostatics (typically unfavorable) occur in all cases.

Similarly, van der Waals terms can be split into direct interactions between the protein and the DNA, which is always favorable, and the van der Waals component of desolvation, which is unfavorable. Ion effects are also double edged, with the decrease in interaction strength of the bound ions due to screening effects, being compensated at varying degrees by the entropic contributions of ion release. There is a general agreement that water release is found to favor binding. Entropy effects of protein and DNA deformation, resulting from the loss of translational, rotational, and vibrational internal degrees of freedom upon complex formation, are generally considered to be unfavorable to binding. Evidently, protein/DNA deformation incurs penalties to the enthalpic energy terms as well when distortions are large.

Protein–DNA interactions are system-specific with regards to the contribution of these competing forces. Moreover, their precise balance is key to calibrating specificity, even in systems where the total binding free energy remains in a narrow interval.

7 Theoretical models for the study of nucleic acids: A modeling hierarchy.

The study of DNA time-dependent properties covers a broad range of different scales, from sub-Angstrom details of the electronic distributions of nucleobases, to the mechanical properties of millimeter-long chromatin fibers. Its very nature makes it extremely challenging for any single theoretical framework to address DNA study in a holistic manner, from fine details of the electronic distributions at a given DNA step, to large chromatin rearrangements occurring throughout the cell cycle. The study of each of these properties is the focus of independent theoretical approaches, although whenever possible and/or needed, integrative hybrid methods can be employed (Figure 1.18). In a nutshell, the theoretical methods applied to the study of nucleic acids include: electronic or quantum mechanical (QM) models, atomistic molecular dynamics (MD), coarse-grained (CG) and mesoscopic models.


Sub-femtoseconds
TIME SCALE
Hours / days
Angstroms
SIZE SCALE
Meters
QUANTUM MECHANICS Few bp. fs to few ps.
QM / MM Dozens of bp. Dozens of ps.
MOLECULAR MECHANICS Dozens to few hundreds of bp. hundreds of ns to few µs.
(a)
(b)
(c)
(d)
MM / CG hundreds to thousands of bp. Dozens of µs.
COARSE GRAIN Thousands of bp. µs to ms.
MESOSCOPIC Thousands to few millions of bp. ms to hours.
MODEL
Electrons
RESOLUTION
Molecular assemblies
Density functionals
METHODOLOGICAL SPACE
Polymer physics

Current Opinion in Structural Biology

Figure 18 Techniques used to study DNA systems depicted with their representative system size, time scale and resolution (taken from [14]**)**

7.1 Electronic models.

QM simulations of small chemical systems provide the highest level of detail and accuracy of all theoretical frameworks and are habitually being used in the study of biological systems. For example, QM based methods provide structural information in excellent agreement with experiment, match experimental barrier heights for chemical reactions, and provide chemically accurate interaction energies for hydrogen-bonded or dispersive systems. However, their high computational cost limits their use to very small systems and introduces the necessity of using several approximations to simplify the complexity of calculations even for small (< 100 atoms) systems. Most commonly simplifications start with the Born-Oppenheimer approximation, disconnecting nuclei (treated as classical particles) and electron movements. The inter-correlation between electron movements can be represented in an average way, like Hartree-Fock approximation (HF), or in a more accurate way, such in the post-HF calculations like Moller-Plesset (commonly to the second order, MP2), Configuration Interaction (CI), Coupled Cluster (CC) or Complete Active Space Multiconfigurational SCF (CASSCF) method. More accurate than HF, but much less computationally demanding than the others, are the DFT (density functional theory) approaches, which are in fact the most widely used in biological applications [140–142].

QM methods are strictly necessary to study processes involving changes in electronic structure, including catalytic, photophysical or spectroscopic properties, but cannot be use when extensive sampling is requiredIn cases where QM description and extensive sampling is required one practical solution is to use combined quantum mechanical and efficient molecular mechanical methods (QM/MM). The idea was originally proposed by Warshel & Levitt [143] and later adopted by many other authors. It combines a QM description of the chemically active region with a MM representation for the surroundings, providing a perfect theoretical framework for systems where the region requiring QM level of theory can be precisely localized.

Both QM and QM/MM methods have already had a major impact on the study of biological systems, by providing invaluable information at the electronic level, which can further be used to parametrize force field representations of biological macromolecules [144,145], but are not efficient to study global properties of DNA in physiological environments, where electronic effects requiring QM description are rare and large sampling are required.

7.2 Classical atomistic models.

Classical MD simulations can be employed for problems where the electronic degrees of freedom can be ignored and the molecule can be represented by an additive energy term of atomic interactions, which integrated through Newton's laws of motion gives the time evolution of the system. Energy terms consist of harmonic potentials for bonded terms, and one or more expression (commonly Lennard–Jones and Coulomb potentials) to reproduce non-bonded interactions [146]. The different terms of the potential energy are dependent on empirical parameters, adjusted to reproduce experimental observables or high-level QM calculations. This severe simplification allows dramatic acceleration in the calculations which can be now done on systems containing even millions of atoms, and therefore increasing their similarity to real processes.

It took however a fair amount of time until classical approaches were successfully applied and then widely used in theoretical studies of DNA, compared to their use in protein studies for example, which has generated a gap between the protein- and the DNA- MD simulation worlds. Thus, the first atomistic-level simulations of a DNA duplex were reported in 1983 by Levitt's [147] and Karplus's [148] groups and involved short trajectories (less than 100 ps), while similar simulations for proteins appeared in 1976. Clearly, the heavily charged polymer of DNA is much more difficult to simulate than globular proteins.

Since the middle nineties, the development of new force-fields and especially the implementation of efficient methods to treat long-range electrostatic effects opened a new era where MD simulations were capable to explore reliable conformational states of nucleic acids in a time scale (10-100 ns) which approached the range of biological importance. Nowadays, routine simulations of DNA in physiological conditions go beyond the microsecond time scale [12,13] and it has been hypothesized that that motions in the 1μs to 1ms range are effectively absent [149], which implies that MD simulations can at the moment capture all biological processes of DNA up to the millisecond.

Evidently, apart from the issue of efficient software and advancing computer capabilities, the core issue of improving MD simulations accuracy is the force field. As efforts in software and hardware development allowed the extension of the size of the simulated models and the length of the trajectories, errors in the different generations of force fields have been gradually detected. Force fields are thus constantly either being refined to include more terms (such as polarization [150,151]) or corrected to better fit known quantities [152,153]. Overall, with an impressive tapestry of aspects of nucleic acid structure, dynamics and interactions to characterize and various approaches, atomistic simulations have had an important impact on our understanding of biological processes involving DNA. More details on MD simulations, as well as the state-of-the-art force field developments and their application for nucleic acids will be given in the following chapter.

7.3 Coarse-grained models.

It is often desirable to further reduce the complexity of the system in order to achieve even longer time scales and to model for example chromatin-level dynamics. *Coarse grain* (CG) is an ambiguous term used to label a family of models, which allow such a reduction in complexity by simplifying the representation of the model and/or the complexity of the potential energy functional. This can be done by representing chemical groups or even entire residues as single interacting centers (beads or grains), which decreases the number of pairwise interactions in the calculations of the potential energies or forces. Additionally, the high frequency vibrations are removed from the system, smoothing the potential energy surface, and allowing one to use a larger simulation time step. Therefore, the CG procedure should be appropriate to simulate the nucleic acid dynamical problems that occur on large timescales and/or in large size systems (thousands of base-pairs). For DNA, depending on the length scale of interest, quite different resolutions of CG modeling can be applied [14,154–156]. In models that aim to preserve base pairing interactions, the number of beads per nucleotide varies from 3 to as much as 8 beads (*particle-based* CG models), whereas other methods work in internal helical coordinates and use a rigid base representation, where the ground state and the stiffness matrixes are taken from MD simulations. When particle-based CG models are used, the beads are chosen in such a way as to reproduce the position and connectivity existing between the backbone, the sugar puckering and the base, as well as hydrogen bonds between bases. The force fields employed in these models can be analogous to those of MD simulations, or simplified by introducing statistical biases.

Regardless of the approach used to derive the force-field, one of the difficulties specific to the DNA coarse graining is the correct handling of long-range electrostatics, something that is crucial to correctly represent the densely packed electronegative charges of DNA [157]. Most models incorporate electrostatics implicitly, using Langevin dynamics for example and treat the solvent as a continuum, but some assign partial charges to the DNA beads and formalize their interaction with the ionic environment in a Debye–Hückel approach [157]. Moreover, some explicit models for water and ions have been developed to work specifically with CG models [158].

7.4 Mesoscopic models.

CG methods allow us to study long DNA segments, but are unable to cross the gap in time scales between DNA fiber dynamics and chromatin organization. Given the DNA's polymeric nature it is very attractive to study large scale DNA dynamics occurring in long DNA fragments through **mathematical models** that take advantage of the restraints imposed by the helical fiber.

The early mesoscopic models of DNA typically ignored the sequence-dependent structure of the double helix and fine structural details, such as bending anisotropy or the correlation between bending and twisting. The elastic rod model [159], based on Kirchhoff elastic rod model, represents DNA by its average macroscopic properties. The model has been used in many studies involving long sequences of DNA, such as DNA loops [160], supercoiled DNAs [161] or DNA mini-circles [162]. However, it only holds for sequences shorter than the persistence length (around 156 bp). For semi-flexible polymers of lengths several orders of magnitude higher than their persistence length, the worm-like chain model [163,164] has been developed and traditionally used to characterize the average elastic properties of long sequences of DNA [165].

More recently, with the development of highly accurate atomistic simulation methods as well as the advances made in a number of experimental techniques, the mesoscopic models have adopted one of two methodologies for their refinement: either bottom-up or top-down (although many times a mix of the two is the more reliable approach). Bottom-up models generally rely upon atomistic MD simulations on small duplexes from which DNA properties are extracted and applied to the study of entire chromosomes. In the top-down models the chromatin structure is derived by implementing experimental restraints coming from chromosome conformation capture techniques (such as Hi-C) into a simple model of the chromatin fiber. Each of the methods has its advantages and limitations and are used depending on the interest and bimolecular system. For better view on broad scale of computational methods, the reader is advised to look into [14,156,166].

7.5 Bioinformatics approaches

In the case or RNA, where experimental structure determination proves to be strenuous and time-consuming, not to mention the fact that high-resolution methods tend to bias the database towards similar and highly ordered structures [96], non *ab-initio* modeling methods cannot be employed before an accurate determination of a close-to-equilibrium starting structure, which can be extremely valuable for understanding the molecular mechanisms behind function [167,168].

The past decade has seen remarkable advances in the development of a new generation of RNA folding theories and models, but we are still far from a consensus scalable model. The prediction accuracy drops significantly for medium (50–100 nucleotides) to large (longer than 100 nucleotides) structures, even with the input from experimental data (101). After almost two decades of advances in computer models for RNA folding, we are now at the cusp of reliable predictions of large RNA 3D structures [169,170]. The biggest challenges to accurate prediction are structures with multi-branched loops, noncanonical interactions, and long-range tertiary contacts. There are several notable approaches to the *de-novo* RNA folding problem (see Figure 1.19).

Comparative or *homology modeling* is based on the empirical observation that evolutionarily related macromolecules usually retain similar 3D structure despite the divergence on the sequence level [171]. It has the advantage to be easily scaled up with no general size limit. However, the predicted model depends crucially on the template structure and on the quality of the sequence alignment, which mostly still requires expert manual preparation. The scarcity of the known and irredundant RNA structures imposes yet another challenge to homology modelling of certain classes of RNAs, which are not yet well represented in the PDB/NDB database. The inherent limitation of all homology modeling approaches is the fact that they are bound to a good alignment between sequences and a close 3D structure homolog – that is, preferably with >60% of sequence identity according with [172].

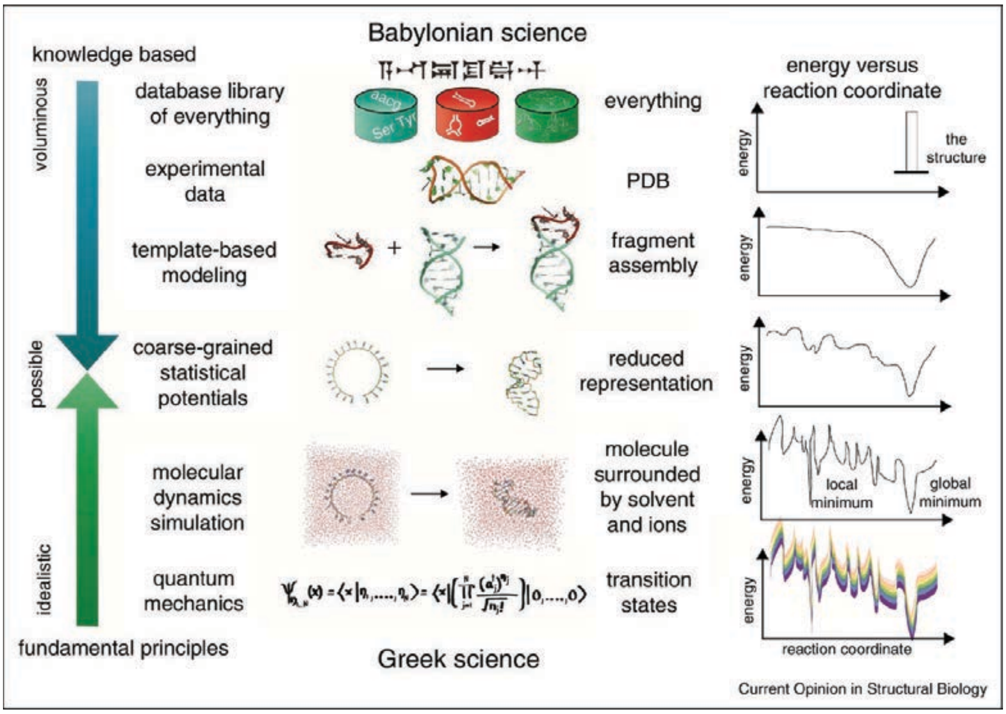


Figure 19 General description of the difference between science built on fundamental principles (Greek science) and knowledge based principles (Babylonian science), the nature of the details they employ, examples, and the energy landscapes (energy versus reaction coordinate) that they entail (taken from [170]**).**

To circumvent some of the shortcoming of global structure homology modelling, many groups took advantage of the strongly supported hypothesis of RNA hierarchical folding [93–95]. Although in most cases they (still) rely on alignment to templates from libraries of RNA motifs, the alignment is generally a structural one, based on secondary structure obtained either experimentally or predicted. Current approaches for secondary structure prediction [173] are already highly efficient and can thus be taken as a starting point for the 3D prediction. In addition, each module is aligned independently and the global architecture is obtained as a subsequent step with the assumption of high stability of these elements. The step of generating 3D RNA architectures from secondary structure elements has been approached in several different ways [174–176]. It is worth-emphasizing that the hierarchical folding approaches allow very easily the implementation of 'cheap' experimental data such as SHAPE, DMS and others (see review by [177]), which will be the real first step in obtaining comparable results with X-Ray, NMR or Cryo-EM techniques.

An altogether different school of thought has emerged separately and its motivation was to facilitate the unbiased simulation of the dynamics of an RNA molecule, which was rendered impossible by the discretization of the conformational space in fragment assembly methods. The tools developed for this purpose rely on very local structural features (from 1 to 3 nucleotides) to build a huge ensemble of possible conformations and then find the lowest energy conformation for a given RNA sequence. The energy function can have several forms and if designed properly can give very accurate results, but at the moment the high complexity and dimensionality of these models restricts them to rather small RNA structures. They still need to rely on knowledge-based terms for the potential, but these are made highly versatile by modeling the interdependency between the local conformations of only two or three adjacent nucleotides. Although these methods solve the uncertainty of accurate coverage of structural fragments in the solved RNA structure database, which currently contains only a limited number of non-redundant structures, their performance is still hindered by the need to sample a huge conformational space, even with the locally imposed restrictions.

Bibliography for Chapter I

1. WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 1953;**171**:737–8.

2. HERSHEY AD, CHASE M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* 1952;**36**:39–56.

3. FRANKLIN RE, GOSLING RG. Molecular Configuration in Sodium Thymonucleate. *Nature* 1953;**171**:740–1.

4. WILKINS MHF, STOKES AR, WILSON HR. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature* 1953;**171**:738–40.

5. CHARGAFF E, ZAMENHOF S, GREEN C. Human Desoxypentose Nucleic Acid: Composition of Human Desoxypentose Nucleic Acid. *Nature* 1950;**165**:756–7.

6. PERUTZ MF, ROSSMANN MG, CULLIS AF *et al.* Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* 1960;**185**:416–22.

7. Arnott S, Hukins DWL. Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. *J Mol Biol* 1973;**81**:93–105.

8. Palmer AG, Patel DJ. Kurt Wüthrich and NMR of biological macromolecules. *Structure* 2002;**10**:1603–4.

9. Ts'o POP, Kan L-S. Nuclear Magnetic Resonance Studies of Nucleic Acids and Proteins. *Chromatin Structure and Function*. Boston, MA: Springer US, 1979, 217–49.

10. Beer M, Stern S, Carmalt D *et al.* Determination of Base Sequence in Nucleic Acids with the Electron Microscope. V. The Thymine-Specific Reactions of Osmium Tetroxide with Deoxyribonucleic Acid and Its Components [*]. *Biochemistry* 1966;**5**:2283–8.

11. Douglas SM, Dietz H, Liedl T *et al.* Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* 2009;**459**:414–8.

12. Pérez A, Luque FJ, Orozco M. Frontiers in Molecular Dynamics Simulations of DNA. *Acc Chem Res* 2012;**45**:196–205.

13. Cheatham TE, Case DA. Twenty-five years of nucleic acid simulations. *Biopolymers* 2013;**99**:n/a-n/a.

14. Dans PD, Walther J, Gómez H *et al.* Multiscale simulation of DNA. *Curr Opin Struct Biol* 2016;**37**:29–45.

15. Georgiadis MM, Singh I, Kellett WF *et al.* Structural Basis for a Six

Nucleotide Genetic Alphabet. *J Am Chem Soc* 2015;**137**:6947–55.

16. Zhang Y, Ptacin JL, Fischer EC *et al.* A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* 2017;**551**:644–7.

17. Johnson R. Xeno-nucleic acids: Unnatural biocatalysts. *Nat Chem 2015 72* 2015.

18. E. Stofer, C. Chipot † and, Lavery* R. Free Energy Calculations of Watson−Crick Base Pairing in Aqueous Solution. 1999, DOI: 10.1021/JA991092Z.

19. Nikolova EN, Gottardo FL, Al-Hashimi HM. Probing Transient Hoogsteen Hydrogen Bonds in Canonical Duplex DNA Using NMR Relaxation Dispersion and Single-Atom Substitution. *J Am Chem Soc* 2012;**134**:3667–70.

20. Nikolova EN, Kim E, Wise AA *et al.* Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* 2011;**470**:498–502.

21. Zhou H, Hintze BJ, Kimsey IJ *et al.* New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. *Nucleic Acids Res* 2015;**43**:3420–33.

22. Schübeler D. Function and information content of DNA methylation. *Nature* 2015;**517**:321–6.

23. Hashimoto H, Liu Y, Upadhyay AK *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* 2012;**40**:4841–9.

24. Kriaucionis S, Heintz N. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science (80-)* 2009;**324**:929–30.

25. Keshet I, Lieman-Hurwitz J, Cedar H. DNA methylation affects the formation of active chromatin. *Cell* 1986;**44**:535–43.

26. Dickerson RE, Drew HR. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. *J Mol Biol* 1981;**149**:761–86.

27. Drew HR, Wing RM, Takano T *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *Proc Natl Acad Sci U S A* 1981;**78**:2179–83.

28. Olson WK, Bansal M, Burley SK *et al.* 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* 2001;**313**:229–37.

29. Blanchet C, Pasi M, Zakrzewska K *et al.* CURVES+ web server for

analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* 2011;**39**:W68-73.

30. Lu X-J, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* 2003;**31**:5108–21.

31. Pasi M, Maddocks JH, Lavery R. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res* 2015;**43**:2412–23.

32. Lu X-J, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* 2003;**31**:5108–21.

33. Westhof E, Sundaralingam M. 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* 1983;**105**:970–6.

34. Altona C, Sundaralingam M. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *J Am Chem Soc* 1972;**94**:8205–12.

35. Olson WK, Sussman JL. How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. *J Am Chem Soc* 1982;**104**:270–8.

36. Schneider B, Neidle S, Berman HM. Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers* 1997;**42**:113–24.

37. Blackburn GM, Gait MJ, Loakes D *et al.* eds. *Nucleic Acids in Chemistry and Biology*. Cambridge: Royal Society of Chemistry, 2007.

38. Oh D-B, Kim Y-G, Rich A. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proc Natl Acad Sci* 2002;**99**:16666–71.

39. Yang C-G, Garcia K, He C. Damage Detection and Base Flipping in Direct DNA Alkylation Repair. *ChemBioChem* 2009;**10**:417–23.

40. Nair DT, Johnson RE, Prakash S *et al.* Replication by human DNA polymerase-ι occurs by Hoogsteen base-pairing. *Nature* 2004;**430**:377–80.

41. Schwartz T, Behlke J, Lowenhaupt K *et al.* Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat Struct Biol* 2001;**8**:761–5.

42. Chung WJ, Heddi B, Schmitt E *et al.* Structure of a left-handed DNA G-quadruplex. *Proc Natl Acad Sci U S A* 2015;**112**:2729–33.

43. Arthanari H, Bolton PH. Functional and dysfunctional roles of quadruplex

DNA in cells. *Chem Biol* 2001;**8**:221–30.

44. Tiner WJ, Potaman VN, Sinden RR *et al.* The structure of intramolecular triplex DNA: atomic force microscopy study. *J Mol Biol* 2001;**314**:353–7.

45. Cubero E, Luque FJ, Orozco M. Theoretical study of the Hoogsteen-Watson-Crick junctions in DNA. *Biophys J* 2006;**90**:1000–8.

46. Huang M, Giese TJ, Lee T-S *et al.* Improvement of DNA and RNA Sugar Pucker Profiles from Semiempirical Quantum Methods. *J Chem Theory Comput* 2014;**10**:1538–45.

47. Wang X, Woods RJ. Insights into furanose solution conformations: beyond the two-state model. *J Biomol NMR* 2016;**64**:291–305.

48. Harvey SC, Prabhakaran M. Ribose puckering: structure, dynamics, energetics, and the pseudorotation cycle. *J Am Chem Soc* 1986;**108**:6128–36.

49. Fonseca R, Pachov D V., Bernauer J *et al.* Characterizing RNA ensembles from NMR data with kinematic models. *Nucleic Acids Res* 2014;**42**:9562–72.

50. Levitt M, Warshel A. Extreme conformational flexibility of the furanose ring in DNA and RNA. *J Am Chem Soc* 1978;**100**:2607–13.

51. Várnai P, Djuranovic D, Lavery R *et al.* Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res* 2002;**30**:5398–406.

52. Berman HM. Crystal studies of B-DNA: The answers and the questions. *Biopolymers* 1997;**44**:23–44.

53. Šponer J, Mládek A, Šponer JE *et al.* The DNA and RNA sugar–phosphate backbone emerges as the key player. An overview of quantum-chemical, structural biology and simulation studies. *Phys Chem Chem Phys* 2012;**14**:15257.

54. Tian Y, Kayatta M, Shultis K *et al.* ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J Phys Chem B* 2009;**113**:2596–603.

55. Klimasauskas S, Kumar S, Roberts RJ *et al.* HhaI methyltransferase flips its target base out of the DNA helix. *Cell* 1994;**76**:357–69.

56. Michael Trieb, Christine Rauch, Bernd Wellenzohn *et al.* Dynamics of DNA: BI and BII Phosphate Backbone Transitions. 2004, DOI: 10.1021/JP037079P.

57. Hartmann B, Piazzola D, Lavery R. BI-BII transitions in B-DNA. *Nucleic Acids Res* 1993;**21**:561–8.

58. Madhumalar A, Bansal M. Sequence Preference for BI/BII Conformations in DNA: MD and Crystal Structure Data Analysis. *J Biomol Struct Dyn* 2005;**23**:13–27.

59. Brahim Heddi §, Nicolas Foloppe *,†, Nadia Bouchemal ‡ *et al.* Quantification of DNA BI/BII Backbone States in Solution. Implications for

DNA Overall Structure and Recognition. 2006, DOI: 10.1021/JA061686J.

60. Imeddourene A Ben, Xu X, Zargarian L *et al.* The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res* 2016;**44**:3432–47.

61. Saenger W. *Principles of Nucleic Acid Structure*. New York, NY: Springer New York, 1984.

62. Dickerson RE, Drew HR, Conner BN *et al.* The Anatomy of A-, B-, and Z-DNA. *Science (80-)* 1982;**216**:475–85.

63. Battistini F, Hunter CA, Gardiner EJ *et al.* Structural Mechanics of DNA Wrapping in the Nucleosome. *J Mol Biol* 2010;**396**:264–79.

64. Tung CS, Harvey SC. Base sequence, local helix structure, and macroscopic curvature of A-DNA and B-DNA. *J Biol Chem* 1986;**261**:3700–9.

65. Goodsell DS, Dickerson RE. Bending and curvature calculations in B-DNA. *Nucleic Acids Res* 1994;**22**:5497–503.

66. Whelan DR, Hiscox TJ, Rood JI *et al.* Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation. *J R Soc Interface* 2014;**11**:20140454.

67. Pérez A, Noy A, Lankas F *et al.* The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res* 2004;**32**:6144–51.

68. Thamann TJ, Lord RC, Wang AH *et al.* The high salt form of poly(dG-dC).poly(dG-dC) is left-handed Z-DNA: Raman spectra of crystals and solutions. *Nucleic Acids Res* 1981;**9**:5443–57.

69. Pohl FM, Jovin TM. Salt-induced co-operative conformational change of a synthetic DNA: equilibrium and kinetic studies with poly (dG-dC). *J Mol Biol* 1972;**67**:375–96.

70. Zhang H, Yu H, Ren J *et al.* Reversible B/Z-DNA transition under the low salt condition and non-B-form polydApolydT selectivity by a cubane-like europium-L-aspartic acid complex. *Biophys J* 2006;**90**:3203–7.

71. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 1999;**98**:285–94.

72. Crick F. On Protein Synthesis. *Crick, Fr "On Protein Synth Symp Soc Exp Biol 12, 138-163 Artic 13 Images* 1958.

73. CRICK F. Central Dogma of Molecular Biology. *Nature* 1970;**227**:561–3.

74. Simons RW, Grunberg-Manago M. *RNA Structure and Function*. Cold Spring harbor Laboratory Press, 1997.

75. Neidle S. *Principles of Nucleic Acid Structure*. Elsevier, 2008.

76. Klostermeier D, Hammann C. *RNA Structure and Folding : Biophysical Techniques and Prediction Methods*.

77. Xu D, Landon T, Greenbaum NL *et al.* The electrostatic characteristics of G.U wobble base pairs. *Nucleic Acids Res* 2007;**35**:3836–47.

78. Varani G, McClain WH. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep* 2000;**1**:18–23.

79. Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA* 2001;**7**:499–512.

80. Moore PB. Structural Motifs in RNA. *Annu Rev Biochem* 1999;**68**:287–300.

81. Hendrix DK, Brenner SE, Holbrook SR. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys* 2005;**38**:221.

82. Leontis NB, Westhof E. The annotation of RNA motifs. *Comp Funct Genomics* 2002;**3**:518–24.

83. Lescoute A, Leontis NB, Massire C *et al.* Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res* 2005;**33**:2395–409.

84. Flores JK, Ataide SF. Structural Changes of RNA in Complex with Proteins in the SRP. *Front Mol Biosci* 2018;**5**:7.

85. Lescoute A, Westhof E. Topology of three-way junctions in folded RNAs. *RNA* 2006;**12**:83–93.

86. Lamiable A, Barth D, Denise A *et al.* Automated prediction of three-way junction topological families in RNA secondary structures. *Comput Biol Chem* 2012;**37**:1–5.

87. Laing C, Jung S, Kim N *et al.* Predicting Helical Topologies in RNA Junctions as Tree Graphs. Najmanovich RJ (ed.). *PLoS One* 2013;**8**:e71947.

88. Leontis NB, Westhof E. *RNA 3D Structure Analysis and Prediction*. Springer, 2012.

89. Washietl S, Hofacker IL, Stadler PF. From The Cover: Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* 2005;**102**:2454–9.

90. Weinreb C, Riesselman AJ, Ingraham JB *et al.* 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* 2016;**165**:963–75.

91. De Leonardis E, Lutz B, Ratz S *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res* 2015;**43**:10444–55.

92. Holbrook SR. Structural Principles From Large RNAs. *Annu Rev Biophys* 2008;**37**:445–64.

93. Tinoco I, Bustamante C. How RNA folds. *J Mol Biol* 1999;**293**:271–81.

94. Reiter NJ, Chan CW, Mondragón A. Emerging structural themes in large

RNA molecules. *Curr Opin Struct Biol* 2011;**21**:319–26.

95. Li PTX, Vieregg J, Tinoco I. How RNA Unfolds and Refolds. *Annu Rev Biochem* 2008;**77**:77–100.

96. Coimbatore Narayanan B, Westbrook J, Ghosh S *et al.* The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res* 2014;**42**:D114-22.

97. Suzuki M, Amano N, Kakinuma J *et al.* Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. *J Mol Biol* 1997;**274**:421–35.

98. Subirana JA, Faria T. Influence of sequence on the conformation of the B-DNA helix. *Biophys J* 1997;**73**:333–8.

99. Calladine CR. Mechanics of sequence-dependent stacking of bases in B-DNA. *J Mol Biol* 1982;**161**:343–52.

100. Olson WK, Gorin AA, Lu XJ *et al.* DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 1998;**95**:11163–8.

101. Sarai A, Jernigan RL, Mazur J. Interdependence of conformational variables in double-helical DNA. *Biophys J* 1996;**71**:1507–18.

102. de Lorenzo V. From the *selfish gene* to *selfish metabolism* : Revisiting the central dogma. *BioEssays* 2014;**36**:226–35.

103. Savelyev A, MacKerell AD. Differential Deformability of the DNA Minor Groove and Altered BI/BII Backbone Conformational Equilibrium by the Monovalent Ions Li $^{+}$, Na $^{+}$, K $^{+}$, and Rb $^{+}$ via Water-Mediated Hydrogen Bonding. *J Chem Theory Comput* 2015;**11**:4473–85.

104. Grokhovsky SL, Il'icheva IA, Nechipurenko DY *et al.* Sequence-specific ultrasonic cleavage of DNA. *Biophys J* 2011;**100**:117–25.

105. Maehigashi T, Hsiao C, Woods KK *et al.* B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res* 2012;**40**:3714–22.

106. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.

107. Zgarbová M, Jurečka P, Lankaš F *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* 2017;**57**:275–87.

108. Rohs R, Jin X, West SM *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem* 2010;**79**:233–69.

109. Djuranovic D, Hartmann B. Conformational Characteristics and

Correlations in Crystal Structures of Nucleic Acid Oligonucleotides: Evidence for Sub-states. *J Biomol Struct Dyn* 2003;**20**:771–88.

110. Anderson CF, Record MT. Salt-Nucleic Acid Interactions. *Annu Rev Phys Chem* 1995;**46**:657–700.

111. McConnell KJ, Beveridge DL. DNA Structure: What's in Charge? *J Mol Biol* 2000;**304**:803–20.

112. Feig M, Pettitt BM. Sodium and Chlorine Ions as Part of the DNA Solvation Shell. *Biophys J* 1999;**77**:1769–81.

113. Rueda M, Cubero E, Laughton CA *et al.* Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys J* 2004;**87**:800–11.

114. Atzori A, Liggi S, Laaksonen A *et al.* Base sequence specificity of counterion binding to DNA: what can MD simulations tell us? *Can J Chem* 2016;**94**:1181–8.

115. Donald Hamelberg †, Lori McFail-Isom ‡, Loren Dean Williams ‡ and *et al.* Flexible Structure of DNA: Ion Dependence of Minor-Groove Structure and Dynamics. 2000, DOI: 10.1021/JA000707L.

116. Kanhere A, Bansal M. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res* 2005;**33**:3165–75.

117. Heddi B, Foloppe N, Oguey C *et al.* Importance of Accurate DNA Structures in Solution: The Jun–Fos Model. *J Mol Biol* 2008;**382**:956–70.

118. Lankaš F, Šponer J, Hobza P *et al.* Sequence-dependent elastic properties of DNA 1 1Edited by I. Tinoco. *J Mol Biol* 2000;**299**:695–709.

119. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

120. Lavery R, Zakrzewska K, Beveridge D *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* 2010;**38**:299–313.

121. Gonzalez O, Petkevičiūtė D, Maddocks JH. A sequence-dependent rigid-base model of DNA. *J Chem Phys* 2013;**138**:55102.

122. Afek A, Lukatsky DB. Nonspecific Protein-DNA Binding Is Widespread in the Yeast Genome. *Biophys J* 2012;**102**:1881–8.

123. von Hippel PH, Berg OG. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A* 1986;**83**:1608–12.

124. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A* 1976;**73**:804–

8.

125. Harteis S, Schneider S. Making the Bend: DNA Tertiary Structure and Protein-DNA Interactions. *Int J Mol Sci* 2014;**15**:12335–63.

126. Wecker K, Bonnet MC, Meurs EF *et al.* The role of the phosphorus BI-BII transition in protein-DNA recognition: the NF-kappaB complex. *Nucleic Acids Res* 2002;**30**:4452–9.

127. Hunter CA. Sequence-dependent DNA Structure: The Role of Base Stacking Interactions. *J Mol Biol* 1993;**230**:1025–54.

128. Lavery R, Pullman B. The electrostatic field of DNA: the role of the nucleic acid conformation. *Nucleic Acids Res* 1982;**10**:4383–95.

129. Siggers T, Gordan R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res* 2014;**42**:2099–111.

130. Mirny L, Slutsky M, Wunderlich Z *et al.* How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J Phys A Math Theor* 2009;**42**:434013.

131. Kalodimos CG, Biris N, Bonvin AMJJ *et al.* Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science (80-)* 2004;**305**:386–9.

132. Ansari A, Kuznetsov S V. Dynamics and Mechanism of DNA-Bending Proteins in Binding Site Recognition. Springer, New York, NY, 2010, 107–42.

133. Zandarashvili L, Iwahara J. Temperature Dependence of Internal Motions of Protein Side-Chain NH $_3$ $^+$ Groups: Insight into Energy Barriers for Transient Breakage of Hydrogen Bonds. *Biochemistry* 2015;**54**:538–45.

134. Zandarashvili L, Esadze A, Iwahara J. NMR Studies on the Dynamics of Hydrogen Bonds and Ion Pairs Involving Lysine Side Chains of Proteins. *Adv Protein Chem Struct Biol* 2013;**93**:37–80.

135. Chen C, Esadze A, Zandarashvili L *et al.* Dynamic Equilibria of Short-Range Electrostatic Interactions at Molecular Interfaces of Protein–DNA Complexes. *J Phys Chem Lett* 2015;**6**:2733–7.

136. Etheve L, Martin J, Lavery R. Decomposing protein-DNA binding and recognition using simplified protein models. *Nucleic Acids Res* 2017;**45**:10270–83.

137. Etheve L, Martin J, Lavery R. Protein-DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors. *Nucleic Acids Res* 2016;**44**:9990–10002.

138. Deng Y, Roux B. Computations of standard binding free energies with molecular dynamics simulations. *J Phys Chem B* 2009;**113**:2234–46.

139. Donald JE, Chen WW, Shakhnovich EI. Energetics of protein-DNA

interactions. *Nucleic Acids Res* 2007;**35**:1039–47.

140. Sakthikumar K, Dhaveethu Raja J, Rajadurai Vijay S *et al.* Density Functional Theory Molecular Modelling, DNA interactions, Antioxidant, Antimicrobial, Anticancer and Biothermodynamic Studies of Bioactive Water Soluble Mixed Ligand Complexes. *J Biomol Struct Dyn* 2018:1–55.

141. Fox SJ, Dziedzic J, Fox T *et al.* Density functional theory calculations on entire proteins for free energies of binding: Application to a model polar binding site. *Proteins Struct Funct Bioinforma* 2014;**82**:3335–46.

142. Georgieva P, Himo F. Density functional theory study of the reaction mechanism of the DNA repairing enzyme alkylguanine alkyltransferase. *Chem Phys Lett* 2008;**463**:214–8.

143. Warshel A, Levitt M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol* 1976;**103**:227–49.

144. Šponer J, Riley KE, Hobza P. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys Chem Chem Phys* 2008;**10**:2595.

145. Banáš P, Jurečka P, Walter NG *et al.* Theoretical studies of RNA catalysis: Hybrid QM/MM methods and their comparison with MD and QM. *Methods* 2009;**49**:202–16.

146. Levitt M, Hirshberg M, Sharon R *et al.* Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comput Phys Commun* 1995;**91**:215–31.

147. Levitt M. Computer simulation of DNA double-helix dynamics. *Cold Spring Harb Symp Quant Biol* 1983;**47 Pt 1**:251–62.

148. Tidor B, Irikura KK, Brooks BR *et al.* Dynamics of DNA Oligomers. *J Biomol Struct Dyn* 1983;**1**:231–52.

149. Galindo-Murillo R, Roe DR, Cheatham TE. On the absence of intrahelical DNA dynamics on the μs to ms timescale. *Nat Commun* 2014;**5**:5152.

150. Vanommeslaeghe K, MacKerell AD, Jr. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim Biophys Acta* 2015;**1850**:861–71.

151. Lemkul JA, MacKerell AD. Polarizable Force Field for DNA Based on the Classical Drude Oscillator: I. Refinement Using Quantum Mechanical Base Stacking and Conformational Energetics. *J Chem Theory Comput* 2017;**13**:2053–71.

152. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:gkw1355.

153. Galindo-Murillo R, Robertson JC, Zgarbová M *et al.* Assessing the

Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* 2016;**12**:4114–27.

154. de Pablo JJ. Coarse-Grained Simulations of Macromolecules: From DNA to Nanocomposites. *Annu Rev Phys Chem* 2011;**62**:555–74.

155. Dršata T, Lankaš F. Multiscale modelling of DNA mechanics. *J Phys Condens Matter* 2015;**27**:323102.

156. Potoyan D, Papoian GA, Papoian GA. The Need for Computational Speed. *Coarse-Grained Modeling of Biomolecules*. CRC Press, 2017, 271–96.

157. Savelyev A, Papoian GA. Chemically accurate coarse graining of double-stranded DNA. *Proc Natl Acad Sci* 2010;**107**:20340–5.

158. Hadley KR, McCabe C. Coarse-Grained Molecular Models of Water: A Review. *Mol Simul* 2012;**38**:671–81.

159. Landau LD (Lev D, Lifshit□s□ EM (Evgeniĭ M, Kosevich AM *et al.* *Theory of Elasticity*. Butterworth-Heinemann, 1986.

160. Balaeff A, Mahadevan L, Schulten K. Elastic Rod Model of a DNA Loop in the Lac Operon. *Phys Rev Lett* 1999;**83**:4900–3.

161. Bouchiat C, Mezard M. Elastic Rod Model of a Supercoiled DNA Molecule. 1999, DOI: 10.1007/s101890050020.

162. Swigon D, Coleman BD, Tobias I. The Elastic Rod Model for DNA and Its Application to the Tertiary Structure of DNA Minicircles in Mononucleosomes. *Biophys J* 1998;**74**:2515–30.

163. Bustamante C, Marko JF, Siggia ED *et al.* Entropic elasticity of lambda-phage DNA. *Science* 1994;**265**:1599–600.

164. Marko JF, Siggia ED. Stretching DNA. *Macromolecules* 1995;**28**:8759–70.

165. Baumann CG, Smith SB, Bloomfield VA *et al.* Ionic effects on the elasticity of single DNA molecules. *Proc Natl Acad Sci U S A* 1997;**94**:6185–90.

166. Potoyan DA, Savelyev A, Papoian GA. Recent successes in coarse-grained modeling of DNA. *Wiley Interdiscip Rev Comput Mol Sci* 2013;**3**:69–83.

167. Shi Y-Z, Wu Y-Y, Wang F-H *et al.* RNA structure prediction: progress and perspective. 2014, DOI: 10.1088/1674-1056/23/7/078701.

168. Cragnolini T, Derreumaux P, Pasquali S. Ab initio RNA folding. 2014, DOI: 10.1088/0953-8984/27/23/233102.

169. Miao Z, Westhof E. RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annu Rev Biophys* 2017;**46**:483–503.

170. Dawson WK, Bujnicki JM. Computational modeling of RNA 3D structures and interactions. *Curr Opin Struct Biol* 2016;**37**:22–8.

171. Krieger E, Nabuurs SB, Vriend G. Homology modeling. *Methods Biochem*

Anal 2003;**44**:509–23.

172. Capriotti E, Marti-Renom MA. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics* 2010;**11**:322.

173. El Fatmi A, Chentoufi A, Bekri MA *et al.* RNA Secondary Structure an Overview. 2018, 379–88.

174. Miao Z, Adamiak RW, Blanchet M-F *et al. RNA-Puzzles* Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* 2015;**21**:1066–84.

175. Miao Z, Adamiak RW, Antczak M *et al.* RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* 2017;**23**:655–72.

176. Cruz JA, Blanchet M-F, Boniecki M *et al.* RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* 2012;**18**:610–25.

177. Magnus M, Matelska D, Lach G *et al.* Computational modeling of RNA 3D structures, with the aid of experimental restraints. *RNA Biol* 2014;**11**:522–36.

OBJECTIVES

The main objective of this thesis is to offer a comprehensive and consensual view of the structural and dynamic properties of DNA under physiological conditions. The works presented here have gradually built on one another, with results slowly gathering up to form a truly convoluted fresco of interdependent mechanisms. The order in which the results are presented is not always chronological. In the process of a PhD thesis there are many situations when one is stuck or cannot see the forest for the trees, but in retrospective, analyzing the final outcome, it is easy to build a succession of gradual achievements. Below I present the logical build-up of objectives, as seen in the aftermath of a PhD thesis.

- **Benchmarking** of the parmbsc1 state-of-the-art DNA force field by testing it on a large variety of DNA systems under various conditions. This is clearly a prerequisite for using MD simulations in the study of B-DNA with confidence. Trajectories have to be proven to sample the B-DNA conformational basin thoroughly and extensively.

- **Explaining B-DNA polymorphisms** is likely to be the key for elucidating the puzzle of its intricate sequence-dependent mechanical properties that ultimately govern most of the biologically relevant functions of the double helix.

- Developing an exhaustive set of rules that govern **B-DNA sequence-effects at the tetranucleotide level**. We combine the new parmBSC1 force-field and the latest knowledge in the area of polymorphisms in the helical space, to bring a complete description and explanation at the tetranucleotide level of the different base, base pair, and base pair step polymorphisms, and their interconnections.

- Deciphering the **higher-than-tetramer effects** on the conformational landscape of the B-DNA, in order to be aware of their contribution to DNA dynamics. We aim to figure out the strength, relevance and ultimately the mechanisms of long-range conformational modulation by specific sequence patterns.

- Applying knowledge of intrinsic DNA properties to the study of protein-DNA recognition and **cooperative protein binding to the DNA**. We finally set out to uncover how long-range communication through the DNA, as demonstrated from sequence effects, impact on its role in protein-DNA interactions.

- Sketching out a compendium of **computational approaches for the modeling of RNA**, which forces researchers nowadays to look beyond common classical or quantum simulation schemes. We aim to summarize scope and challenge of the most recent approaches created to characterize the large conformational landscape of RNA, which should help guide the development of a new generation of methods able to make quantitative predictions on the structure and physical properties of RNA.

CHAPTER II | Classical Atomistic Methods in Computer Simulations of Nucleic Acids

1 Molecular Dynamics Algorithms

Scientific investigation involves both observation and comprehension, where observation amounts to setting up an experiment to obtain information regarding a specifically formulated question and comprehension is usually achieved by developing theoretical models to describe the observed behaviors. Computer simulation is a third way of doing science, different from both experiment and theory. It is in fact a numerical experiment being run in the virtual laboratory of a computer's processing unit. It can provide in the same time new observational data, facilitate understanding of the contributing factors for the ensemble properties and it additionally has predictive power in a large number of problems. Simulation is, as a consequence, part of a feedback loop, and interpretation of its results has value only in a context where a theoretically plausible basic model has been successfully employed to reproduce and predict experimental observation.

Understanding matter at the microscopic level can be reduced to a classical many-body problem that can be treated, at least conceptually, within the framework of statistical mechanics. This approach provides a formal description – based on the partition function – of a system in equilibrium. However, with a few notable exceptions, there are no quantitative answers unless severe approximations are introduced in sampling, in the definition of the Hamiltonian and in the size of the system. Calculating the partition function and associated thermodynamic and equilibrium properties for a general many-body potential that includes nonlinear interactions becomes an insurmountable task if only analytical techniques are employed. Molecular dynamics (MD) simulations are an attempt to avoid much of the approximation normally associated with analytical theory, replacing it by a numerical solution.

There are three principal aspects to a MD calculation: 1) the model describing molecular interactions; 2) the calculation of energies and forces from the model, which should be done accurately and efficiently; 3) the algorithm used to integrate the equations of motion. In the simplest form of MD, the trajectories of atoms and molecules are determined by numerically solving Newton's equations of motion for a system of interacting particles [1–3]. By default equilibrium MD corresponds to the microcanonical ensemble of statistical mechanics (constant total energy), but by coupling the system with an external heat or pressure environment (i.e. by coupling simulations to thermostats and/or barostats) isobaric and/or isothermal conditions can be simulated. Once all the details of a real-like system are introduced, the equations of motion can only be solved numerically due to the complex time dependence of forces.

1.1 The model.

Molecular dynamics simulation requires the definition of a potential energy function. This can be a quantum Hamiltonian (ab initio or QM molecular dynamics simulations), but in most cases for obvious computational reasons a classical representation of the system is used (the force field). The reduction from a fully quantum description to a classical potential entails two main approximations. The first one is the Born–Oppenheimer approximation, which states that the dynamics of electrons are so fast that they can be considered to react instantaneously to the motion of their nuclei. As a consequence, they may be treated separately. The second one treats the nuclei, classical particles following Newtonian dynamics. The quantum-mechanical effects of the electrons are represented implicitly, as functional approximations containing parameters such as atomic partial charges, van der Waals atomic radii and hardness estimates, equilibrium bond lengths and angles with the stiffness associated to their perturbations, and Fourier expansions associated to the energy profile of torsions (Figure 2.1). The values of the parameters are obtained by fitting against detailed electronic calculations (quantum mechanical simulations) or a variety of experimental physical properties. The most common force fields consist of a summation of *bonded forces* (associated with chemical bond lengths, bond angles, and bond dihedrals) and *non-bonded forces* (associated electrostatic interactions and Van der Waals interactions).

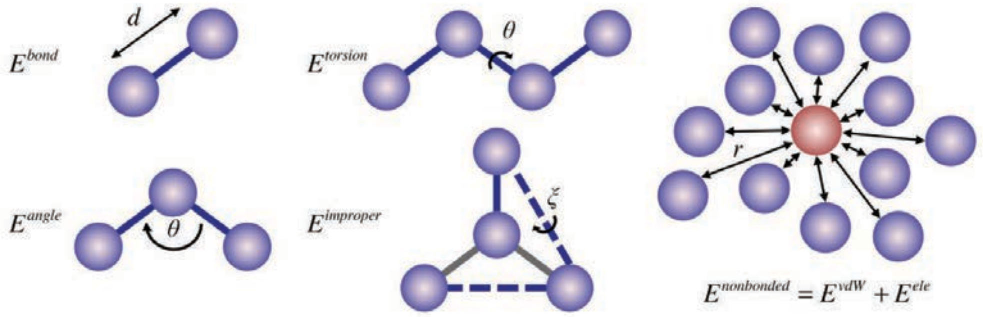


Figure 1 Schematic illustration of the terms in a classical fixed-charge force field, i.e. bond stretching (Ebond), bond-angle bending (E_{angle}), dihedral- angle torsion ($E_{torsion}$), and improper dihedral-angle bending ($E_{improper}$) as well as van der Waals (E_{vdW}) and electrostatic (E_{ele}) interactions.

The ***bonded potential terms*** generally consist of a sum of 2-, 3-, and 4-body interactions of covalently bonded atoms, which are summarized below:

• *bond stretching* The standard way to approximate the potential energy for a covalent bond in a molecule is to use a Hooke's law term, thus representing the bond as a spring linking the two atoms. The energy potential well is parabolic in this approximation, so it only holds for small fluctuations around the equilibrium value:

$$E_{bond} = \sum_{bonds} k_r \cdot (r - r_0)^2 \qquad (2.1)$$

which states that the force acting between particles is proportional to the force constant kr and the square relative distance (r-r_0)^2, from the equilibrium position r_0.

• *bond angle bending* is the fluctuation of the angle between two consecutive bonds (defined by three atoms). As bond angles are found (experimentally and theoretically) to vary around a single value it is sufficient in most applications to use a harmonic representation (in a similar manner to the bond potential) using Hooke's law as:

$$E_{angle} = \sum_{angle} k_a \cdot (a - a_0)^2 \qquad (2.2)$$

where k_a is the force constant acting on the relative angle $(a - a_0)$with a given equilibrium value a_0.

The equilibrium values and force constants (for both bond and angle terms) can be obtained from vibrational analysis of the molecule (experimental or theoretical).

• *torsion angle twisting* around a bond of a dihedral ω (defined by four adjacent atoms) cannot be described by a harmonic term, because of its periodicity and low internal rotation barriers. The form of the torsional potential term is also extremely versatile, depending on the atoms forming it, and the chosen functional has to reflect this fact. It is most common to model the torsional interaction using a Fourier series:

$$E_{dih} = \sum_{dihedrals} \sum_{n=1}^{4} \frac{V_{dih}}{2} \cdot (1 - \cos(n \cdot w - \phi)) \qquad (2.3)$$

where V_{dih} represents the torsional barrier, n the periodicity (when modeling organic compounds between 1 and 4 terms are generally used in the series) and ϕ the phase angle. Closely related to the torsional interaction are the out-of-plane distortions, or improper dihedral angles. They can be accounted for in force fields in one of two main ways, either treated harmonically in a similar way to bending terms or using a two-fold Fourier term.

Non-bonded interactions act practically between all pairs of atoms in the system, both intra- and inter-molecular. Force fields usually divide non-bonded interactions in two contributions:

- The *electrostatic interaction* arises due to distribution of charge in a molecule, which can be modeled by placing point charges on each particle and evaluating the interaction between each pair of atoms i and j by a Coulomb potential:

$$E_{elec} = \sum_{i,j} \frac{1}{4\pi\varepsilon} \cdot \frac{q_i q_j}{r_{ij}} \qquad (2.4)$$

where ε is the dielectric constant of the medium and r_{ij} the distance between two charges q_i and q_j, associated to particles i and j.

- The *van der Waals interaction* consists of the residual attractive and repulsive forces that count for dispersion interaction and Pauli repulsion. A good approximation of van der Waals interactions is Lennard-Jones potential:

$$E_{vdW} = \sum_{i,j} \varepsilon^* \left[\left(\frac{r_m}{r_{ij}}\right)^{12} - \left(\frac{r_m}{r_{ij}}\right)^{6} \right] \qquad (2.5)$$

where ε* is the depth of the potential well, r_m is the distance at which the potential it's minimum for the given pair i and j and r_{ij} is the distance between particle i and j. In the Lennard-Jones potential the r^{-12} term accounts for the short distance repulsion, while r^{-6} term is used to represent dispersion interactions. The r^{-6} dependence of the attractive term arises from induced dipole-induced dipole dispersive attraction, averaged this over all orientations. There are no physical arguments for choosing the repulsive term to vary as r^{-12}: this arises due to computational expediency. In fact, alternative algorithms using exponential functions of other polynomial expansions have been suggested in the literature [4–6].

The nonbonded terms of the ubiquitous fixed-charge model approximations do not explicitly include polarizability, the process by which the charge distribution in an atom or molecule changes in response to its environment. Polarization can be introduced at the classical level by using induced dipoles, fluctuating charges, or Drude oscilators, but in most cases it is included implicitly by assigning partial atomic charges that overestimate molecular dipoles [7,8], simulating then the increase in dipole interactions produced by polarization.

1.2 Evaluation of Energies and Forces.

The stretching bending and torsion terms are straightforward and computationally inexpensive to evaluate. By far, the most expensive part of the calculation is evaluating the nonbonded (nb) forces and energies given by the Lennard-Jones and Coulomb terms. Because of the non-local nature of non-bonded interactions, they involve all particles in the system, i.e. the cost of calculation increases with the second power of the number of particles. The problem is magnified as periodic boundary conditions are used to simulate an infinite system. Approximations are then used to reduce the number of interactions computed and reduce the computational cost accordingly. The van der Waals term falls off rapidly with distance, so the numerical solution is fairly insensitive to introducing a shifted "cut-off" in their calculation. However, the Coulomb potential decays slowly, with r^{-1}, and would suffer from major truncation artifacts if imposed a cutoff [9]. In order to obtain rapid convergence of the energy without a high penalty on accuracy, mathematical tricks can be employed. A very common approach is the use of the particle-mesh Ewald (PME) method [10] – or more recent variations of it such as particle-particle-particle-mesh (P3M) or smooth PME – that makes use of the fact that any long-range function in real space becomes short-ranged in reciprocal Fourier space. Therefore, the long-range interaction is divided into two parts: a short-range potential, calculated in the real space for close neighbors, and a long-range potential, which is calculated in the Fourier space. The method requires charge neutrality of the molecular system in order to calculate accurately the total Coulombic interaction and, more importantly, it implicitly assumes periodic symmetry. Luckily, as noted above, molecular dynamics simulations use, in many cases, periodic boundaries, in order to mimic an infinite system, which fits well with the intrinsic requirements of Ewald's methods. The configuration of a system with periodic boundaries is achieved by considering an infinite array of identical copies of the simulation region (a tetrahedron, dodecahedron or truncated octahedron) that extends in every direction (Figure 2.2). Any molecule leaving the unit cell through a particular bounding face immediately reenters the region through the opposite face, thus keeping the total number of atoms in the central simulation box constant.

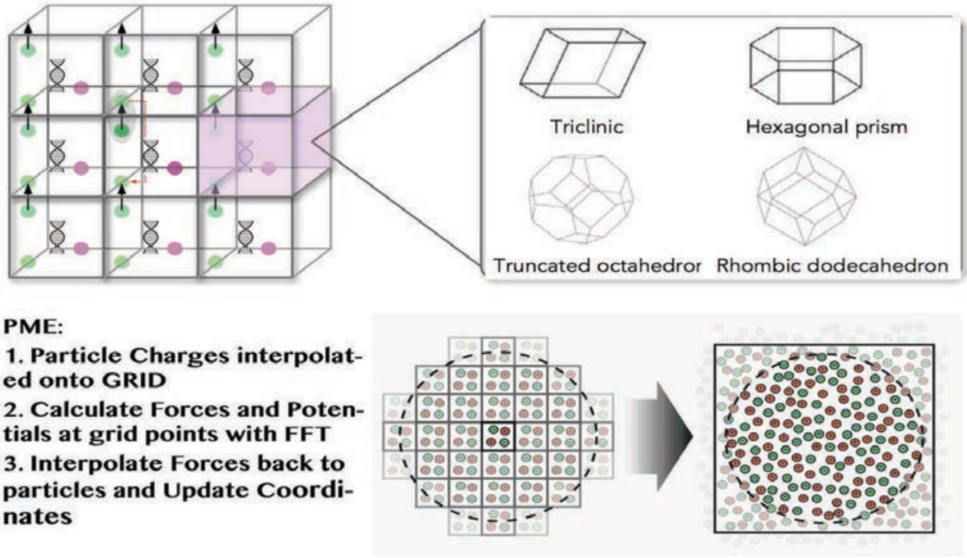

Triclinic
Hexagonal prism
Truncated octahedror
Rhombic dodecahedron
PME:
1. Particle Charges interpolated onto GRID
2. Calculate Forces and Potentials at grid points with FFT
3. Interpolate Forces back to particles and Update Coordinates

Figure 2 An illustration of the PBC and PME in MD simulations. Top: Basic concept and most commonly used 3D unit cells in PCB; Bottom: Basic steps for PME integration of nonbonded forces.

1.3 Numerical Integrators.

The Verlet Integrator (or variants of it such as Leap-Frog or velocity Verlet), introduced by in 1967 [9], is still the most popular algorithm for solving Newton's equations of motions of a simulated system. Verlet and any other numerical methods for integration are *symplectic* and *time-reversible*. Sympletic means that when the methods are guaranteed to conserve total energy (more correctly, the Hamiltonian) in conservative simulation problems. Time-reversibility is a fundamental symmetry of Hamilton's equations that should be preserved by a numerical integrator.

Starting from a given form of the potential energy together with a set of initial conditions (including atomic positions and velocities), the time evolution of the system can be obtained by iterative numerical integration. First, the forces acting on each particle can be computed as the negative gradient of its potential energy U:

$$\vec{F}(X) = -\nabla U(X) \tag{2.6}$$

Next, making use of Newton's equations of motion

$$\vec{F}(X) = m\vec{a}(X) = m\frac{d^2\vec{x}}{dt^2} \tag{2.7}$$

and expressing a particle's position iteratively in increments of time step Δt, the position at time t+Δt is obtained using a Taylor series expansion in terms of its position, velocity, and acceleration at time t according to:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\partial^3 r}{\partial t^3}\frac{\Delta t^3}{3!} + \emptyset(\Delta t^4) \tag{2.8}$$

where the expansion goes up to the second order derivative in Δt. The Verlet algorithm can then generate atomic positions for an arbitrary length of time at each time step with:

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \frac{f(t)}{m}\Delta t^2 + \emptyset(\Delta t^4) \tag{2.9}$$

If needed, the velocities can be constructed at any point in the trajectory via

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} + \emptyset(\Delta t^2) \tag{2.10}$$

More elegantly, the velocity Verlet algorithm explicitly evolves the velocities along with the positions, thus fully defining each time point in the phase space.

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t \tag{2.11}$$

$$v(t+\Delta t) = v(t) + \frac{f(t+\Delta t) + f(t)}{2m2\Delta t}\Delta t \tag{2.12}$$

Similarly, the 'Leap-Frog' algorithm generates phase space vectors (composed of both positions and velocities) at discrete times, but coordinates and velocities are evaluated at different times, a detail that gave it its name.

So how important is the initial state for an iterative numeric integrator? Theoretically, it is of no importance whatsoever as, given an infinite amount of time, the system would be able to visit all configurations on the constant energy hypersurface (ergodicity), and accordingly the results of a simulation of adequate duration are insensitive to the initial state, so that any convenient initial state is allowed. However, in real systems this is not usually the case and local barriers in potential energy often appear, even when the total energy is conserved, which means that when using finite sampling times, original coordinates are important to guarantee that our simulation is sampling reliable regions of the conformational space. For complex systems such as biological macromolecules, it is usually necessary to obtain initial coordinates from an experimental X-ray or NMR structure. Once initial coordinates are specified, the initial velocities are typically assigned randomly from the Maxwell-Boltzmann distribution, taking care to ensure that they are consistent with any constraints imposed on the system (such as temperature).

The integration time step (Δt) is chosen to be smaller than the fastest motion in the system, which for biological system is the bond stretching of a hydrogen atom, happening on below the femtosecond timescale. Increasing the time step beyond this number potentially makes the simulation unstable. However, in order to speed up calculations, methods for removing or slowing down the highest-frequency motions of the macromolecule under study have been developed, that can therefore afford a longer time step. The use of such methods introduces, apart from simulation stability issues, also some formal issues of preserving an integration scheme that gives exact solutions, which are addressed within these algorithms.

The most common of these constraining methods consist of the freezing of bonds involving hydrogen by means of special algorithms, most notably SHAKE [11] (used in AMBER simulation packages), LINCS [12] (used in GROMACS package) or RATTLE [12] (used in NAMD package), which allow the increase of integration steps by up to a factor of 2 (2 femtoseconds). This is done by imposing a set of holonomic constraints, that is, constraints that depend only on the positions of the particles involved. The 2-fs limit is due to limitations in the algorithms themselves. The constraint conditions have to be exactly satisfied within a particular numerical integration scheme, a feat formally achieved by computing at each step a set of Lagrange multipliers for enforcing the constraints.

A different approach involves slowing down the high frequency vibrations by repartitioning the mass of heavy atoms into the bonded hydrogens, a method called hydrogen mass repartitioning (HMR). The idea of changing atomic masses in order to speed up MD simulations can be traced back to the 1970s [12], but has only more recently been implemented in popular MD simulation software (GROMACS, NAMD, AMBER or ACEMD)

[13]. The main idea behind the method is that equilibrium thermodynamic averages of observables are not dependent on the exact mass distribution of the system. This is due to the fact that in classical MD force fields without magnetic terms, the Hamiltonian is separable in position and momentum. Importantly, the total mass of the system should be kept constant when repartitioning, in order to achieve a true speed up of the simulation, as shown in the Feenstra et al. seminal study [14]. Some authors have used this type of methods to perform simulations with a 4 fs time scale, increasing the throughput of MD calculations.

Finally, some simulation packages (notably NAMD, but also AMBER) have implemented multiple time scale (MTS) methods that divide the computation into "slow" and "fast" portions, assigning appropriate time steps to each segment. Most commonly, the scheme is distance-based and partitions between bonded, short-range nonbonded and long-range electrostatic interactions. Here too is it important to ensure that the integration scheme gives an exact solution to Hamilton's equations, remains reversible and evolves in a symplectic manner. A very powerful implementation of the MTS method is through the reversible reference system propagator algorithm (r-RESPA) [14], derived from the Trotter factorization of the Liouville propagator.

As mentioned before, MD simulations rely on integrating the classical equations of motion for a molecular system and thus, sample a microcanonical ensemble by default, where the number of particles N, the volume V and the total energy E are conserved (NVE ensemble). However, for compatibility with experiment, it is often desirable to sample configurations with constant temperature and/or pressure. Thus, conditions more similar to experimental ones (other ensembles) can be obtained by applying specific modifications to the system Hamiltonian or equations of motion. A modification of the basic MD scheme with the purpose of maintaining the average temperature constant (NVT ensemble) is called a thermostat algorithm. Popular techniques to control temperature include velocity rescaling [15], the Andersen thermostat [15], the Nosé–Hoover thermostat [16,17], Nosé–Hoover chains, the Berendsen thermostat [18] and Langevin dynamics [19,20]. But most experimental observations are performed at constant temperature and pressure, so it is desirable to run MD simulations in the isothermal-isobaric (NPT) ensemble. Similarly to the temperature coupling schemes, an extra term is added to the equations of motion that additionally effects a pressure change (barostat algorithms). Notable examples are the Berendsen barostat, Nose-Hoover bath, or Parrinello-Rahman barostat [21,22]. Generally, stochastic models have the drawback of non-reproducibility of the trajectory and lack of a conserved quantity. Deterministic algorithms can equilibrate very slowly depending on the scheme and even worse, tend to lose ergodicity. Langevin dynamics, although very efficient, are extremely sensitive to the choice of friction coefficient and might also be slow to equilibrate. Knowing the advantages and disadvantages of each coupling method is important for designing a simulation problem.

Solvent-induced effects are extremely important for the thermodynamics and conformational properties of biomolecules, since have evolved to function in an environment of water and ions. Solvent can be represented either implicitly, as a continuous medium, by approximating the mean force exerted by the media on the solute, or explicitly, by inclusion of water molecules and ions. Implicit water models can be considerably faster to compute, because the implicit solvent contributes no or few degrees of freedom to the simulation. 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, up to certain system sizes, it is common to use explicit solvents, composed of a water model in combination with an ion parametrization. Water models that represent a good compromise between accuracy and computational cost are the TIP3P, TIP4P [23], or SPC/E [24]. More elaborate models introducing extra centers are still not much employed in the DNA field, but those retuning of current 3- or 4- point models are starting to be used [25].

2 Force field development methods.

The algorithmic advances discussed in the previous Section, such as the extended and rigorous representation of the potential energy function, together with appropriate boundary methods and suitable integrators have greatly increased the quality of force fields for biological macromolecules over the years. However, the algorithmic set-up is not a force field without the assignment of parameters and atom types that ultimately dictate its quality and applicability. The general potential function

$$
\begin{aligned}
U(R) = \sum_{bonds} k_r \cdot (b - b_0)^2 + \sum_{angle} k_a \cdot (a - a_0)^2 \\
+ \sum_{dihedrals} \sum_{n=1}^{4} \frac{V_{dih}}{2} \cdot (1 - \cos(n \cdot w - \phi)) \\
+ \sum_{i,j} \frac{1}{4\pi\varepsilon} \cdot \frac{q_i q_j}{r_{ij}} + \sum_{i,j} \varepsilon^* \left[\left(\frac{r_m}{r_{ij}}\right)^{12} - \left(\frac{r_m}{r_{ij}}\right)^{6} \right]
\end{aligned}
\qquad (2.13)
$$

needs properly adjusted parameters for each functional to yield an accurate representation of a biomolecule conformational space and thermodynamics. In eq. 2.13, b is the bond length, θ is the valence angle, ω is the dihedral or torsion angle, ϕ is the improper angle, and r_{ij} is the distance between atoms i and j. Parameters, the terms that represent the actual force field, include force constants and equilibrium values for distances, valence angles and improper dihedrals, the torsional force constants, multiplicities and phase angles for the dihedral rotations potential. Collectively, these represent the bonded parameters. Nonbonded parameters between atoms i and j include the partial atomic charges for the Coulomb potential, and the LJ well-depth and minimum interaction radius used to treat the van der Waals (vdW) interactions. Different parameters need to be set for different atom types. Atom types assignment depends on the functional group that the atom is part of (molecular environment) and/or hybridization state. The dielectric constant, ε, is typically set to 1, corresponding to the permittivity of vacuum, in calculations that incorporate explicit solvent representations. An overview of the common assumptions and methodologies employed to determine both bonded and nonbonded parameters is given below.

Bonded parameter determination.

The equilibrium bond and angles can be derived from experimental spectroscopic measures (X-Ray, neutron diffraction, IR and Raman spectra), while stretching and bending constant can be obtained from analysis of IR and Ramn spectra. Alternatively, both equilibrium and force constants can be obtained by fitting to high-level QM calculations.

For torsion parameters the dihedral parameters can be fitted in some cases to reproduce NMR J-couplings, but in most cases torsional terms are derived by fitting to QM conformational energy profiles of rotation about selected bonds to Potential of Mean Force (PMF) simulations with gradually adjusted parameters. Exact reproduction of gas-phase QM data may yield geometries not appropriate for condensed phase MD simulations, therefore it is important to either use solvent corrections in the QM calculations directly, or to further optimize parameters taking experimental data into account. Experimental observables obtained from X-Ray (structural data stored in the protein or nucleic acids database) or NMR data are often used to assess accuracy of the parameters.

Non-bonded parameter optimization. Proper optimization of nonbonded parameters is both essential and more complicated than that of their bonded counterparts. For simple liquids and simple systems a parametrization scheme developed by Jorgensen and others is the simultaneous fitting of electrostatic and van der Waals terms in an iterative process where force-fields are adjusted to guarantee that the simulation reproduces a set of experimental observables of the system (density, compressibility, permittivity, radial distribution functions, heat of vaporization and many others). Alternatively, for more complex systems, independent parametrization of electrostatic and van der Waals terms is required. For the later transferability is assumed and atomic parameters (hardness and radii) are transferred from crystal lattice measures: while electrostatic parameters (charges) are derived from QM calculations.

Two main strategies have emerged to fit charges: namely the Electrostatic Potential (ESP) fitting methods and supramolecular approaches. *ESP-based methods* optimize atomic charges to reproduce a QM determined ESP mapped onto a grid surrounding a model compound [26,27]. Typically it requires the use of restraints during fitting in order to properly determine charges of "buried" atom (RESP fitting). In *supramolecular approaches* the charges are optimized to reproduce QM determined energies and geometries of interacting pairs of molecules, usually the model compound with individual water molecules [28]. This approach is quite laborious and therefore less popular, but in some situations it might be preferable as guarantee a good coupling of the molecular and water force-fields. Both the ESP and supramolecular approaches tend to overestimate dipole moments [29] and interaction energies, but such overestimations are desirable for additive force fields, as they lead to partial charge distributions that include the implicit polarization required for condensed phase simulations.

Interdependence of force field parameters. Finally, it should be emphasized that the parameters within a force field are, to various degrees, correlated to eachother. The most significant interdependence is between vdW parameters and the partial atomic charges, which implies that consistence between the two sets of parameters should be evaluated before adopting a force-field. Additionally, some of the internal bond parameters are dependent on the nonbonded parameters. This is most obvious in the so-called 1,4 interactions, between atoms three bonds away from eachother, where the energy surface of rotation about a bond will be determined not only by the dominant dihedral term, but will also contain contributions from the electrostatic and LJ terms [29]. Actually, the nonbonded terms sometimes need to be scaled down in such configurations, in order not to significantly alter the shape of the potential surface after direct fitting to QM profiles. In 1,2 (covalently bonded atoms) and 1,3 interactions (atoms separated by two bonds), the nonbonded terms are usually completely removed. Such correlations make it generally inadvisable to combine parameters from different force fields, since there is no guarantee to still maintain the proper balance of the intra- and intermolecular forces. Finally, the importance of using the correct water model with a given force field must be mentioned, since the nonbonded parameters in a force field are optimized to be compatible with a specific water model.

3 Optimization of Force Fields for Nucleic Acids.

In the particular case of NA simulations, force field optimization, and thus parameter adjustment, can rely not only on a subset of atom types from small systems that are then assembled to describe the macromolecule, but the four types of nucleotide units can be parametrized independently, taking advantage of the simple oligomeric nature of the double helix. However, the polyanionic nature of NAs has posed a great challenge for empirical force field development, requiring accurate treatment of interactions with the solvent environment and great care to the balance between stability and proper sampling of conformational landscape. As a result, early force fields for DNA had very little success in producing stable MD simulations. Intuitive tricks were used in order to improve stability, such as removing the charges on the phosphates, the inclusion of "solvated" sodium ions and the use of a distance dependent dielectric constant to mimic the solvent environment [30–33]. Particularly, early force field attempts to perform simulations of DNA with an explicit representation of the solvent (thus trying to mimic physiological conditions), have been consistently unsuccessful.

Significant progress in DNA simulations occurred in the mid 1990s with the development of the "second generation" force fields for nucleic acids, facilitated by the development of Ewald methods to treat long range electrostatic interactions, and the use of periodic boundary conditions [10]. These force fields included the Cornell et al AMBER (PARM94) [34] and CHARMM all-atom [35] force fields, both of which produced stable simulations in the nanosecond scale, but still had critical systematic problems. With CHARMM22 there was a strong tendency towards A-form duplex DNA structures, even at low ion concentration [36], while AMBER PARM94 had problems with sugar puckering and under-twisting of duplexes [37]. Once these deficiencies started to emerge, developers of both force fields made attempts at improving their parameter set, with CHARMM going through a full reoptimization process, while AMBER focused on the glycosidic and torsion parameters. Although changes in dihedral parameters were found to influence other parameters and lead to structural inconsistencies, and high-level QM calculation were not yet available to be used as reference, these force-fields produced equilibrated trajectories up to 10 ns that sampled conformations of DNA not far from the experimental ones. CHARMM27 still had problems with B to A-form transitions as a function of salt concentration; but real problems with the AMBER ff99 were only detected once improvement in computer capabilities allowed for somewhat longer time scales of 50-100 ns [38–40], where big distortions in the structure emerge. Analysis of the trajectories determined that these distortions were related to a disproportionate α/γ populations of the gauche+/trans geometry, in detriment of the canonical gauche-/gauche+ state in AMBER f99, producing "ladder-like" structures of DNA duplexes and unnatural widening of the minor grooves in CHARMM27. The most notable improvement at that time came in the form of a force field correction to AMBER ff9: the PARMBSC0 parameterization [38][, which allowed the simulation of stable DNAs in the multi-nanosecond regime. For a decade, parmbsc0 was the "gold standard" of nucleic acids force fields, extensively used to simulate a variety of nucleic acids in the sub-microsecond timescale [41], producing almost 1500 citations up to date and significant contributions to the general understanding of DNA structure and dynamics [42,43].

Gradually, as multi-microsecond simulations became available, several errors in the parmbsc0 parametrization emerged and required to be addressed [44–50]. Among the noted inaccuracies was an underestimation of average twist values compared to NMR and X-Ray data that can build up to significant structural error in long oligos. The BI-BII equilibrium, which has been shown to be connected to the bimodality of twist distributions, especially in RpY step [45,50] was also misrepresented, with a bias towards the canonical BI state [46]. The puckering was still not up to par with an abnormal East population, and low flexibility of the glycosidic torsion [47] meant applications to non-canonical structures of DNA were unreliable [44,48,49]. Terminal base fraying was too large, giving unrealistic configurations at the ends of the duplex [45].

As expected, corrections to the parmbsc0 force field prompted new developments, either addressing particular cases of exotic DNA forms, or trying to come up with a new standard for DNA simulations. Worth noting are the OL1 [51] parameter set, created to improve ε/ζ representation for a better consensus of BI/BII population, or the OL4 [49] patch that aimed to correct the χ angle distribution, followed by the novel OL15, which incorporated all the previous OL corrections for DNA and included additional adjustments of the β torsion [52]. Other approaches meant imposing harmonic restraints derived from NMR measures to guarantee a good representation of the BI/BII equilibrium [46].

The proposed modifications and corrections to parmbsc0 mentioned above have proved useful to correct some, but not all the problems of DNA simulations, making it necessary to develop a new general-purpose AMBER force-filed for DNA simulations. Our own contribution to this effort in the form of the parmBSC1 force field will be presented and discussed in the Results section (Chapter III).

In parallel with DNA force-field efforts several groups have focused on developing force field corrections specifically for RNA, especially by means of reparametrization of the backbone and glycosidic torsion parameters [53], on the basis of QM calculations. While native states could generally be correctly described by these force fields, significant imbalances arose from several key issues, such as the overestimation of nucleobase stacking and underestimation of base pairing strength, could only be addressed by modifying electrostatic and van der Waals (vdW) parameters [54,55]. Chen and García [56] scaled down vdW interactions of nucleobase atoms to weaken stacking and strengthen base pairing. The Shaw group have implemented a variety of corrections, modifying electrostatic, vdW, and torsional parameters of the AMBER ff14 RNA force field [57] to more accurately reproduce the energetics of nucleobase stacking, base pairing, and key torsional conformers obtained from ab-initio and empirical methods. The overstacking is not exclusively an effect in RNAs [54,55], therefore these modifications developed for RNAs can in some cases be applied to study exotic forms of DNA structures [58].

4 Applications for MD simulations of Nucleic Acids

From an extensive sampling of an equilibrium conformation, both time-averaged and dynamic properties can be extracted. The information of interest needs to be carefully revealed from noisy data, such that in many cases the limiting step of conducting *in-silico* molecular experiments is not the generation of the trajectory, but its analysis.

4.1 MD-Averaged Information

A trajectory is a time-dependent sampling of the global conformational space of a molecular system, where an important class of ensemble properties corresponds to time-averaged structural descriptors. The average conformations are expected to represent global minima and can be both directly compared to experimental observables and used to check the convergence of the trajectory. Convergence is usually evaluated by comparing average structural parameters calculated for the first or second part of the trajectory with those calculated over the full trajectory. Typically, for DNA there are two major types of conformational representations: those based on Cartesian coordinates and those based on the internal helical coordinates. In either case, the first step of extracting MD-averaged structures is the root-mean square fitting of the snapshots to define a common reference system, which is necessary either to compute average Cartesian coordinates or to build the average helical axis.

Analysis in Cartesian Space.

A common measure in Cartesian space is the root-mean-square deviation (RMSd), which quantifies the minimum deviation of atomic positions of a given structure from those of a reference structure.

$$RMSd = min\left(\sqrt{\sum_{i=1}^{N}(\overrightarrow{x_i} - \overrightarrow{y_i})^2}\right) \tag{2.14}$$

where v_i and w_i are the coordinates of each of the N equivalent atoms in the structure of interest and the reference structure. It is often used to perform translations and rotations of one structure with respect to another, which minimize the RMSd with a simple least squares fitting algorithm, in order to superimpose the two. It can also be used as a quantitative measure of the similarity between simulation data and experimentally obtained structures, or to track the stability along a trajectory compared to a reference conformation, typically the average or starting one.

Alternatively, when a dynamical system fluctuates about some well-defined average position, the root-mean-squared fluctuations (RMSf) can be computed instead. In contrary to RMSd, RMSf gives the average fluctuations over time for each atom i:

$$RMSf_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\vec{x_i}(t) - \langle\vec{x_i}\rangle)^2} \qquad (2.15)$$

where T is the overall trajectory time, t is the selected time frame, $\vec{r_i}$ is the position of atom i after superposition on the reference structure, and $\langle\,\vec{r_i}\rangle$ is the average reference position over time T. Usually RMSf is measured separately for each residue, as DNA terminal residues tend to fluctuate more than others.

The Cartesian space average structure can be also used to analyze properties of DNAs using GRID-based approaches [59,60], such as the non-bonded interaction energy between the DNA and different probes placed in thousands of points around the molecule.

In many cases, a trajectory represents transitions between a number of different states, instead of sampling small fluctuations around a single conformation of minimum energy. Clearly, averaging of such ensembles will provide a "transition state-like" structure, which will be not realistic of the most prevalent regions of the conformational space. In Cartesian coordinates, a solution is to divide the configurational space into several clusters of structures that can define reference states. This information can be obtained from two-dimensional cross-RMSd plots, which computes the RMSd deviation of each snapshot with respect to the remaining ones. Once a restricted number of different states have been identified by clustering techniques [61,62], the configurational space can be seen as a combination of those MD-averaged reference states.

$$\{X_i\} \approx \sum_k P_k^t X_k^{av} \qquad (2.16)$$

where $\{X_i\}_t$ stands for the global configurational space sampled along the trajectory for time t, X_k^{av} stands for the MD-averaged conformation obtained from snapshots pertaining to cluster k, and P_k^t is the time-population of cluster k.

Analysis in Helical Space.

Nevertheless, the use of Cartesian coordinates can produce structures with distorted internal geometries. This problem can be alleviated by performing the averaging in a set of internal coordinates, which is especially feasible for standard nucleic acids, where the internal coordinates chosen are the helical parameters. Replacing atomic positions with rigid body translational and rotational parameters additionally allows a dramatic reduction in the degrees of freedom of the system. These parameters have been standardized by definitions set in the Tsukuba convention [63] and generally direct comparisons can be made between analyses performed by the different excellent programs such as Curves, 3DNA or NewHelix. Throughout this work we use the internal coordinate set as defined by the Curves+ software [64], which provides information on the helical parameters, the helical groove geometry and backbone conformation (described in terms of bond torsions and pseudorotational parameters).

A reference frame as defined in the Tsukuba convention [63] is attached to each base starting from the atoms involved in the glycosidic bond, defining the base plane by three atoms, N9, C1′ and C4 for purines and N1, C1′ and C2 for pyrimidines. Base pair reference frames and mid-frames of a base pair step are also constructed following the same scheme. Once the reference frames are constructed, the intra- and inter-base pair helical parameters are obtained by calculating rigid body transformations that map one reference frame into another, separated further into rotations and translations (three rotational parameters and three translational parameters, both intra- and inter-base pair – See Chapter 1 for details). The helical axis is calculated from the screw axes, which link successive base pair reference frames, and therefore can be curvilinear. Groove width measurements are based on distance between spline curves through the phosphorus atoms, reduced by 2 × 2.9 A ˚ to allow for the size of the phosphate groups. Groove depths involve the long axis of the base pairs and are reduced by 3.5 A ˚ to allow for the half-width of the base pairs.

4.2 Dynamic Information

DNA Dynamics.

A properly sampled MD trajectory of adequate length allows the extraction of both time-dependent and time-independent dynamic information. Time-independent dynamical information includes equilibrium distributions of helical parameters in one or more dimensions, together with their first and second moments (means and covariances), flexibility properties, principal component analysis and essential dynamics. Time-dependent information can consist of transition rates and residence times, correlations and causation between different transition motions.

In a first approximation, extracting average values and standard deviations of helical parameters will give a sense of the sequence-dependent conformational space available for specific DNA structures. MD simulations can be used to provide equilibrium samplings from which means and covariance matrices in helical space can be determined. Further, from the DNA representation in reduced helical coordinates, stiffness constants associated to deformation in helical space can be derived, as first proposed by Olson and Lankas groups [65–67]. In this representation, the inversion of covariance matrices provides the corresponding harmonic stiffness matrices for helical DNA deformations. This allows for a description of the dynamics of long DNA fragments with a potential based on elastic deformation energies from stiffness matrices:

$$K = kTC^{-1} \tag{2.17}$$

where C is the stiffness matrix with elements Kij.

$$E_{def} = \sum_i \frac{K_{ii}}{2}(X_i - X_{i0})^2 + \sum_{i\neq j} \frac{K_{ij}}{2}(X_i - X_{i0})(X_j - X_{j0}) \tag{2.18}$$

where Kij are force-constant and Xi and Xi0 are helical coordinates and their equilibrium values. Thus, DNA dynamic behavior can be modeled in the harmonic approximation, assuming small fluctuation around helicoidal coordinates minima. . Early elastic models rely on the use of a nearest neighbors representation of DNA, which was proven to not be accurate enough, forcing the use of tetramer-adjusted parameters [42,43,68–71]. Recent studies have also raised concerns on the use of harmonic approach [45,50,72,73], as DNA samples different conformational substates. Careful characterization of these substates at the tetramer level has been extensively performed in the present thesis.

Solvent Dynamics.

Analysis of ion distribution around DNA demonstrated that dependable sequence specific coordination studies should be based on simulations in the microsecond time scale, since ion convergence and ergodicity is only achieved after hundreds of nanoseconds [73,74]. It is therefore only recently that routine simulations can capture converged ion distributions. Another major milestone in the theoretical analysis of ion-DNA interactions was the development of an unambiguous method for the analysis of ion distributions in the curvilinear helicoidal frame given by the DNA axis [74]. The method uses the natural coordinate system for DNA, namely its helical axis and the curvilinear helicoidal coordinates (CHC) to determine the positions of solvent molecules around DNA. This allows the calculation of average ion populations that do not suffer from the DNA conformational fluctuations and removes the need for atomic references when investigating time-dependent information such as residence times (See Figure 2.3).

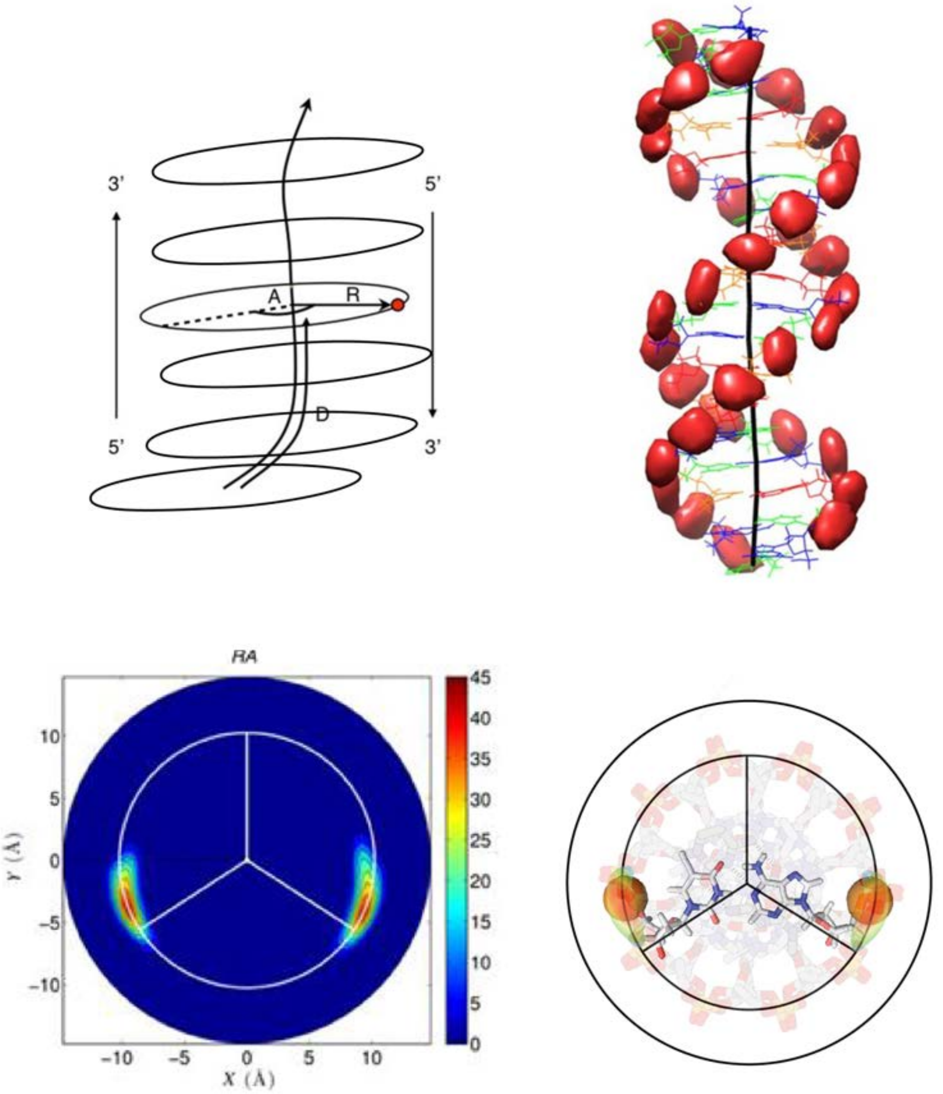


Figure 3 Schematic view of the curvilinear helicoidal coordinates (CHC) and phosphorus positions calculated in the CHC and mapped back into Cartesian space using the average helical axis of the oligomer.

Essential Dynamics.

An extremely powerful approach to describe global DNA flexibility is through its Essential Dynamics (ED) [75,76]. Following the principal component analysis (PCA) method, the approach determines the natural motions of a structure, that is, those motions that explain most of the variance detected during the trajectory. Practically, the method implies extracting the ensemble of conformations from an MD simulation, calculating the covariance matrix of the atomic movements, which is then diagonalized to yield a set of 3N-6 eigenvectors and their associated set of 3N-6 eigenvalues. The eigenvectors represent the nature of the essential motions, while the eigenvalues represent the percentage of variance explained by each eigenvector. Harmonic deformation constants along the essential modes can easily be derived from the eigenvalues (λi) with:

$$K_i = \frac{k_B T}{\lambda_i} \tag{2.19}$$

where the index i stands for an essential movement, k_B is Boltzmann factor, T is the absolute temperature and λ is the eigenvalue associated to deformation i in distance^2 units. The eigenvalues represent the amount of displacement along a mode expected at a given temperature. Note that once the force constant is known, the deformation energy along the essential mode i can be easily determined from eq 2.

$$E = \sum_{i=1}^{m} \frac{k_B T}{\lambda_i} \Delta X_i{}^2 \tag{2.20}$$

where ΔX_i is a Cartesian deformation along the eigenvector i.

An ED representation permits to reduce the flexibility of the DNA to a set of harmonic potentials applied on essential deformation modes. This opens the possibility to use essential deformation movements to perform MC or MD simulations on these essential modes. This approach is not expected to yield conformations very different to those sampled during the atomistic trajectory from which essential movements were determined, so a mere re-run of a simulation in reciprocal space is not very helpful. However, in a properly designed approach, this massive amount of information can be used to predict the dynamics of related sequences, or those of large DNA polymers of different sequences without the need to perform thousands of additional simulations.

Entropy Estimation.

Entropic factors are arguably the most difficult quatitity to estimate from molecular dynamics simulations, since they do not appear explicitly in the formulation of the force field. However, it has been shown [77–80] that they can be very important in determining affinity and selectivity of binding events. Interactions can modify the DNA freedom and accordingly its intramolecular entropy, but in most cases the largest entropic term in most cases relates to the reorganisation of solvent that accompanies the binding process. The expulsion of ordered water molecules from the binding interface are entropically highly favorable processes.

The intramolecular entropy change related to binding can be decomposed in two terms: the loss of rotational and translational degrees of freedom that accompany binding, ΔS_{r+t}, and the configurational entropy of the molecules' internal degrees of freedom, ΔS_{conf}. The calculation of ΔS_{conf} can be done directly from MD data, as proposed by Schlitter [81] (eqn. 3) and then by

Andricioaei and Karplus [82] (eqn. 4). Both require the calculation of the eigenvalues, ω, of the mass-weighted coordinate covariance matrix from the simulation:

$$S \approx \frac{1}{2}k\sum_i ln\left(1+\frac{e^2}{\alpha_i{}^2}\right) \tag{2.21}$$

$$S \approx \frac{1}{2}k\sum_i \frac{\alpha_i}{r^{\alpha_i}-1} - ln(1-e^{-\alpha_i}) \tag{2.22}$$

where
$$\alpha_i = \frac{\hbar w_i}{kT} \tag{2.23}$$

and the sum is over all non-trivial vibrations. Although there are slight differences in the derivation of these two methods, in practice they give very similar results.

In order for the diagonalisation procedure to generate (3N-6) 'true' eigenvalues – after the removal of the translational and rotational degrees of freedom – it theoretically requires that the covariance matrix be constructed from the analysis of at least (3N-6) independent snapshots. The definition of the time step between two independent snapshot is not always straightforward and depends on the system of interest, but what is generally observed is that the calculated entropy rises with trajectory length until it stabilizes at a system-dependent time scale. This makes sense intuitively, as any system will keep accessing unsampled areas of the configurational space with longer simulation time, until it samples it completely. The empirical relationship for this dependency has been formulated by Harris' group [83], as:

$$S(t) = S_{inf} - \alpha \cdot t^{-n} \tag{2.24}$$

where S_{inf} is the entropy for a simulation of infinite length, and α and n are fitting parameters.

In summary, a full understanding of what drives biological processes, such as molecular recognition, requires modelling methods that can probe both the enthalpic and entropic components of the system. The relationship between structure and enthalpy is quite straightforward to grasp (although high accuracy absolute values are only achieved at significant computational cost), but the relationship between entropy and dynamics can be less obvious and requires high-quality simulation data for quantitation.

Bibliography Chapter II

1. Lifson S, Warshel A. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and *n* - Alkane Molecules. *J Chem Phys* 1968;**49**:5116–29.

2. Rahman A. Correlations in the Motion of Atoms in Liquid Argon. *Phys Rev* 1964;**136**:A405–11.

3. Alder BJ, Wainwright TE. Studies in Molecular Dynamics. I. General Method. *J Chem Phys* 1959;**31**:459–66.

4. Lim T-C. The Relationship between Lennard-Jones (12-6) and Morse Potential Functions. *Zeitschrift für Naturforsch A* 2003;**58**:615–7.

5. Galliéro G, Boned C, Baylaucq A *et al.* Molecular dynamics comparative study of Lennard-Jones α -6 and exponential α -6 potentials: Application to real simple fluids (viscosity and pressure). *Phys Rev E* 2006;**73**:61201.

6. Migliorati V, Serva A, Terenzio FM *et al.* Development of Lennard-Jones and Buckingham Potentials for Lanthanoid Ions in Water. *Inorg Chem* 2017;**56**:6214–24.

7. Arieh Warshel *, Mitsunori Kato and, Pisliakov A V. Polarizable Force Fields: History, Test Cases, and Prospects. 2007, DOI: 10.1021/CT700127W.

8. Luque FJ, Dehez F, Chipot C *et al.* Polarization effects in molecular interactions. *Wiley Interdiscip Rev Comput Mol Sci* 2011;**1**:844–54.

9. Saito M. Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation. *J Chem Phys* 1994;**101**:4055–61.

10. Darden T, York D, Pedersen L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J Chem Phys* 1993;**98**:10089–92.

11. Ryckaert J, Ryckaert J, Ciccotti G *et al.* Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput PHYS* 1977:327--341.

12. Hess B, Hess B, Bekker H *et al.* LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput CHEM* 1997;**18**:18--1463.

13. Hopkins CW, Le Grand S, Walker RC *et al.* Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* 2015;**11**:1864–74.

14. Feenstra KA, Hess B, Berendsen HJC. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* 1999;**20**:786–98.

15. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys* 2007;**126**:14101.

16. Nosé S, Shuichi. A unified formulation of the constant temperature

molecular dynamics methods. *J Chem Phys* 1984;**81**:511–9.

17. Hoover WG. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* 1985;**31**:1695–7.

18. Berendsen HJC, Postma JPM, van Gunsteren WF *et al.* Molecular dynamics with coupling to an external bath. *J Chem Phys* 1984;**81**:3684–90.

19. Brünger A, Brooks CL, Karplus M. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem Phys Lett* 1984;**105**:495–500.

20. Grønbech-Jensen N, Farago O. A simple and effective Verlet-type algorithm for simulating Langevin dynamics. *Mol Phys* 2013;**111**:983–91.

21. Parrinello M, Rahman A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys Rev Lett* 1980;**45**:1196–9.

22. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 1981;**52**:7182–90.

23. Jorgensen WL, Chandrasekhar J, Madura JD *et al.* Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;**79**:926–35.

24. Berendsen HJC, Grigera JR, Straatsma TP *et al.* The missing term in effective pair potentials. *J Phys Chem* 1987;**91**:6269–71.

25. Izadi S, Anandakrishnan R, Onufriev A V. Building Water Models: A Different Approach. *J Phys Chem Lett* 2014;**5**:3863–71.

26. Cornell WD, Cieplak P, Bayly CI *et al.* Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J Am Chem Soc* 1993;**115**:9620–31.

27. Singh UC, Kollman PA. An approach to computing electrostatic charges for molecules. *J Comput Chem* 1984;**5**:129–45.

28. Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 1988;**110**:1657–66.

29. Matczak P, Matczak, Piotr. A Test of Various Partial Atomic Charge Models for Computations on Diheteroaryl Ketones and Thioketones. *Computation* 2016;**4**:3.

30. Srinivasan J, Withka JM, Beveridge DL. Molecular dynamics of an in vacuo model of duplex d(CGCGAATTCGCG) in the B-form based on the amber 3.0 force field. *Biophys J* 1990;**58**:533–47.

31. Seibel GL, Singh UC, Kollman PA. A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proc Natl Acad Sci U S A* 1985;**82**:6537–40.

32. Tidor B, Irikura KK, Brooks BR *et al.* Dynamics of DNA Oligomers. *J*

Biomol Struct Dyn 1983;**1**:231–52.

33. Levitt M. Computer simulation of DNA double-helix dynamics. *Cold Spring Harb Symp Quant Biol* 1983;**47 Pt 1**:251–62.

34. Cornell WD, Cieplak P, Bayly CI *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* 1995;**117**:5179–97.

35. MacKerell AD, Wiorkiewicz-Kuczera J, Karplus M. An all-atom empirical energy function for the simulation of nucleic acids. *J Am Chem Soc* 1995;**117**:11946–75.

36. MacKerell AD, Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* 2000;**56**:257–65.

37. Cheatham TE, Cieplak P, Kollman PA. A Modified Version of the Cornell *et al.* Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J Biomol Struct Dyn* 1999;**16**:845–62.

38. Pérez A, Marchán I, Svozil D *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 2007;**92**:3817–29.

39. Dixit SB, Beveridge DL, Case DA *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* 2005;**89**:3721–40.

40. Beveridge DL, Barreiro G, Suzie Byun K *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* 2004;**87**:3799–813.

41. Pérez A, Luque FJ, Orozco M. Frontiers in Molecular Dynamics Simulations of DNA. *Acc Chem Res* 2012;**45**:196–205.

42. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

43. Lavery R, Zakrzewska K, Beveridge D *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* 2010;**38**:299–313.

44. Lane AN, Chaires JB, Gray RD *et al.* Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res* 2008;**36**:5482–515.

45. Dršata T, Pérez A, Orozco M *et al.* Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J Chem Theory Comput* 2013;**9**:707–21.

46. Heddi B, Foloppe N, Oguey C *et al.* Importance of Accurate DNA

Structures in Solution: The Jun–Fos Model. *J Mol Biol* 2008;**382**:956–70.

47. Pérez A, Lankas F, Luque FJ *et al.* Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res* 2008;**36**:2379–94.

48. Fadrná E, Špačková N, Sarzyñska J *et al.* Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J Chem Theory Comput* 2009;**5**:2514–30.

49. Krepl M, Zgarbová M, Stadlbauer P *et al.* 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* 2012;**8**:2506–20.

50. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.

51. Zgarbová M, Luque FJ, Šponer J *et al.* Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J Chem Theory Comput* 2013;**9**:2339–54.

52. Zgarbová M, Šponer J, Otyepka M *et al.* Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J Chem Theory Comput* 2015;**11**:5723–36.

53. Zgarbová M, Otyepka M, Šponer J *et al.* Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput* 2011;**7**:2886–902.

54. Banáš P, Mládek A, Otyepka M *et al.* Can We Accurately Describe the Structure of Adenine Tracts in B-DNA? Reference Quantum-Chemical Computations Reveal Overstabilization of Stacking by Molecular Mechanics. *J Chem Theory Comput* 2012;**8**:2448–60.

55. Morgado CA, Jurečka P, Svozil D *et al.* Balance of Attraction and Repulsion in Nucleic-Acid Base Stacking: CCSD(T)/Complete-Basis-Set-Limit Calculations on Uracil Dimer and a Comparison with the Force-Field Description. *J Chem Theory Comput* 2009;**5**:1524–44.

56. Chen AA, Garcia AE. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc Natl Acad Sci* 2013;**110**:16820–5.

57. Tan D, Piana S, Dirks RM *et al.* RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc Natl Acad Sci U S A* 2018;**115**:E1346–55.

58. Yang C, Kim E, Pak Y. Free energy landscape and transition pathways from Watson–Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Res* 2015;**43**:7769–78.

59. Gelpí JL, Kalko SG, Barril X *et al.* Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins* 2001;**45**:428–37.

60. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985;**28**:849–57.

61. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. *J Comput Chem* 2004;**25**:865–71.

62. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci U S A* 1998;**95**:11158–62.

63. Olson WK, Bansal M, Burley SK *et al.* 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* 2001;**313**:229–37.

64. Blanchet C, Pasi M, Zakrzewska K *et al.* CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* 2011;**39**:W68-73.

65. Olson WK, Gorin AA, Lu XJ *et al.* DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 1998;**95**:11163–8.

66. Lankaš F, Šponer J, Hobza P *et al.* Sequence-dependent elastic properties of DNA 1 1Edited by I. Tinoco. *J Mol Biol* 2000;**299**:695–709.

67. Olson WK, Zhurkin VB. Modeling DNA deformations. *Curr Opin Struct Biol* 2000;**10**:286–97.

68. Imeddourene A Ben, Xu X, Zargarian L *et al.* The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res* 2016;**44**:3432–47.

69. Ben Imeddourene A, Elbahnsi A, Guéroult M *et al.* Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. MacKerell A (ed.). *PLOS Comput Biol* 2015;**11**:e1004631.

70. Tian Y, Kayatta M, Shultis K *et al.* ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J Phys Chem B* 2009;**113**:2596–603.

71. Zgarbová M, Jurečka P, Lankaš F *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* 2017;**57**:275–87.

72. Maehigashi T, Hsiao C, Woods KK *et al.* B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res* 2012;**40**:3714–22.

73. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.

74. Pasi M, Maddocks JH, Lavery R. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res* 2015;**43**:2412–23.

75. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. *Proteins Struct Funct Genet* 1993;**17**:412–25.

76. Alberto Pérez †,‡, José Ramón Blas †, Manuel Rueda †,§ *et al.* Exploring the Essential Dynamics of B-DNA. 2005, DOI: 10.1021/CT050051S.

77. Jayaram B, McConnell K, Dixit SB *et al.* Free-energy component analysis of 40 protein-DNA complexes: A consensus view on the thermodynamics of binding at the molecular level. *J Comput Chem* 2002;**23**:1–14.

78. Dragan AI, Klass J, Read C *et al.* DNA binding of a non-sequence-specific HMG-D protein is entropy driven with a substantial non-electrostatic contribution. *J Mol Biol* 2003;**331**:795–813.

79. Liang F, Cho BP. Enthalpy-entropy contribution to carcinogen-induced DNA conformational heterogeneity. *Biochemistry* 2010;**49**:259–66.

80. Haq I, Ladbury JE, Chowdhry BZ *et al.* Specific binding of hoechst 33258 to the d(CGCAAATTTGCG)2 duplex: calorimetric and spectroscopic studies. *J Mol Biol* 1997;**271**:244–57.

81. Schlitter J, Jürgen. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem Phys Lett* 1993;**215**:617–21.

82. Andricioaei I, Karplus M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J Chem Phys* 2001;**115**:6289–92.

83. Harris SA, Gavathiotis E, Searle MS *et al.* Cooperativity in drug-DNA recognition: a molecular dynamics study. *J Am Chem Soc* 2001;**123**:12658–63.

CHAPTER III | ParmBSC1

Our group has established their expertise in deciphering the physical properties of nucleic acids with the use of MD simulations over more than a decade, with significant contributions to this field [1–11]. Part of their notable efforts was the development of parmbsc0 [11], a force field for nucleic acids that, until rather recently, had provided the gold standard in MD simulations of DNA.

At the beginning of this thesis, as the sampling problem and thus convergence were being steadily improved [5,6,9,12,13], a number of inconsistencies in the equilibrium conformations obtained with parmbsc0 had started to be reported, both from within the group and outside [14–20]. Notable among them were slight but significant deviations of helical parameter averages from experimental values (especially in twist and roll), coming from the underestimation of BI/BII equilibrium (stemming from ε/ζ coupled distribution) [15,20], excessive terminal base fraying [15] and difficulties in accessing exotic DNA structures which was related to a stiff χ torsion [17].

I experienced such biases myself, and decided to join the already on-going effort in the group of reparameterizing the parmbsc0 force field with regard to the sugar puckering, ε/ζ and χ torsions, using high-level QM calculations both in gas phase and solution. The final version of this reparametrization was named parmBSC1.

In the meantime, specific-purpose corrections to the parmbsc0 were already starting to appear, mainly from the Zgarbová et al. [21,19,22], with modifications targeting the χ distribution (χOL4 – for simulation of DNA quadruplexes), the ε/ζ representation (ε/ζOL1) and finally the β profile as well (βOL1 – for improved Z-DNA structures). The last generation of these force fields incorporated all the previous OL corrections for DNA (OL15) and was developed in parallel to the parmBSC1 force-field.
By now, several review papers [23,24] have assessed the accuracy of parmBSC1 against other proposed force fields and have concluded that, along with the similarly performing OL15, it should be the new standard in DNA simulations (see Figure 3.1).

Our own extensive testing of the new parameter set was done over the course of four years before publication. We aimed to thoroughly validate not only the correct reproduction of experimental structures, but also against experimental observables, such as NOEs and RDCs from NMR experiments, persistence lengths and transition rates in different solvent types. Due to the fact that experimentally obtained structures have in many cases unclear accuracy, it was very significant that the MD ensembles produced with our force field faithfully reproduced direct experimental observables. Other dynamic properties were assessed through folding and unfolding experiments and conformational transitions (such as A- to B-DNA). We also tested its applicability not only to canonical DNA, but also various non-canonical forms, other exotic DNA configurations like triplexes and quadruplexes, to DNA in complexes with proteins and ligands, and in a number of solvent environments. This meant a cumulative simulation time of ~140 μs, and over 100 different simulated systems.

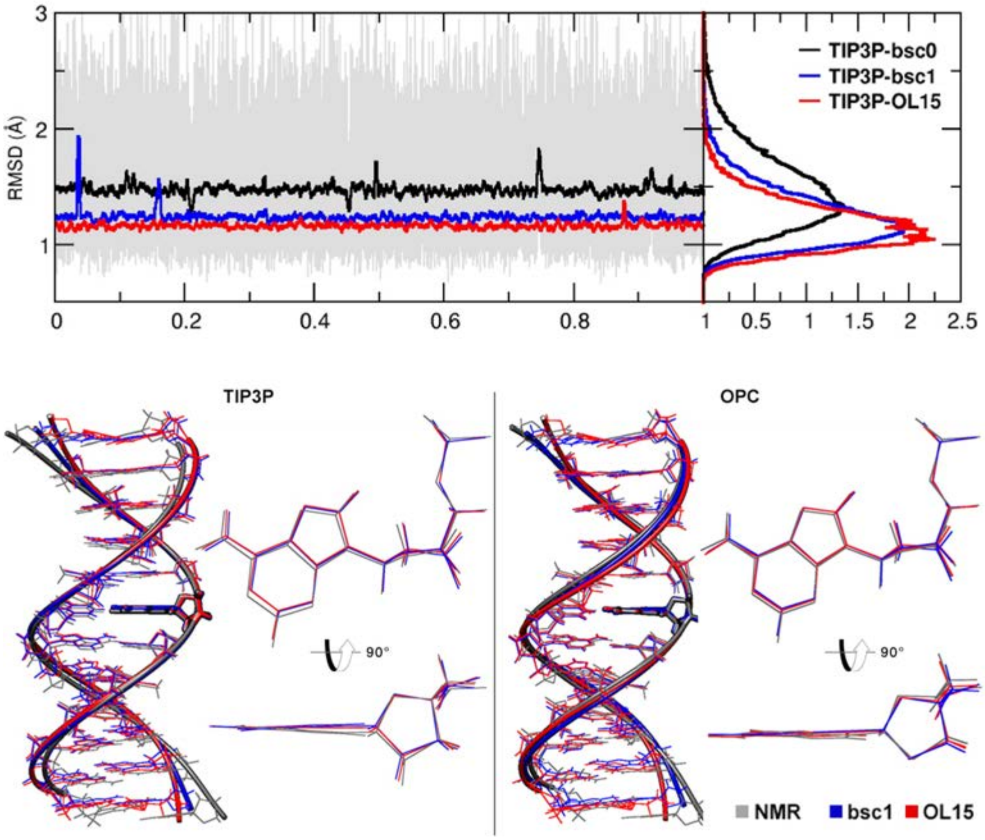


Figure 1 RMSd (top) and average structures (bottom) calculated from 1 ms aggregated trajectories: comparison between parmbsc0, parmbsc1, OL15 and NMR average structure (adapted from [23])

First, by correcting the coupled ε/ζ profile, we aimed not only to improve BI/BII equilibrium representation (which is by definition defined by these two dihedrals, with trans/gauche- representing the canonical BI state, while BII is given by gauche-/trans of ε/ζ), but we justly assumed this correction would rectify the twist and roll distributions, which are tightly correlated to the backbone state. Validation of the improved ε/ζ distribution revealed that with this correction known digressions from normality of certain base pair steps in their helical parameters were also correctly reproduced.

The glycosidic torsion correction accounted for the anti/syn equilibrium of the base orientation and allowed for accurate simulations of non-canonical DNA structures, and also reduced end-base fraying. Trajectories collected for testing this modification showed structures with better-conserved terminal hydrogen bonding and lowed RMSd of terminal bases with respect to experimental data. Figure 3.2 illustrates the excellent agreement of the final parmbsc1 ε/ζ and χ dihedral energy profiles with quantum mechanical data compared to the previous parmbsc0 force field.

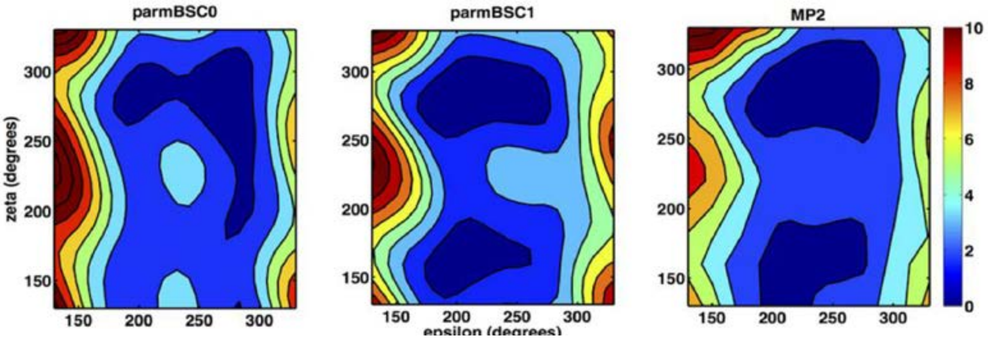


Figure 2 Contour profiles of epsilon/zeta distributions from QM MP2 calculations (right) and PMF profiles using parmbsc0 (left) and parmbsc1 (middle) force fields.

Finally, as a consequence of changing these parameters, the puckering required a new refinement. We noticed this when simulations of a small duplex d(CGATCG)$_2$ incorporating the two corrections showed small imperfections with puckering profiles. With this reparametrization of puckering pseudorotation, we also got the opportunity to ameliorate the parmbsc0 bias towards East conformations.

This study is presented in the publication *Parmbsc1: a refined force field for DNA simulations* attached in the following pages. There is little doubt that the parmbsc1 force-field provides accurate insights into the richness and complexity of dynamics on the submillisecond time sclale. The level of agreement with experimental data of last generation empirical force fields (parmbsc1 and OL15), when properly used, is often within the accuracy of the experimental method, legitimizing the great potential of these virtual microscopes for providing valuable insights into important biological processes.

Publications from this Chapter:

Ivan Ivani, Pablo D. Dans, Agnes Noy, Alberto Perez, Ignacio Fausno, Adam Hospital, Jurgen Walther, Pau Andrio, Ramon Goni, Alexandra Balaceanu, Guillem Portella, Federica Basni, Josep Lluis Gelpi, Carlos Gonzalez, Michele Vendruscolo, Charles A. Laughton, Sarah A. Harris, David A. Case and Modesto Orozco; Parmbsc1: a refined force field for DNA simulations. Nature Methods, 13, 55-58, 2016.

PARMBSC1: A REFINED FORCE-FIELD FOR DNA SIMULATIONS

Ivan Ivani[1,2], Pablo D. Dans[1,2], Agnes Noy[3], Alberto Pérez[4], Ignacio Faustino[1,2], Adam Hospital[1,2], Jürgen Walther[1,2], Pau Andrio[2,5], Ramon Goñi[2,5], Alexandra Balaceanu[1,2], Guillem Portella[1,2,6], Federica Battistini[1,2], Josep Lluis Gelpí[2,7], Carlos González[8], Michele Vendruscolo[6], Charles A. Laughton[9], Sarah A. Harris[3], David A. Case[10],and Modesto Orozco[1,2,7]

[1]Institute for Research in Biomedicine (IRB Barcelona), the Barcelona Institute of Science and Technology, Barcelona, Spain.

[2]Joint BSC-IRB Research Program in Computational Biology, Barcelona, Spain.

[3]School of Physics and Astronomy, University of Leeds, Leeds, UK.

[4]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, USA.

[5]Barcelona Supercomputing Center, Barcelona, Spain.

[6]Department of Chemistry, University of Cambridge, Cambridge, UK.

[7]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain.

[8]Instituto de Química Física "Rocasolano", CSIC, Madrid, Spain.

[9]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK.

[10]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, USA.

Corresponding author: Modesto Orozco, modesto.orozco@irbbarcelona.org

Editorial summary:
Parmbsc1, a force-field for DNA simulations, is presented. It has been broadly tested on nearly 100 DNA systems and overcomes simulation artifacts that affect previous force-fields.

We present parmbsc1, a force-field for DNA atomistic simulation, which has been parameterized from high-level quantum mechanical data and tested for nearly 100 systems (representing a total simulation time of ~140 μs) covering most of DNA structural space. Parmbsc1 provides high quality results in diverse systems. Parameters and trajectories are available at http://mmb.irbbarcelona.org/ParmBSC1/.

The Force-field, the energy functional used to describe the dependence between system conformation and energy, is the core of any classical simulation including molecular dynamics (MD). Its development is tightly connected to the extension of simulation time scales. As MD trajectories are extended to longer timescales, errors previously undetected in short simulations emerge, creating the need to improve the force-fields[1]. For example, AMBER (Assisted Model Building with Energy Refinement) parm94-99 was the most used force-field in DNA simulations until multi-nanosecond simulations revealed severe artifacts[2,3], thus fueling the development of parmbsc0[4], which, in turn, started to show deviations from experimental data in the μsec regime (for example an underestimation of the twist, deviations in sugar puckering, biases in ε and ζ torsions, excessive terminal fraying[2,5], and severe problems in representing certain non-canonical DNAs[1,6]). Various force-field modifications have been proposed to address these problems, such as the Olomouc (OL)-ones[5,6] designed to reproduce specific forms of DNA. While these and other tailor-made modifications are useful, there is an urgent need for a new general-purpose AMBER force-field for DNA simulations to complement recent advances in the CHARMM (Chemistry at HARvard Macromolecular Mechanics) family of force-fields (Online Methods). We designed theparmbsc1 force-field presented here to solve these needs, with the aim of creating a general-purpose force-field for DNA simulations. We demonstrate its performance by testing its ability to simulate a wide variety of DNA systems (**Supplementary Table 1**).

Parmbsc1 shows good ability to fit quantum mechanical (QM) data (QM data fitting section in **Supplementary Discussion**), improving on previous force-field results (Online Methods, **Supplementary Table 2**). We first tested QM-derived parameters on the Drew-Dickerson dodecamer (DDD), a well-studied DNA structure[2,7], typically used as benchmark in force-field developments.Parmbsc1 trajectories sampled a stable B-type duplex that remained close to the experimental structures (**Fig. 1** and **Supplementary Table 2**), preserving hydrogen bonds and helical characteristics, even at the terminal base pairs, where fraying artifacts are common using other force-fields[2,8] (see Online Methods and **Supplementary Discussion**). The average sequence-dependent helical parameters (**Fig. 1** and **Supplementary Figs. 1** and **2**), and BI/BII conformational preferences (**Supplementary Table 2** and **Supplementary Fig. 3**) matched experimental values (for the comparisons with estimates obtained with other force-fields see Online Methods). Furthermore, parmbsc1 reproduced residual dipolar couplings (Q-factor = 0.3) and NOEs (Nuclear Overhauser Effect; only two violations), yielding success metrics similar to those obtained in the NMR (Nuclear Magnetic Resonance)-refined structures (**Supplementary Table 3**).

We next evaluated the ability of parmbsc1 to represent sequence-dependent structural features from simulations on 28 B-DNA duplexes (**Supplementary Table 4**). The agreement between simulation and experiment was excellent (Root Mean Square deviation (RMSd) per base pair of 0.1 or 0.2 Å). Almost no artifacts arising from terminal fraying were present, and the average helical parameters (twist and roll from simulations: 33.9 º and 2.5 º respectively), matched values from the analysis of the PDB (33.6 º and 2.9 º)[9]. Moreover, parmbsc1 was able to reproduce the unique properties of A-tract sequences[10] (**Supplementary Figs. 4–6**), and capture sequence-dependent structural variability (**Supplementary Fig. 7**). We also studied longer duplexes (up to 56 bp) to ensure that a possible accumulation of small errors given by the force-field did not compromise the description of the DNA, finding excellent results (**Supplementary Table 5**). The expected spontaneous curvature was clearly visible in both static and dynamical descriptors, demonstrating that parmbsc1 trajectories were able to capture complex polymeric effects (**Supplementary Table 5**).

We also explored the ability of parmbsc1 to represent unusual DNAs, such as a Holliday junction, a complex duplex-quadruplex structure which was fully preserved in µsec-long trajectories (**Supplementary Figs. 8** and **9**); or the Z-DNA, a *levo* duplex containing nucleotides in *syn*, for which parmbsc1 not only provided stable trajectories (**Fig. 2a**), but also reproduced the experimentally known salt dependence, confirming that the conformation is stable only at high (4 M) salt concentration[11]. For Hoogsteen-DNA (H-DNA), simulations with parmbsc1 showed a stable duplex for more than 150 ns (**Fig. 2b**), and severe distortions in longer simulation periods (**Supplementary Fig. 10**), as expected from its metastable nature[12]. We obtained equivalent results for another metastable structure, the parallel poly-d(AT) DNA (**Supplementary Fig. 11**)[13]. Parmbsc1 simulations not only reproduced the known structure of parallel d(T-A·T) and d(G-G·C) triplexes (**Figs. 2c,d**), but also showed correctly that the equivalent antiparallel structures are unstable in normal conditions (**Fig. 2e**)[14]. Finally, parmbsc1 was able to reproduce experimental structures of both parallel and antiparallel DNA quadruplexes with RMSd < 2 Å (**Figs. 2f,g**).

We explored also the ability of parmbsc1 to reproduce the complex conformation of hairpins and loops, exceptionally challenging structures for force-fields[15]. We performed µs simulations of the d(GCGAAGC) hairpin (PDB: 1PQT), the 4T-tetraloop in *Oxytricha nova* quadruplex d($G_4T_4G_4$)$_2$ (OxyQ; PDB: 1JRN), and the junction loops in the human telomeric quadruplex (HTQ; PDB: 1KF1). Parmbsc1 provided excellent representations (RMSd around 1 Å) of the d(GCGAAGC) hairpin (**Fig. 2h**), and of the OxyQ quadruplex (**Fig. 2i**). For the very challenging HTQ structure, parmbsc1 maintained the stem structure 20 times longer than in previous simulations[15], and recognized the large flexibility of the loops in the absence of the lattice-contacts (**Supplementary Fig. 12**), showing that, as predicted[16], not only the crystal, but also other loop conformations were sampled (**Fig. 2j**).

As an additional critical test of the new force-field we predicted NMR observables from parmbsc1 trajectories (Online Methods). We obtained equivalent NOE violation statistics to those determined from NMR-derived ensembles (**Supplementary Tables 6** and **7**, and **Supplementary Fig. 13**). This agreement was maintained in *de novo* predictions, *i.e.* in those cases where NMR observables were collected in one of our laboratories after parmbsc1 development (**Supplementary Table 8**). Finally, it is worth noting that parmbsc1 trajectories reproduced the structure of DNA in crystal environments, yielding a RMSd between the simulated and crystal structures of only 0.7 Å, and average twist differences below one degree, improving on previous calculations (Online Methods and **Supplementary Figs. 14** and **15**).

In our final structural test we explored the ability of parmbsc1 to reproduce the conformation of DNA in complex with other molecules. We studied four diverse protein DNA complexes (PDB: 1TRO, 2DGC, 3JXC and 1KX5), and two prototypical drug DNA complexes. In all cases, we found excellent agreement (RMSd for DNA around 2–3 Å in protein-DNA complexes, and 1–2 Å in drug-DNA complexes) with experiments (**Fig. 3** and **Supplementary Figs. 16** and **17**).

A force-field should not only reproduce the structure of DNA, but also its mechanical properties[1]. To evaluate the performance of parmbsc1 we firstly evaluated the µs-scale dynamics of the central 10 base pairs of the DDD. The agreement between parmbsc0 and parmbsc1 normal modes and entropy estimates (Online Methods and **Supplementary Table 9**) demonstrated that parmbsc1 does not "freeze" the DNA structure, a risk for a force-field reproducing well average properties. This was further confirmed by the ability of parmbsc1 to reproduce the DNA dielectric constant (8.0 ± 0.3 for DDD *versus* the experimental estimate of 8.5 ± 1.4; see **Supplementary Fig. 18**), and also the cooperative binding (around 0.7 kcal mol^{-1}) of Hoechst 33258 to DNA. We then computed the helical stiffness matrices for the ten unique base pair steps[17,18]. Parmbsc1 values were intermediate between parmbsc0 and CHARMM27 stiffness parameters[18], and substantially smaller than those suggested by Olson and coworkers[17] (**Supplementary Table 10** and **Supplementary Fig. 19**); the dependence of the stiffness parameters on sequence were similar for parmbsc1 and parmbsc0[17].

The persistence length, the torsional, and the stretching modules were obtained from simulations of long (up to 56 bp) duplexes (Online Methods). Parmbsc1 predicted persistence lengths in the range of 40–57 nm (**Supplementary Table 11**), close to the generally accepted value of 50 nm. The computed static persistence length, stretch and twist torsion modules were around 500 nm, 1,100–1,500 pN, and 50–100 nm respectively, also in agreement with experimental values (**Supplementary Table 11**). Finally, we explored the ability of parmbsc1 to describe relaxed and stressed DNA minicircles. We performed three 100 ns simulations of a 106-bp minicircle with ten turns (106t10), which should have zero superhelical density (σ =0) and therefore no denatured regions[19,20] (**Supplementary Fig. 20**). A kink was observed only in a single replica for one of the register angles, while in the remaining simulations the DNA remained intact (**Supplementary Fig. 20**). On the contrary, negatively supercoiled 100-bp (100t9; σ = –0.05) and 106-bp (106t9, σ = –0.10) minicircles formed distortions due to the superhelical stress, as previously reported experimentally using enzymes that digest single stranded DNA[19,20].

Having demonstrated the ability of parmbsc1 to describe stable and metastable DNA structures and DNA flexibility, we finally studied conformational transitions. Parmbsc1 reproduced the spontaneous A to B-form DNA transition in water, and the A form was found, as expected, to be stable in 200 ns control simulations in a 85% ethanol and 15% water mixture (**Supplementary Fig. 21**). Parmbsc1 also reproduced the unfolding of DNA d(GGCGGC)$_2$ in a 4 Molar pyridine solution (**Supplementary Fig. 21**), and the effective folding of d(GCGAAGC) in water (**Supplementary Fig. 22**), suggesting the ability to capture long-scale conformational changes in DNA.

Based on the wide series of tests we report, we conclude that parmbsc1 provides good representations of the static and dynamic properties of DNA. We anticipate that parmbsc1 will be a valuable reference force-field for atomistic DNA simulations under a diverse range of conditions.

METHODS

Methods and associated references are available in the online version of the paper.

ACKNOWLEDGEMENTS

MO thanks Spanish Ministry of Science (BIO2012-32868), the Catalan SGR, the Instituto Nacional de Bioinformática, and the European Research Council (ERC SimDNA) for support.MO is an ICREA academia researcher. MO thanks CPU-GPU time on MareNostrum-MinoTauro (BSC).CAL, SAH and AN thanks the UK HECBioSim Consortium for HPC time on ARCHER (Grant EP-L000253-1).AN was supported by the Biotechnology and Biological Sciences Research Council (BBSRC, grant number BB-I019294-1), and thanks ARC Leeds for computational resources. PDD is a PEDECIBA and SNI (ANII, Uruguay) researcher. DAC thanks C. Liu for assistance with the crystal simulation analysis.

AUTHOR CONTRIBUTION

II derived the parmbsc1 force-field parameter set. II, PDD, AN, AP, IF, AH, JW, AB, GP, FB, CAL, and SAH performed validation simulations. CG, MV, and GP validate results from NMR. CG did *de novo* NMR measures. DAC performed crystal MD simulations. RM, PA, AH, and JLG created the database infrastructure and web application. All authors contributed to the analysis of data. MO had the idea, directed the project, and wrote the manuscript which was improved by the rest of the authors.

FINANCIAL STATEMENT

The authors declare no competing financial interests.

REFERENCES

1. Pérez, A., Luque, F. J. & Orozco, M. *Acc. Chem. Res.***45,** 196–205 (2011).
2. Pérez, A., Luque, F. J. & Orozco, M. *J. Am. Chem. Soc.***129,** 14739–14745 (2007).
3. Varnai, P.& Zakrzewska, K. *Nucleic Acids Res.***32,** 4269-4280 (2004).
4. Pérez, A. *et al. Biophys. J.***92,** 3817–3829 (2007).
5. Zgarbová, M. *et al. J. Chem. Theory Comput.***9,** 2339–2354 (2013).
6. Krepl, M. *et al. J. Chem. Theory Comput.***8,** 2506–2520 (2012).
7. Wing, R. *et al. Nature***287,** 755–758 (1980).
8. Lavery, R. *et al. Nucleic Acids Res.***38,** 299–313 (2010).
9. Dans, P.D., Pérez, A., Faustino, I., Lavery, R. & Orozco, M. *Nucleic Acids Res.***40,** 10668–10678 (2012).
10. Lankaš, F., Špačková, N., Moakher, M., Enkhbayar, P. & Šponer, J. *Nucleic Acids Res.***38,** 3414–3422 (2010).

11. Thamann, T.J., Lord, R.C., Wang, A.H.J. & Rich, A. *Nucleic Acids Res.***9,** 5443–5458 (1981).

12. Abrescia, N.G.A., González, C., Gouyette, C. & Subirana, J.A. *Biochemistry***43,** 4092–4100 (2004).

13. Cubero, E., Luque, F.J. & Orozco, M. *J. Am. Chem. Soc.***123,** 12018–12025 (2001).

14. Soyfer, V.N. & Potaman, V.N. in*Triple-helical nucleic acids* 1st edn. (Springer - Verlag New York, 1996).

15. Fadrná, E. *et al. J. Chem. Theory Comput.***5,** 2514–2530 (2009).

16. Martín-Pintado, N. *et al. J. Am. Chem. Soc.***135,** 5344–5347 (2013).

17. Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M. & Zhurkin, V.B. *Proc. Natl. Acad. Sci.***95,** 11163–11168 (1998).

18. Pérez, A., Lankas, F., Luque, F.J. & Orozco, M. *Nucleic Acids Res.***36,** 2379–2394 (2008).

19. Moroz, J.D. & Nelson, P. *Proc. Natl. Acad. Sci.***94,** 14418–14422 (1997).

20. Du, Q., Kotlyar, A. & Vologodskii, A. *Nucleic Acids Res.***36,** 1120–1128 (2008).

FIGURE LEGENDS

Figure1|Analysis of the Drew-Dickerson dodecamer. (**a**) Visual comparison of MD average structure (brown) and NMR structure (PDB id: 1NAJ) (light blue) and X-ray structure (PDB id: 1BNA) (green). (**b**) RMSd of 1.2 µs trajectory of DDD compared with B-DNA (blue) and A-DNA (green) form (coming from the standard geometries derived from fiber diffraction, see Online Methods section Validation of MD simulations). (**c**) RMSd compared to experimental structures (with (dark) and without (light) ending base-pairs): X-ray (green) and NMR (blue). Linear fits of all RMSd curves are plotted on top. (**d**) Evolution of total number of hydrogen bonds formed between base pairs in the whole duplex. (**e**) Helical rotational parameters (twist, roll, and tilt) comparison of average values per base-pair step (standard deviations are shown by error bars) coming from NMR (cyan), X-ray (dark green), 1 µs parmbsc0 trajectory[2] (black) and 1.2 µs parmbsc1 trajectory (violet).

Figure2|Analysis of non-canonical DNA structures. (**a**) Comparison of Z-DNA (PDB id: 1I0T) simulations in neutralized conditions (green) and in 4 M solution of $Na^{+}Cl^{-}$ (blue). Structural comparisons at given time points are shown above the RMSd curves. (**b**) Simulation of anti-parallel H-DNA (PDB id: 2AF1) showing deviation of the structure over time (highlighted in red). RMSd of (**c**) parallel $d(T\text{-}A{\bullet}T)_{10}$, (**d**) parallel $d(G\text{-}G{\bullet}C)_{10}$,and (**e**) antiparallel $d(G\text{-}G{\bullet}C)_{10}$ triplexes. (**f**) Parallel (PDB id: 352D) and (**g**) anti-parallel (PDB id: 156D) quadruplex showing stable structures over time. (**h**) Structural stability of d(GCGAAGC) hairpin (PDB id: 1PQT) and (**i**) OxyQ quadruplex (PDB id: 1JRN) with ions, over time. (**j**) Human Telomeric Quadruplex (PDB id: 1KF1) with highlighted loops. RMSd of HTQ backbone, loop 1, loop 2 and loop 3 regions are shown below. In all panels, parmbsc1 (final, averaged or at a given trajectory point) structures (light blue; also green for Z-DNA) are overlapped over experimental structure (grey) for comparison. See **Supplementary Table 1** for information on the PDB structures.

Figure3|Analysis of DNA-protein complexes. Structural details of microsecond trajectories of four complexes with PDB id: 1TRO (**a**), 2DGC (**b**), 3JXC (**c**) and 1KX5 (**d**) (500 ns trajectory). Each plot shows overlap of the MD starting (red) and final (blue) structures, time dependent mass-weighted root mean square deviation (RMSD in Å) of all DNA (red) and protein (cyan) heavy atoms, and comparison of the values of rotational helical parameter roll (in degrees) at each base pair step calculated from the X-ray crystal structure (cyan) and averaged along the MD simulation (red line with the standard deviation envelope in light red). For clarity, in the 1KX5 plot of the roll value, the base pair steps are defined by the number of the position along the DNA strand and not by the base pair step name.

ONLINE METHODS

General parameterization strategy.

AMBER charges and van der Waals parameters for DNA are able to reproduce high-level QM data[21–23] and hydration free energies[24–26], as well as producing reasonable hydrogen bond stabilities[2, 21–23, 27] and complex properties such as sequence-dependent stabilities of duplex DNA[2, 28, 29]. Thus, we decided to keep the non-bonded parameters unaltered in this force-field revision, and focus our efforts in the parameterization of the backbone degrees of freedom: sugar puckering, glycosidic torsion, and ε and ζ rotations (taking the recently re-parameterized α and γ torsions from parmbsc0[4]). Parameterization of the different torsion angles (see below) was done from high-level QM calculations using the refined gas phase fitted parameters as initial guesses for the refinement of parameters in solution taken now as reference high level Self-Consistent Reaction Field (SCRF)-QM data. In cases where fitting of one force-field parameter requires the knowledge of another parameter for the optimization, an iterative procedure using parmbsc0 parameters in the first iteration was employed.

Quantum mechanical calculations.

Model compounds, shown in **Supplementary Fig. 23**, were first geometrically optimized at the B3LYP/6-31++G(d,p) level[30] from which single-point energies were calculated at the MP2/aug-cc-pVDZ level[31]. To reduce errors in the fitting, optimizations were done while selected backbone and sugar dihedral angles were constrained to typical values obtained from a survey of DNA crystal structures[9]. We obtained both vacuum and solvent profiles for all structures calculated. 3D profiles of ε and ζ were sampled with 10 º increment in the region of interest (ε = [175 º, 275 º], ζ = [220 º, 330 º]), and with 40 º increment in the rest of the profile. Profiles of χ were sampled with 15 º increment and profiles of sugar pucker by 10 º in the range of phase angles from 0 º to 180 º, and considering the four nucleosides. To increase the accuracy of the profiles, we performed CCSD(T)/complete basis set (CBS) calculations[32, 33] on key point along the Potential Energy Surface (for ε and ζ these points were B_I, B_{TRANS} and B_{II} states; for χ minima of *anti* and *syn* regions, and maximum between them; and minima of *North*, *East* and *South* conformations for the sugar pucker). These calculations were performed first by optimization at the MP2/aug-cc-pVDZ level, followed by single-point calculations at the MP2/aug-cc-pVXZ (X = Triplex and Quadruplex) levels. CBS energies were obtained by extrapolating to infinite basis set, from the scheme of Halkier *et al.*[32], and adding the correction term of the difference from CCSD(T) and MP2 with the 6-31+G(d) basis set. These high level points were introduced with increased weights in the global fitting (see below). All QM calculations were performed with Gaussian09 (http://www.gaussian.com).

Solvation corrections in QM calculations.

The solvent calculations were done at the single-point level using our version of the polarizable continuum model (PCM) from Miertus, Scrocco and Tomasi (MST)[34–40]. For comparison, test calculations were performed using Cramer and Truhlar SMD (Solvent Model based on Density) model[41], and the standard Integral Equation Formalism (IEF)-PCM[36] as implemented in the Gaussian09 package, obtaining very similar results (data not shown). Consequently, only MST values were used in this work.

Molecular mechanics and Potential of Mean Force calculations.

Molecular mechanics (MM) reference calculations of the QM-optimized structures *in vacuo* were obtained from MM single-point energy calculations using the AMBER 11 package (http://ww.ambermd.org). MM profiles in solution were recovered from potential of mean force (PMF) calculations created with umbrella sampling (US)[42] procedures in explicit solvent conditions (no restraints were used on any dihedrals out of the reaction coordinate in these calculations). US calculations were carried out with a weak biasing harmonic potential of 0.018 kcal mol^{-1} deg^{-2}. The resulting populations were integrated using the Weighted Histogram Analysis Method (WHAM, http://membrane.urmc.rochester.edu/content/wham).US calculations typically involve 40–100 windows, each consisting of 2–5 ns of equilibration and sampling times in the order of 1–2 ns. Simulation details in PMF-US calculations were the same as those outlined below in the validation of MD simulations section.

Force-field fitting.

The procedure of force-field fitting was similar to parmbsc0 parameterization process[4]. In order to avoid altering other torsional parameters of the general force-field, we introduced new atom types depending on the parameterization. For ε, ζ, and sugar pucker parameterization we assigned the atom type *CE* to C3' atom. For χ parameterization we assigned *C1* to the C8 atom of adenine and *C2* to the C6 atom of thymine, while keeping unchanged the atom types *CK* for guanine and *CM* for cytosine. Charges for model systems used in the parameterization were calculated from standard RESP methods mimicking the original amber parameterization. We used the standard torsions definition, *i.e.* ε = C4'-C3'-O3'-P, ζ = C3'-O3'-P-O5', χ = O4'-C1'-N9-C8 (for dA and dG) and χ = O4'-C1'-N1-C6 (for dC and dT). For sugar pucker parameterization we chose ν_1=O4'-C1'-C2'-C3', the δ backbone and the ν_2=C1'-C2'-C3'-C4' dihedrals, since they connect the two corrections: ε/ζ and χ[43–45].

As in the parmbsc0 parameterization, we used a Monte Carlo method for fitting residual energy, or QM-MM difference (Eq. I), to a Fourier series limited to the third order to maintain the AMBER force-field philosophy (Eq. II). The rotational barrier V_n and the phase angle α of each periodicity (n = 1, 2, 3) were fitted to obtain the minimal error in:

$$E_{\mathrm{dih},x} = E_{QM} - E_{ffbsc0(x=0)} \qquad \text{(I)}$$

where *x* stands for a specific torsion or a combination of torsions (in the case of ε and ζ) and *ffbsc0(x=0)* refers to the standard parameters and the specific *x* torsion set to zero (that used in reference MM or US calculations noted above). The dihedral term is defined as:

$$E_{\mathrm{dih}} = \sum_{torsions} \sum_{n}^{3} \frac{V_n}{2}\left[1 + \cos(n\varphi - \alpha)\right] \quad \text{(II)}$$

where *torsions* denotes a torsion, *n* stands for the periodicity of the torsion, V_n is the rotational barrier,φ is the torsion angle, and α is the phase angle.

Our flexible Metropolis Monte Carlo algorithm allows the introduction of different weights in the fitting for each point of the profile, as well as weighting of energy slopes to guarantee smooth transitions, or even mixing information from different profiles obtained in different conditions or with different levels of QM data. Fittings were done taking all the data in consideration, but with increased weighting at the profile minima (typically five times more than others) specially at the key points computed through the most accurate CCSD(T)/CBS approach (typically weighted nine times more than others). For certain cases like the sugar puckering, detailed attention was needed to properly reproduce the transition region, which was achieved by increasing the importance of the energy maximum and by also introducing weights to the slopes in the calculations. As described before[4], around 5–10 acceptable solutions of the Monte Carlo refinement were tested on short MD simulations (around 50–100 ns) for one small duplex d(CGATCG)$_2$ rejecting those leading to distorted structures. The optimum parameter set (see **Supplementary Discussion** and **Supplementary Table 12**), without additional refinement was extensively tested against experimental data. Note that the way in which the parameters were derived does not guarantee their validity for RNA simulations, for which the use of others already validated RNA force-fields are recommended[45].

Validation of MD simulations.

We performed MD simulations with the PMEMD code from the programs AMBER 11-12 (http://www.ambermd.org), or with GROMACS[46], depending on the given simulation. As shown in **Supplementary Fig. 24**, results are insensitive to the simulation engine or to the use of CPU or GPU-adapted codes[47]. Unless otherwise noted NPT conditions with default temperature and pressure setting, at 300 K and pressure of 1 atm, where used. Calculations employed an integration step of 2 fs in conjunction with SHAKE[48] (or LINCS[49] in the case of GROMACS), to constrain X-H bonds with the default values. The TIP3P[50] or SPCE[51] water models were used, with a minimum buffer of 10 Å solvation layer beyond the solute, and the negatively charged DNA was neutralized with Na^+or K^+ ions[52]. Test simulations with added salt (Na^+Cl^-) showed that DNA helical conformations were not much dependent on the surrounding ionic strength in the 0 to 0.5 M range (**Supplementary Discussion** and **Supplementary Fig. 25**). Long range electrostatic interactions were calculated using the particle mesh Ewald method (PME)[53] with default grid settings and tolerance. All structures were first optimized, thermalized and pre-equilibrated for 1 ns using our standard protocol[8] and were subsequently equilibrated for an additional 10 ns period. Conformational snapshots were saved every 1, 10, 20, or even 100 ps depending on the system size, the objective of the simulation, and its length. Simulations mimicking crystal environments were carried out as described elsewhere[54] for $d(CGATCGATCG)_2$ (PDB: 1D23) using 2 µsec simulation with 12 unit cells (or 32 duplexes) in the simulation periodic box (**Supplementary Fig. 14**), for a total of 64 µsec of duplex simulation.

For annotation of conformational regions at the nucleotide level we used standard criteria. Sugar puckering (C3'-endo for P between 0 º and 36 º (canonical North) C4'-exo for P between 36 º and 72 º, O4'-endo for P between 72 º and 108 º (canonical East), C1'-exo for P between 108 º and 144 º, C2'-endo for P between 144 º and 180 º (canonical South), C3'-exo for P between 180 º and 216 º, C4'-endo for P between 216 º and 252 º, O4'-exo for P between 252 º and 288 º (canonical West), C1'-endo for P between 288 º and 324 º, and C2'-exo for P between 324 º and 360 º), glycosidic torsion (*anti* for 90º to 180 º or -60 º to -180 º, and *syn* for -60 º to 90 º). BI (ε trans, ζ gauche-) and BII (ε gauche-, ζ trans). An H-bond is annotated using standard GROMACS rules and was considered broken when donor-acceptor distance was greater than 3.5 Å for at least ten consecutive picoseconds. Reference A-DNA and B-DNA fiber conformations were taken from Arnott's values[55]. Whenever possible, the simulations were validated against experimental data obtained in solution.

A variety of analyses were performed to characterize the mechanical properties of DNA based on MD simulations. Flexibility analysis was performed using essential dynamics algorithms[56–58], base step stiffness analysis[17, 59, 60], and quasi-harmonic entropies computed by using either Andricioaei-Karplus[61] or Schlitter[62] procedures. Similarities between essential deformation movements were determined using standard Hess's metrics[63] as well as energy-corrected Hess-metrics[59]. The calculation of polymer deformation parameters (persistence length, stretch and twist torsion modules) was done following different approaches to reduce errors associated to the use of a single method to move from atomistic simulations to macroscopic descriptors: i) extrapolation of base step translations and rotations[17, 59], ii) analysis of the correlations in the conformations and fluctuations of the DNA at different lengths[64], and iii) an implementation of Olson's hybrid approach, which requires additional Monte Carlo simulations using MD-derived stiffness matrices[65]. Dielectric constants of DNA were computed using Pettit's procedure[66, 67].

The trajectories were analyzed using AMBERTOOLS (http://www.ambermd.org), GROMACS[46], MDWeb[68], NAFlex[69], and Curves+[70], as well as with in-house scripts (http://mmb.irbbarcelona.org/www/tools).

NMR analysis.

Analysis of the ability of MD trajectories to reproduce NMR observables (NOE-derived interatomic distances and residual dipolar couplings) was done using the last 950 ns of microsecond trajectories. We used the Single Value Decomposition (SVD) method implemented in the program PALES[71] to obtain the orientation tensor that best fitted the calculated and observed RDC values. Violations of the NOE data were computed using the tool *g_disre*, included in the GROMACS package, using distance restraints derived from the deposited BioMagResBank database[72], or as described below when NOEs were collected *de novo* using full relaxation matrix experiments.

The *novo* NMR experiments.

Samples (3 mM oligonucleotide concentration) were suspended in 500 µL of either D2O or H2O/D2O 9:1 in 25 mM sodium phosphate buffer, 125 mM Na^+Cl^-, pH 7. NMR spectra were acquired in Bruker spectrometers operating at 800 MHz, and processed with Topspin software. DQF-COSY (Double Quantum Filter – Correlation spectroscopy), TOCSY (Total Correlation spectroscopy), and NOESY (Nuclear Overhauser effect spectroscopy) experiments were recorded in D2O and H2O/D2O 9:1. The NOESY spectra were acquired with mixing times of 75, 100, 200, and 300 ms, and the TOCSY spectra were recorded with standard MLEV 17 spin lock sequence, and 80 ms mixing time. NOESY spectra were recorded at 5 and 25 ºC.

The spectral analysis program Sparky (https://www.cgl.ucsf.edu/home/sparky) was used for semi-automatic assignment of the NOESY cross-peaks and quantitative evaluation of the NOE intensities**.** Quantitative distance constraints were obtained from NOE intensities by using a complete relaxation matrix analysis with the program MARDIGRAS[73]. Error bounds in the inter-protonic distances were estimated by carrying out several MARDIGRAS calculations with different initial models, mixing times and correlation times (2.0, 4.0 and 6.0 ns). Final constraints were obtained by averaging the upper and lower distance bounds in all the MARDIGRAS runs.

Availability of force-field parameters and porting to different MD codes.

The refined parameters are incorporated in amber-format libraries accessible from http://mmb.irbbarcelona.org/ParmBSC1/. Porting to GROMACS format was done from amber topology files using external utilities (amb2gmx[74] and acpype[75] tools accessible at https://simtk.org/home/mmtools and https://github.com/choderalab/mmtools). Porting to NAMD (http://www.ks.uiuc.edu/Research/namd) is not required since direct reading of AMBER topology files is possible.

Data Management.

Trajectories and the analysis performed were placed in a novel dual database framework for nucleic acid simulations using Apache's Cassandra to manage trajectory data, and MongoDB to manage trajectory metadata and analysis. Results are available at http://mmb.irbbarcelona.org/ParmBSC1/. Details on the Barcelona's nucleic acids database will be presented elsewhere.

Online Methods references

21. Šponer, J., Jurecka, P. & Hobza, P. *J. Am. Chem. Soc.***126,** 10142–10151 (2004).

22. Hobza, P., Kabeláč, M., Šponer, J., Mejzlík, P. & Vondrášek, J. *J. Comput. Chem.***18,** 1136–1150 (1997).

23. Šponer, J. *et al.Chem. Eur. J.***12,** 2854–2865 (2006).

24. Orozco, M. & Luque, F.J. *Chem. Phys.***182,** 237–248 (1994).

25. Colominas, C., Luque, F.J. & Orozco, M. *J. Am. Chem. Soc.***118,** 6811–6821 (1996).

26. Orozco, M., Cubero, E., Hernández, B., López, J.M. & Luque, F.J. in *Computational Chemistry. Reviews of Current Trends***4,** 191–225 (World Scientific Publishing, 1999).

27. Pérez, A. *et al.Chem. Eur. J.***11,** 5062–5066 (2005).

28. Beveridge, D.L. *et al.Biophys. J.***87,** 3799–3813 (2004).

29. Portella, G., Germann, M.W., Hud, N.V. & Orozco, M. *J. Am. Chem. Soc.***136,** 3075–3086 (2014).

30. Krishnan, R., Binkley, J.S., Seeger, R. & Pople, J.A. *J. Chem. Phys.***72,** 650–654 (1980).

31. Woon, D.E. & Dunning Jr, T.H. *J. Chem. Phys.***98,** 1358–1371 (1993).

32. Halkier, A. *et al.Chem. Phys. Lett.***286,** 243–252 (1998).

33. Halkier, A., Helgaker, T., Jørgensen, P., Klopper, W. & Olsen, J. *Chem. Phys. Lett.***302,** 437–446 (1999).

34. Miertuš, S., Scrocco, E. & Tomasi, J. *Chem. Phys.***55,** 117–129 (1981).

35. Miertus, S. & Tomasi, J. *Chem. Phys.***65,** 239–245 (1982).

36. Cances, E., Mennucci, B. & Tomasi, J. *J. Chem. Phys.***107,** 3032–3041 (1997).

37. Bachs, M., Luque, F. J. & Orozco, M. *J. Comput. Chem.***15,** 446–454 (1994).

38. Soteras, I., Curutchet, C., Bidon-Chanal, A., Orozco, M. & Luque, F.J. *J. Mol. Struct. THEOCHEM***727,** 29–40 (2005).

39. Soteras, I., Forti, F., Orozco, M. & Luque, F.J. *J. Phys. Chem. B***113,** 9330–9334 (2009).

40. Soteras, I., Orozco, M. & Luque, F.J. *J. Comput. Aided Mol. Des.***24,** 281–291 (2010).

41. Marenich, A.V., Cramer, C.J. & Truhlar, D.G. *J. Phys. Chem. B***113,** 6378–6396 (2009).

42. Torrie, G.M. & Valleau, J.P. *J. Comput. Phys.***23,** 187–199 (1977).

43. Hart, K. *et al.J. Chem. Theory Comput.***8,** 348–362 (2011).

44. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B. & Bax, A. *J. Biomol. NMR***26,** 297–315 (2003).

45. Zgarbová, M. *et al.J. Chem. Theory Comput.***7,** 2886–2902 (2011).

46. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. *J. Chem. Theory Comput.***4,** 435–447 (2008).

47. Galindo-Murillo, R., Roe, D.R. & Cheatham III, T.E. *Nat. Commun.***5,** (2014).

48. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.C. *J. Comput. Phys.***23,** 327–341 (1977).

49. Hess, B., Bekker, H., Berendsen, H.J.C. &Fraaije, J.G.E.M. *J. Comp. Chem.***18**, 1463–1472(1997).

50. Jorgensen, W.L., Chandrasekhar, J., Madura, J. D., Impey, R.W. & Klein, M.L. *J. Chem. Phys.***79,** 926–935 (1983).

51. Berendsen, H.J.C., Grigera, J.R. &Straatsma, T.P. *J. Phys. Chem.***91**, 6269-6271 (1987).

52. Smith, D.E. & Dang, L. X. *J. Chem. Phys.***100,** 3757–3766 (1994).

53. Darden, T., York, D. & Pedersen, L. *J. Chem. Phys.***98,** 10089–10092 (1993).

54. Liu, C., Janowski, P.A. & Case, D.A. *Biochim. Biophys. Acta (BBA)-General Subj.***1850**, 1059–1071 (2014).

55. Arnott, S. & Hukins, D.W.L. *Biochem. Biophys. Res. Comm.***47**, 1505-1509 (1972).

56. Orozco, M., Pérez, A., Noy, A. & Luque, F.J. *Chem. Soc. Rev.***32,** 350–364 (2003).

57. Pérez, A. *et al.J. Chem. Theory Comput.***1,** 790–800 (2005).

58. Amadei, A., Linssen, A. & Berendsen, H.J.C. *Proteins Struct. Funct. Bioinforma.***17,** 412–425 (1993).

59. Lankaš, F., Šponer, J., Hobza, P. & Langowski, J. *J. Mol. Biol.***299,** 695–709 (2000).

60. Noy, A., Perez, A., Lankas, F., Luque, F.J. & Orozco, M. *J. Mol. Biol.***343,** 627–638 (2004).

61. Andricioaei, I. & Karplus, M. *J. Chem. Phys.***115,** 6289–6292 (2001).

62. Schlitter, J. *Chem. Phys. Lett.***215,** 617–621 (1993).

63. Hess, B. *Phys. Rev. E***62,** 8438 (2000).

64. Noy, A. & Golestanian, R. *Phys. Rev. Lett.***109,** 228101 (2012).

65. Zheng, G., Czapla, L., Srinivasan, A.R. & Olson, W.K. *Phys. Chem. Chem. Phys.***12,** 1399–1406 (2010).

66. Cuervo, A. *et al. Proc. Natl. Acad. Sci.***111**, E3624–E3630 (2014).

67. Yang, L., Weerasinghe, S., Smith, P.E. & Pettitt, P.M. *Bioph. J.***69**, 1519–1527 (1995).

68. Hospital, A. *et al. Bioinformatics***28,** 1278–1279 (2012).

69. Hospital, A. *et al. Nucleic Acids Res.***41,**W47-W55(2013).

70. Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D. & Zakrzewska, K. *Nucleic Acids Res.***37,** 5917–5929 (2009).

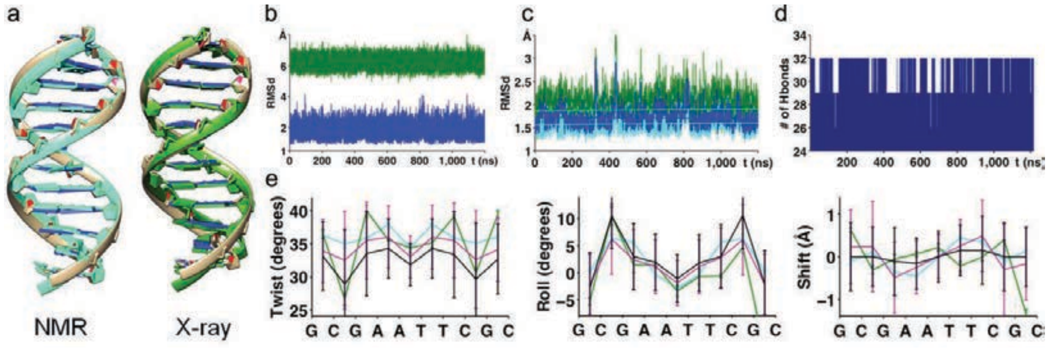
71. Zweckstetter, M. *Nat. Protoc.***3,** 679–690 (2008).

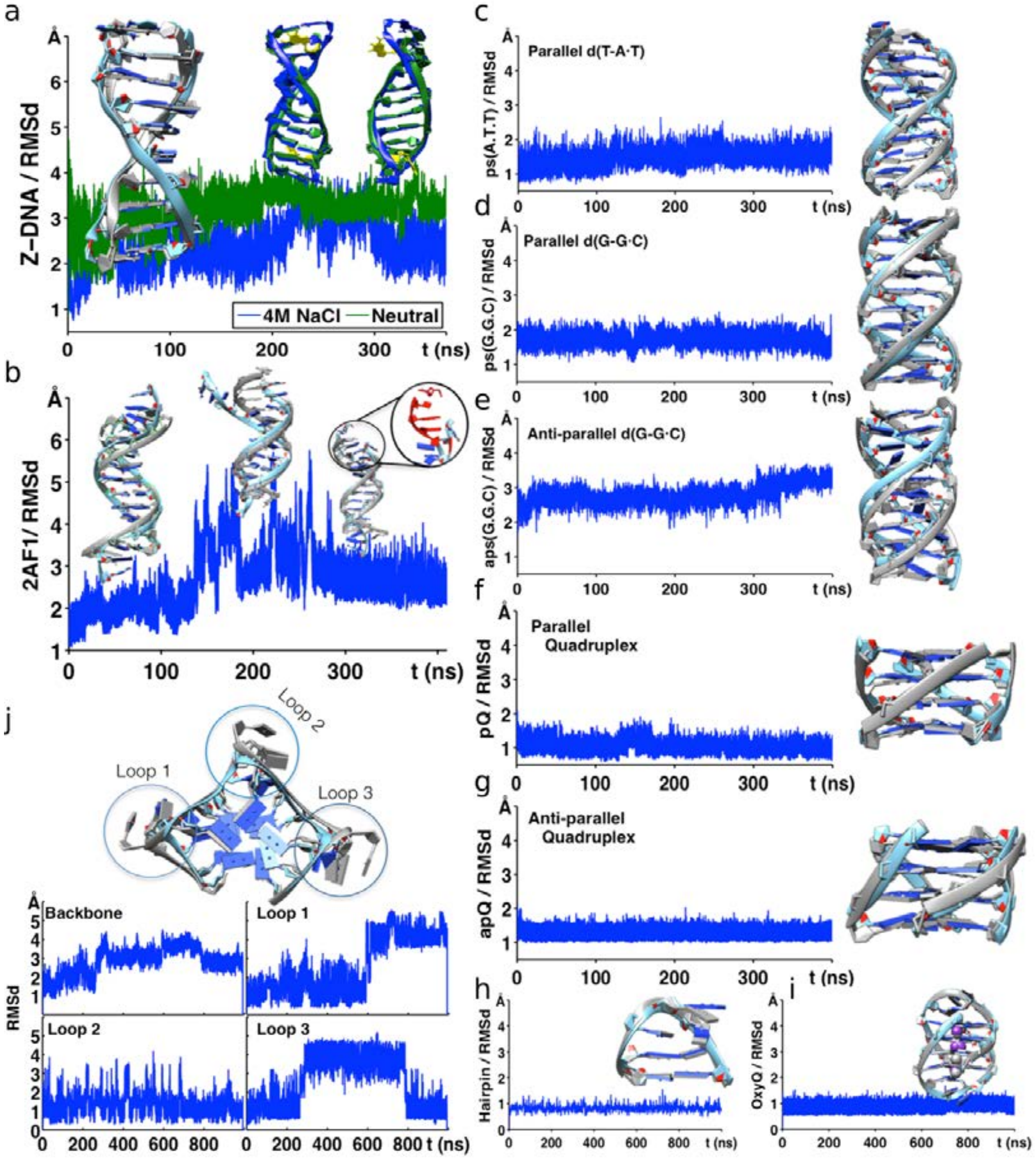
72. Bernstein, F.C. *et al. Eur. J. Biochem.***80,** 319–324 (1977).

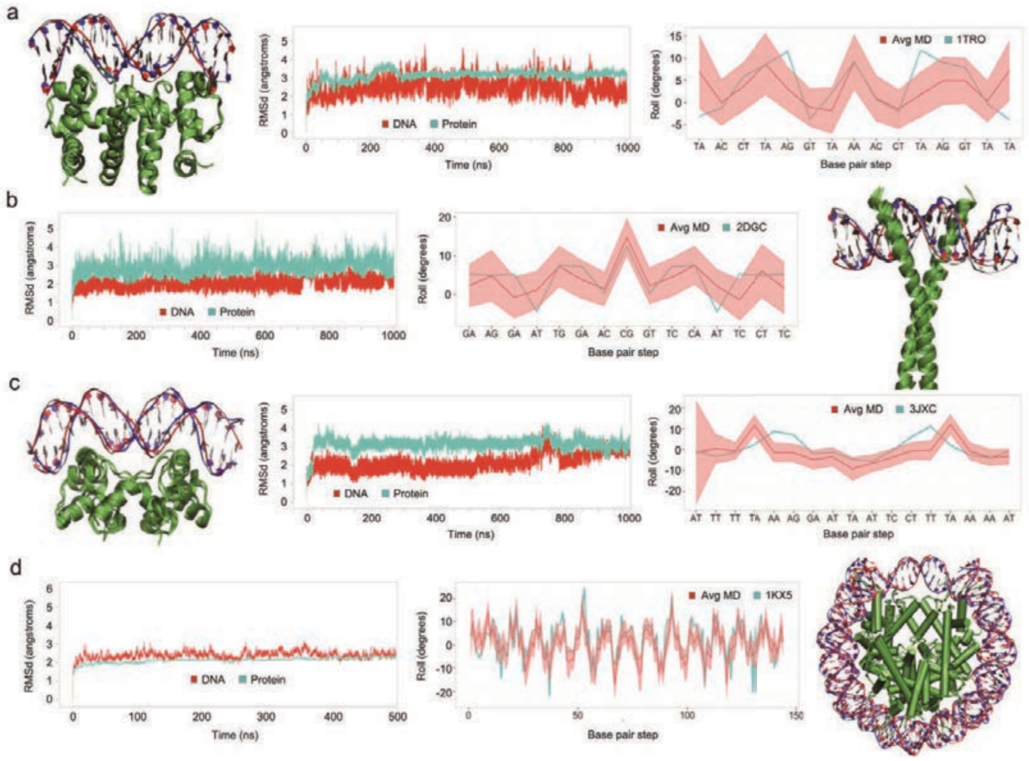
73. Borgias, B.A. & James, T.L. *Journal of Magnetic Resonance (1969)***87,** 475–487 (1990).

74. Mobley, D.L., Chodera, J.D, Dill, K.A., *J.Chem.Phys.***125,** 084902 (2006).

75. Sousa da Silva, A.W. & Vranken, W.F., *BMC Res Notes***5**, 367 (2012).


a
NMR
X-ray
b
Å
RMSd
6
4
2
1
0 200 400 600 800 1,000 t (ns)
c
Å
RMSd
3
2.5
2
1.5
1
0 200 400 600 800 1,000 t (ns)
d
# of Hbonds
34
32
30
28
26
24
200 400 600 800 1,000 t (ns)
e
Twist (degrees)
40
35
30
25
G C G A A T T C G C
Roll (degrees)
10
5
0
−5
G C G A A T T C G C
Shift (Å)
1
0
−1
G C G A A T T C G C


a
Å
Z–DNA / RMSd
6
5
4
3
2
1
0 100 200 300 t (ns)
4M NaCl
Neutral
b
Å
2AF1/ RMSd
6
5
4
3
2
1
0 100 200 300 t (ns)
c
Parallel d(T-A·T)
ps(A.T.T) / RMSd
0 100 200 300 t (ns)
d
Parallel d(G-G·C)
ps(G.G.C) / RMSd
0 100 200 300 t (ns)
e
Anti-parallel d(G-G·C)
aps(G.G.C) / RMSd
0 100 200 300 t (ns)
f
Parallel Quadruplex
pQ / RMSd
0 100 200 300 t (ns)
g
Anti-parallel Quadruplex
apQ / RMSd
0 100 200 300 t (ns)
j
Loop 1
Loop 2
Loop 3
Å
RMSd
Backbone
Loop 1
Loop 2
Loop 3
5
4
3
2
1
0
0 200 400 600 800 0 200 400 600 800 t (ns)
h
Hairpin / RMSd
0 200 400 600 800 t (ns)
i
OxyQ / RMSd
0 200 400 600 800 t (ns)


a
RMSd (angstroms)
0
1
2
3
4
5
0
200
400
600
800
1000
Time (ns)
■ DNA ■ Protein
Roll (degrees)
15
10
5
0
-5
■ Avg MD ■ 1TRO
TA AC CT TA AG GT TA AA AC CT TA AG GT TA TA
Base pair step
b
RMSd (angstroms)
0
1
2
3
4
5
0
200
400
600
800
1000
Time (ns)
■ DNA ■ Protein
Roll (degrees)
20
10
0
■ Avg MD ■ 2DGC
GA AG GA AT TG GA AC CG GT TC CA AT TC CT TC
Base pair step
c
RMSd (angstroms)
0
1
2
3
4
5
0
200
400
600
800
1000
Time (ns)
■ DNA ■ Protein
Roll (degrees)
20
10
0
-10
-20
■ Avg MD ■ 3JXC
AT TT TT TA AA AG GA AT TA AT TC CT TT TA AA AA AT
Base pair step
d
RMSd (angstroms)
0
1
2
3
4
5
6
0
100
200
300
400
500
Time (ns)
■ DNA ■ Protein
Roll (degrees)
20
10
0
-10
-20
■ Avg MD ■ 1KX5
0
50
100
150
Base pair step

Bibliography for Chapter III

1. Dans P, Walther J, Gómez H *et al.* Multiscale simulation of DNA. *Curr Opin Struct Biol - Theory Simul • Macromol Mach 2016* 2016;**37**.
2. Dans PD, Danilāne L, Ivani I *et al.* Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res* 2016;**44**:4052–66.
3. Pérez A, Luque FJ, Orozco M *et al.* Dynamics of B-DNA on the Microsecond Time Scale. *J Am Chem Soc* 2007;**129**:14739–45.
4. Pérez A, Lankas F, Luque FJ *et al.* Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res* 2008;**36**:2379–94.
5. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.
6. Pérez A, Luque FJ, Orozco M *et al.* Frontiers in Molecular Dynamics Simulations of DNA. *Acc Chem Res* 2012;**45**:196–205.
7. Durán E, Djebali S, González S *et al.* Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. *Nucleic Acids Res* 2013;**41**:7220–30.
8. Hospital A, Faustino I, Collepardo-Guevara R *et al.* NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res* 2013;**41**:W47–55.
9. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.
10. Orozco M. A theoretical view of protein dynamics. *Chem Soc Rev* 2014;**43**:5051–66.
11. Pérez A, Marchán I, Svozil D *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 2007;**92**:3817–29.
12. Galindo-Murillo R, Roe DR, Cheatham TE. On the absence of intrahelical DNA dynamics on the μs to ms timescale. *Nat Commun* 2014;**5**:5152.
13. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.
14. Lane AN, Chaires JB, Gray RD *et al.* Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res* 2008;**36**:5482–515.
15. Dršata T, Pérez A, Orozco M *et al.* Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J Chem Theory Comput* 2013;**9**:707–21.
16. Heddi B, Foloppe N, Oguey C *et al.* Importance of Accurate DNA Structures in Solution: The Jun–Fos Model. *J Mol Biol* 2008;**382**:956–70.
17. Pérez A, Lankas F, Luque FJ *et al.* Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res* 2008;**36**:2379–94.
18. Fadrná E, Špačková N, Sarzyñska J *et al.* Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J Chem Theory Comput* 2009;**5**:2514–30.

19. Krepl M, Zgarbová M, Stadlbauer P *et al.* 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* 2012;**8**:2506–20.

20. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.

21. Zgarbová M, Luque FJ, Šponer J *et al.* Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J Chem Theory Comput* 2013;**9**:2339–54.

22. Zgarbová M, Šponer J, Otyepka M *et al.* Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J Chem Theory Comput* 2015;**11**:5723–36.

23. Galindo-Murillo R, Robertson JC, Zgarbová M *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* 2016;**12**:4114–27.

24. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:gkw1355.

CHAPTER IV | DNA Sequence Dependence and Polymorphisms

The sampling problem for duplex B-DNA seems to have been overcome – at least when neglecting slow motions (> ms), such as internal base pair opening. Even over the duration of this thesis, the time scales for small oligomer simulations has quadrupled, with current lengths routinely exceeding the μs limit. With extended sampling time assuring converged ensembles [1,2], highly accurate force fields, like parmbsc1, and novel analysis tools, our understanding of the DNA conformational space has evolved significantly in recent years. The type of B-DNA motions for which we can obtain reliable statistics from μs long trajectories include sampling of different backbone conformational states (most notably BI-BII transitions), rapid fluctuations in the groove widths, sugar repuckering, bending, twisting and sampling of a wide variety of helicoidal parameter distributions, as well as terminal base pair opening [3].

Several groups, including our own [4–7] reported that many of these internal degrees of freedom sample distributions that stray from normality, indicating that the conformational space of DNA is intrinsically polymorphic. These conclusions have recently started to be seriously backed up by experimental evidence [5,8–10], and there is consensus that specific combinations of internal degrees of freedom animated by thermal fluctuations thusly give rise to the different conformational substates.

One of the most intensely studied structural polymorphism of B-DNA in the sub-μs time scale is the BI-BII equilibrium, which has been shown to have implications in protein-DNA binding through the so-called indirect readout mechanism. Many experimental and theoretical studies have established the sequence-dependence of the BII state propensity and its connection to changes in groove width and depth [11]. However, some key aspects of the thermodynamic and kinetic details allowing and driving the BI/BII transitions were still not understood. This motivated us to try to complete the picture.

In a first work on this topic, we addressed the problem of sequence-dependent BI/BII propensities and their stabilizing factors. Previous studies from the group and also in collaboration with the Ascona B-DNA Consortium (ABC) reported that the formation of a C8H8-O3′ contact in RpR and YpR steps was highly correlated to the BII conformation, with BII populations varying as a function of different 3′- and 5′- neighbors [6]. It remained to be clarified what modulates the BII populations in the other two base-pair step types (RpY and YpY). Access to a cohesive set of simulation data where all 136 unique tetranucleotides were represented with statistical significance from the ABC (μABC dataset) allowed the exploration of this question in a systematic way, such that next-to-nearest-neighbor effects were investigated. We so observed a C6H6-O3′ contact – identified as an h-bond type dipole-dipole interaction – was able to explain the BI/BII propensities in RpY and YpY steps, which until then had precluded a comprehensive model in accordance with experimental values. These intra-molecular h-bonds are not only important because they thermodynamically allow the BII substate to exist for certain bps, but also because (i) they connect a structural polymorphism occurring in the backbone with movements in the bases, and (ii) their presence defines the tetranucleotide as the minimum unit necessary to characterize and analyze the sequence dependent BI/BII propensities. These results are presented in the work entitled *The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA* (page ##).

Another well-documented polymorphism at the time, with an important link to the backbone equilibrium, was the twist bimodality, seen particularly in CpG steps [6,7,12]. However, other non-Gaussian and multi-peaked helical parameters distributions had also been obtained in MD simulations for certain base pair steps. As these results were obtained with the parmbsc0 force field the question of a well-represented sequence-dependent equilibrium distribution was still open. This prompted us to perform a new set of simulations with the state-of-the-art parmbsc1 force field [13], covering the sequence space to the same extent as the μABC [12,14–16]. In the second work presented in this Chapter, *The Physical Properties of B-DNA beyond Calladine's rules*, we use this new set of multi-microsecond MD simulations (miniABC dataset) with parmbsc1 to deliver a global view of the polymorphic landscape of each B-DNA tetranucleotide, unifying experimental and theoretical results in a consensus view. Polymorphisms for shift, slide and twist, BI-BII transitions, the formation of the C-H···O h-bonds, and the correlations between all these elements, have been dissected and have allowed us to reformulate Calladine's rules at the tetramer level.

1 The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA.

This article came as a necessary final piece to complete the puzzle of BI/BII state populations in the different B-DNA tetranucleotides. We first observed in the extensive MD trajectories of μABC the formation of a C6H6-O3′ contact in RpY steps, an interaction analogous to the C8H8-O3′ contact of RpR steps. Our analysis clearly related the presence of this interaction in the backbone of the junction between two bases to the backbone transition to BII at the same junction ($R^2 > 0.9$) and the contact was found to stabilize this state.

We therefore set up to provide an exhaustive picture of the mechanism driving the sequence-dependent BI/BII backbone transitions. We show that almost all BI → BII backbone transitions involve the instantaneous formation of the hydrogen bond. We point out that this is accomplished in specific ways for the different bps types. Although the bond formation is in all cases extremely well correlated with the backbone state, the more complex choreography of the transitions is quite different depending on the sequence. The h-bond of different bps involves a combination of changes in helical parameters at the same step and in neighboring steps that is sequence dependent. Along with the transition to BII, water occupancy at the O3′ group decreases dramatically from that in the BI state.

Furthermore, we carried out *ab-initio* MP2 calculations on representative snapshots, which allowed us to quantify the relative strength of these interactions and speculate on the implications to tetramer-level backbone stabilities. The C6–H6···O3′ bonds of RpY steps, although slightly lower than the C8–H8···O3′ interactions reported for RpR bps, are within the expected range of values for canonical hydrogen bond interactions. Our conclusions were supported by analysis of high-resolution experimental structures of unbound DNA, where we she a bimodal distribution of C6···O3′ distances that is also clearly correlated with the BI/BII state in the backbone, in agreement with simulations.

Publication:

- Alexandra Balaceanu, Marco Pasi, Pablo D. Dans, Adam Hospital, Richard Lavery, Modesto Orozco; The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. J. Phys. Chem. Lett., 8, 21-28, 2017.

The role of unconventional hydrogen bonds in determining BII propensities in B-DNA

Alexandra Balaceanu[1,2,&], Marco Pasi[3,4,&], Pablo D. Dans[1,2,&],
Adam Hospital[1,2], Richard Lavery[3,*], Modesto Orozco[1,2,5,*]

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology. Baldiri Reixac 10-12, Barcelona 08028, Spain.

[2] Joint BSC-IRB Program in Computational Biology, Institute for Research in Biomedicine. Baldiri Reixac 10-12, Barcelona 08028, Spain.

[3] MMSB, Univ. Lyon I/CNRS UMR 5086, Institut de Biologie et Chimie des Protéines, 7 Passage du Vercors, Lyon 69367, France.

[4] School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, University Park NG7 2RD, UK.

[5] Department of Biochemistry and Biomedicine, Faculty of Biology, University of Barcelona. Diagonal 643, Barcelona 08028, Spain.

[&] Equally contributing authors.

[*] Correspondence to: Prof. Richard Lavery (richard.lavery@ibcp.fr) or Prof. Modesto Orozco (modesto.orozco@irbbarcelona.org).

ABSTRACT

An accurate understanding of DNA backbone transitions is likely to be the key for elucidating the puzzle of the intricate sequence-dependent mechanical properties that govern most of the biologically relevant functions of the double helix. One factor believed to be important in indirect recognition within protein-DNA complexes is the combined effect of two DNA backbone torsions (ε and ζ) which give rise to the well-known BI/BII conformational equilibrium. In this work we explain the sequence dependent BII propensity observed in RpY steps (R = purine; Y = pyrimidine) at the tetranucleotide level with the help of a previously undetected C-H···O contact between atoms belonging to adjacent bases. Our results are supported by extensive multi-microsecond molecular dynamics simulations from the Ascona B-DNA Consortium, high-level quantum mechanical calculations, and data mining of the experimental structures deposited in the Protein Data Bank.

INTRODUCTION

The fact that DNA's overall conformation is associated with changes in backbone geometry became apparent from the analysis of the first generation of successfully resolved diffraction patterns[1,2]. One of the major backbone structural polymorphisms in B-DNA arises from its ability to populate two distinct conformations, known as BI and BII[3] (see scheme in Figure 1A). Within each DNA strand, the phosphodiester junction between two consecutive bases undergoes fast inter-conversions between two states defined by specific combinations of rotations around the ε and ζ dihedrals. The canonical state, referred to as BI, features ε/ζ in a trans/gauche- (t/g-) conformation, while the other state, BII, has ε/ζ in g-/t conformation. To determine BI/BII equilibrium in B-DNA, proton and Phosphate NMR experiments[4-7], Molecular Dynamics (MD) simulations[8-10], and data mining of crystal structures from databases[11-13] have being historically used as the preferred methods. Following initial observations based on crystal structures showing that BI/BII transitions were associated with base destacking and minor groove widening[14,15], computer MD simulations have shed light on the influence of water and ion dynamics on the propensity of BI/BII states[8,9,16,17]. Destacking of two successive bases along one strand and water migration were found to be necessary, but not sufficient, conditions for the adoption of the BII conformation[16,18]. Also based on crystal structures, successive nucleotide in one strand have been shown to have anti-correlated backbone conformational states[11,13]. From the very beginning, ^{31}P-NMR experiments were crucial to identified sequence dependence as a key modulator of BI/BII transitions, capable of fine-tuning the population of states. Hartmann's group has been particularly active in this field, providing a sequence dependent view of BI/BII at the dinucleotide level (and even tetranucleotide level)[4,7,19,20], but also putting into perspective NMR results with values obtained from crystal structures or MD simulations[21]. Moreover, recent MD studies have also shown[10,22,23] that the impact of sequence in BI/BII equilibrium is more complex than anticipated, supporting the importance of the tetranucleotide environment.

The transition between backbone substates can only be accurately captured in its entirety from a cohesive set of data where all 136 unique tetranucleotides are represented with statistical significance, which precludes the use of experimental information deposited in structural databases[22], having to rely on the use of information gained from atomistic Molecular Dynamics (MD) simulations. To this end, we performed part of the present analysis on trajectories obtained by the Ascona B-DNA Consortium

(ABC, bisi.ibcp.fr/ABC/Welcome.html) who created a database of simulations, called μABC, containing multi-microsecond MD of DNA oligomers containing multiple copies of all the 136 distinct tetranucleotides[9,10]. An impressive amount of information was already extracted from this collection of simulations, extending our vision of the conformational landscape of B-DNA at the tetranucleotide level[10] and leading to some interesting observations, including the fact that YpR and YpY dinucleotide steps rarely populate the BII state, and also that the base pairs flanking a given step can significantly modulate the BI/BII equilibrium in RpR and RpY steps. Interestingly, a previously uncharacterized C-H⋯O hydrogen bond (H-bond) between the C8-H8 atoms of the R base and the O3' atom of the corresponding 5' phosphate turned out to be a key player in the stabilization of BII states at steps featuring a purine in the 3'-position (i.e. RpR and YpR)[10,23]. It follows that the C8-H8⋯O3' H-bond is not expected to have any role in the stabilization of the BII state in RpY and YpY steps, where it cannot be formed. Our analysis of RpY steps from the μABC set has revealed that a similar contact can be formed between the same O3' oxygen in the backbone phosphate and C6-H6 atoms of pyrimidines. In this work, we characterize in detail the nature of this previously undetected contact, first as captured by the classical potentials of MD simulations, and second, by carrying out Quantum Mechanics (QM) calculations at the MP2 level on several representative structures of RpY steps taken from the MD trajectories. The electron density obtained was *a posteriori* analyzed using the Atoms in Molecules[24] (AIM) approach in order to determine the stabilizing nature of this interaction. Finally, we extracted all the RpY steps from the high-resolution X-ray structures of isolated DNA deposited in the Protein Data Bank (PDB). As suggested by our atomistic simulations, RpY steps in PDB found in the BII state showed a distance between the O3' and C6 atoms compatible with the presence of the stabilizing C6-H6⋯O3' H-bond. Our results also demonstrate that the occurrence of this interaction is highly time-correlated with the backbone BI → BII transitions and establish it as a stabilizing factor of the BII state, providing a complete view of the BI/BII equilibrium.

RESULTS AND DISCUSSION

We started by analyzing two of the μABC simulations[10], namely those referring to the CAAG (5'-GCAGCAAGCAAGCAAGGC-3') and TAAG (5'-GCAGTAAGTAAGTAAGGC-3') sequences, consisting of 3.5 repeats of each of these two tetranucleotides with GpC base pairs capping each end. We focused our analysis on the centermost G8pC9 (Watson strand) and G24pC25 (Crick strand) steps from the CAAG oligomer, along with G8pT9

(Watson) and A24pC25 (Crick) steps of the TAAG sequence. These were chosen because GpY and ApC steps show the highest BII propensities among RpY steps from MD simulations (75% and 60% BII respectively, see Figure 1B). The reader should be aware that when comparing NMR, MD simulations, and X-ray structures, differences exist between the specific BII percentages assigned to some base pairs steps[12,21]. The sources of uncertainties from crystal structures have being discussed several times and are clearly related to low resolution (frequently insufficient to define backbone's states), lack of dynamics, and lattice restraints[13,25,26]. In the same way, changes in NMR's refinement protocols or annealing procedures from one experiment to another[27,28], the usually low number of restraints due to the low density of protons in DNA, and the frequent overlap of several NOE peaks[20], are the main sources of uncertainties, being the average standard deviation in the prediction of BII percentages ±8[21]. In addition, both experimental sources of BII values suffer from a sequence bias due to the limited number of tetranucleotide represented in the limited set of sequence available[20,22,29]. For its part, the parmBSC0[30] force field for MD simulations of DNA, the gold standard for the last decade used to produce the μABC dataset, is known to produce an overall underestimation of Twist and a clear underestimation of BII populations[28,31], in particular for YpR steps. Despite these considerations, the stabilizing C-H···O contacts are postulated to exist beyond the apparent discrepancies reported between the different methods or the specific value assigned to a specific base pair step. Nevertheless, we decided to support our conclusions on the role of the C-H···O H-bonds, simulating again the CAAG and TAAG sequences, following the same protocol[10], but with the latest force field for DNA parmBSC1[31] which is known to fix the Twist and BII issues. These new simulations validate our results, and ensure that our conclusions goes beyond the use of a specific force field (results obtained with parmBSC1 are presented in the Supporting Data but discussed throughout the text).

To correctly capture the fast inter-conversion between backbone states[10,23] or related twist/slide states[23], we extracted information from the simulations every 1 ps leading to conformational ensembles comprising more than 10^6 structures for each of the two sequences. From the ensemble of collected conformations, we identified the existence of an interaction between the O3' atom of the backbone phosphate of the GpC, GpT and ApC junctions and the C6-H6 atoms of the pyrimidine (see scheme in Figure 2A). This contact is structurally equivalent to the C8-H8···O3' H-bond identified in RpR or YpR steps[10,23], where it has been shown to play a fundamental role in stabilizing the BII state. This equivalence suggests that this new interaction can explain the BII propensity observed for the two remaining classes of steps, namely RpY and YpY. To evaluate whether this is the case, we turned to the complete μABC dataset, and found that the occurrence of the C6-H6···O3’ hydrogen bond is indeed in perfect sequence-dependent correlation with the BII state population of RpY and YpY steps (compare figures 1B with 2B, and see the correlation in Figure 2C). This interaction can reasonably be termed a H-bond since the partial charges assigned in the force-field (obtained by QM fitting[32]) to the C6-H6 atoms in thymine or cytosine generate a significant bond dipole (Table 1). Furthermore, the geometrical features of the three atoms, whenever this contact occurs, are consistent with the partial covalent nature of hydrogen bonding[33] as assessed from angle and distance distributions within the ensemble of MD structures (Figure 3A shows the angle distribution of structures with H6···O3’ distances below 2.5 Å). We validate this geometrical arrangement by obtaining results totally equivalent with the last generation force field for DNA[31] (Figure S1).

To further confirm our interpretation we carried out *ab initio* MP2 calculations on seven different snapshots from each of the two simulations (see Methods). Five representative structures were taken from the most populated state in the C6-H6···O3’ angle/C6···O3’ distance space, while two other structures belonged to the marginal bins of the distributions depicted in Figure 3A with angles above 170° coupled with donor-acceptor distances below 3.35 Å. The electron densities obtained in this way were analyzed using the AIM approach[24] to determine the stationary points and the gradient paths (obtained from the first derivative of the electron density) connecting them. In particular, we focused on the bond critical points (bcp) generated between hydrogen and acceptor group, as previous studies have demonstrated that canonical H-bonds (X-H···Z; with X and Z being electronegative atoms) are associated with electron densities at the bcp that vary in the range from 0.002 to 0.034 atomic units (a.u.)[34]. Bond critical points and bond paths between the H6 and O3’ atoms were found in all the 21 electron densities analyzed (see Figure 3B for a representative scheme, and Table 1 for numerical description) supporting the existence of H-bonds. On average, the electron densities at the bcp were found to be around 0.011 a.u., within the expected range of values for canonical interactions (see Table 1 and reference 34), even though slightly lower than those reported for the C8-H8···O3’ H-bond in RpR steps (two cases are shown for comparison). The positive value of the Laplacian (second derivative of the electron density) at the bcp indicates a depletion of electron density towards the interacting nuclei from a density maximum, another feature consistent with the formation of hydrogen bonds[24]. We calculated the gas phase stabilization provided by the specific C6-

H6···O3’ contact in all cases, by estimating the interaction energy from the linear relation described by Cubero *et al*[35]. We found that each C6-H6···O3’ hydrogen bond stabilizes the BII state by more than 3 kcal mol^{-1}, a value only slightly lower compared with the equivalent C-H···O bond described for RpR steps (Table 1)[23]. It is worth noting that a 3 kcal mol^{-1} stabilizing effect should completely drive the equilibrium to BII, but part of this stabilization will be compensated by hydration effects, since water occupancy at the O3’ group decreases dramatically from 36% in the BI state (when the C-H···O bond is not formed) to only 1.4% in BII (data from the analysis of 10^5 structures taken from the last 100 ns of trajectory filtered according to the BI/BII state, confirmed by both force fields).

A more detailed analysis of the MD time series shows that the C6-H6···O3’ contact occurs simultaneously with the formation of BII states: our results indicate that almost all BI → BII backbone transitions involve the instantaneous formation of the hydrogen bond (Figure 4A). This sheds new light on the question of whether the hydrogen bond forms prior to the transition, driving the backbone into a BII state slowly, allowing for a period of structural frustration. Our results suggest that this interaction is more of a stabilizing force than a driving element. This is in agreement with the observed average lifetime of hydrogen bond formation (25.6 ps for GpC and 16.5 ps for GpT), compared to the average lifetimes of the corresponding BII states (23.6 ps and 15.8 ps respectively). These results were confirmed by analyzing the time series obtained with parmBSC1[31] force field. Although the percentage occurrence of the BII state and C6-H6···O3’ hydrogen bond at GpC and ApC steps varies slightly (by less than 10%) when using parmBSC1, the extent to which the two events are time-correlated is substantially unchanged (compare the time series and distributions of Figures 4A and S2).

From a mechanical point of view, the formation of the C8-H8···O3’ H-bond in RpR steps has been shown to be coupled to slide polymorphism at the base level in the same junction[10,23]. On the contrary, no helical parameter or torsion angle showed a two-state distribution coupled to the formation of the C6-H6···O3’ hydrogen bond in RpY steps (again confirmed by both force fields). It seems that the mechanical coupling between slide and the backbone with a purine in the 3’ position (which helps to bring closer the C8-H8 atoms to the backbone), is not necessary in the case of pyrimidines in 3’. We also did not find any coupled role of cations in these transitions (in contrast to that reported for CpG steps[23]), while the hydration change around the backbone atoms, synchronized with the BI/BII transition, and the formation of the intra-molecular C-H···O hydrogen bond, produced a local water migration in agreement with that reported experimentally[18].

To confirm our results we performed two ‘proof of concept’ simulations, with the same sequences reported above, labeled TAAG(H6-) and CAAG(H6-), where the H6 atom from the pyrimidine base was removed and its charge was transferred to the C6 atom (see Methods and reference 23). This "alchemical" base is useful in testing the conformational impact of the C6-H6···O3’ H-bond. As expected, in these simulations C6 and O3’ no longer come into close contact and the backbone of RpY steps undergoes significantly fewer transitions to the BII state (up to 69% less BII for GpC), suggesting that without the C-H···O stabilizing interaction the backbone cannot last in time in the BII state (its average lifetime is decreased to 10.6 ps for GpC and 8.8 ps for GpT, half its normal value), or easily access a g-/t state of the ε/ζ torsions (Figure 4B and Figure S3).

Finally, we performed an analysis of high-resolution experimental structures of isolated DNA to find experimental support for our hypothesis. For this purpose, we extracted all high-resolution X-ray structures of isolated dsDNA oligomers deposited in PDB (R <2.5 Å, see Table S1); the resulting 554 experimental structures contain information on 3,991 RpY dinucleotides (37% GpC, 18% GpT, 19% ApC, and 26% ApT). We found that the distribution of C6···O3’ distances is bimodal, in agreement with simulations, and clearly correlated with the BI/BII state in the backbone (Figure S4). For the GpC case, for which we have better statistics, the C6···O3’ distance decreases in average from 5.03 to 3.48 Å (with a s.d. of 0.4 Å), when moving from BI to BII state in the crystal structures. We repeated the analysis for RpR steps (484 structures analyzed, see Table S1 and Figure S4), finding, analogously, a shortening of the C8···O3’ distance from 5.20 (BI) to 3.80 Å (BII). Equivalent results were obtained for YpR and YpY steps (Table S2 and Figure S4). It should be noted that in spite of the extended set of structures used in this work, BII propensities from crystal structures still seems underestimated when compared with NMR and MD results (Table S2)[4-6,12,20,21]. The reasons for the apparent discrepancies between the methods, which are beyond the scope of the present work, are complex, of diverse sources, and have been partially addressed recently[21]. In summary, despite the relative scarcity of experimental structural data, the analysis of crystal structures provides quantitative support for the importance of the C6-H6···O3’ interaction discussed here, as well as for the C8-H8···O3’ H-bond previously reported[10,23].

The observed sequence-dependent BII propensity in RpY steps, as obtained from MD simulations, can now be explained by taking into account the established hierarchy of bond strength, inferred from the populations of H-bond formation corresponding to each purine-pyrimidine combination (Figure 2B). Considering all possible steps, H-bonds in RpR are the strongest, with GpA being the most favorable, while RpG and RpC interactions are of similar strength. RpT and YpR contacts are rather weaker, but a H-bond in YpY steps is indeed very infrequent. We also observed the hindering effect one H-bond has on the formation of a second H-bond in a neighboring step, in agreement with the known anti-correlation between adjacent BII backbone states. Accordingly, applying these simple considerations in a tetranucleotide context would predict, for example, that a YpRpYpY sequence should result in the highest BII content among all RpY steps in MD simulations, since it is unfavorable to form a H-bond in either of the flanking base steps, while the opposite effect should be observed within an RpRpYpR sequence, where both flanking steps will compete for H-bond formation, leading to conformational frustration. This is an important effect to understand sequence-dependent propensities at the tetranucleotide level, since the crankshaft motion of the backbone ensure almost every time that alternate BI/BII/BI/BII states will be observed in successive junctions in the same strand[13], leading to conformational frustration when two or more dinucleotides with high BII content are side-by-side. The confirmation of these predictions (Figure 1B), and the extension of our conclusions to RpR and YpR steps[10], supported by last generation force field, allows us to conclude that the newly detected C-H···O H-bond makes an important contribution to deciphering the sequence-dependent BII propensity within B-DNA, which in turn has an important role to play in protein-DNA recognition processes.

METHODS

Molecular dynamics simulations. Sequence dependence analyses are based on the 39 multi μs simulations collected in 2014 by the Ascona B-DNA Consortium that form the μABC data set[10]. Since this analysis, a new version of the Amber force field for DNA named parmBSC1[31] was published. As this force field has been shown to produce trajectories in even better correlation with experiment[24,36], we validate parmBSC0-derived conclusions by simulating again with parmBSC1 the two double stranded B-DNA oligomers with sequences 5'-GCAGCAAGCAAGCAAGGC-3' (labeled CAAG) and 5'-GCAGTAAGTAAGTAAGGC-3' (labeled TAAG) used in the detailed analysis presented here. Additionally, simulations (using both parmBSC0[30] and parmBSC1[31]) were carried out removing the H6 atom of either the thymine or the two cytosines of interest: labeled TAAG(H6-) and CAAG(H6-) respectively. To maintain the total charge of the system in these model calculations, the H6 charge was transferred to the C6 atom[23]. All simulations were carried out using the protocol described in Pasi *et al*[10]. The results obtained with parmBSC1 are equivalent to those discussed in the main text, and are presented in the Supporting Data.

Analysis of trajectories. Trajectories were pre-processed with the cpptraj module of the AmberTools15 package[37]. Conformational analysis was performed using the Curves+ and Canal programs[38], which provide a full set of helical, backbone and groove geometry parameters, and further dissection of these quantities was done with the use of NaFleX server[39] and in-house tools. We consider a H-bond was formed when the distance between C6/C8 and O3' was below 4 Å. Trajectories will be deposited in the BigNASim database[40] of the European MuG Virtual Research Environment (www.multiscalegenomics.eu/MuGVRE/).

Quantum mechanical calculations. To make a first principles confirmation of the existence of the C-H···O intra-molecular hydrogen-bond, Bader's atoms in molecules (AIM)[24] electron topology analysis was used. Seven representative snapshots from the three selected dinucleotides (GpC, GpT and ApC) were extracted from the MD simulations to perform single-point MP2 calculations. Waters and ions were removed and only the dinucleotide step was kept and subjected to single-point calculations at the MP2(FC)/6–31G(d,p) level of theory using Gaussian 09[41]. H atoms were used to complete the valency of the 5' and 3' oxygen atoms. The electron density, the gradient and its Laplacian at the bcp were computed and analyzed using the program AIM-UC[42].

ACKNOWLEDGEMENTS

MINECO Severo Ochoa Award of Excellence, Government of Spain (to IRB Barcelona); Spanish Ministry of Science (BIO2015-64802-R, BFU2014-61670-EXP) (to M.O.); Catalan SGR (to M.O.); Instituto Nacional de Bioinformática (to M.O.); European Research Council (ERC SimDNA) (to M.O.); H2020 program (MuG and BioExcel projects) (to M.O.). Agence de la Recherche National, France (CHROME ANR-12-BSV5-0017-01) (to M.P. and R.L.). M.O. is an ICREA academia researcher. P.D.D. is a PEDECIBA and SNI (ANII, Uruguay) researcher. We thank the Ascona B-DNA Consortium for the μABC data set of trajectories. We also thank Prof. John Maddocks, Daiva Petkevičiūtė, and Alessandro Patelli for invaluable *scientific discussion.*

SUPPORTING INFORMATION

Codes of the structures analyzed from the PDB, analyses from the simulations performed without the H6 atom, and results obtained with the parmBSC1 force-field are included in the supporting data.

REFERENCES

1. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proceedings of the National Academy of Sciences of the United States of America, 95(19), 11163–11168.
2. Matthews, B. W. (1988). No code for recognition. Nature, 335(6188), 294–295.
3. Hartmann, B., Piazzola, D. & Lavery, R. (1993). BI-BII transitions in B-DNA. Nucleic Acids Res., 21(3), 561-568.
4. Heddi, B., Foloppe, N., Bouchemal, N., Hantz, E. & Hartmann, B. (2006) Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. J. Am. Chem. Soc., 128, 9170–9177.
5. Schwieters, C. D. & Clore G. M. (2007). A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. Biochemistry, 46(5), 1152-1166.

6. Tian, Y., Kayatta, M., Shultis, K ., Gonzalez, A., Mueller, L.J. & Hatcher, M.E. (2008). 31P NMR investigation of backbone dynamics in DNA binding sites. J. Phys. Chem. B, 113, 2596–2603.

7. Abi-Ghanem, J., Heddi, B., Foloppe, N. & Hartmann, B. (2010). DNA structures from phosphate chemical shifts. Nucleic Acids Res., 38(3), e18.

8. René, B., Masliah, G., Antri, S. El, Fermandjian, S., & Mauffret, O. (2007). Conformations and dynamics of the phosphodiester backbone of a DNA fragment that bears a strong topoisomerase II cleavage site. The Journal of Physical Chemistry. B, 111(16), 4235–4243.

9. 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 Research, 38(1), 299–313.

10. 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 Research, 42(19), 12272–12283.

11. Djuranovic,D. and Hartmann,B. (2003) Conformational characteristics and correlations in crystal structures of nucleic acid oligonucleotides: evidence for sub-states. J. Biomol. Struct. Dyn., 20, 771–788.

12. Madhumalar, A. & Bansal, M. (2005). Sequence preference for BI/BII conformations in DNA: MD and crystal structure data analysis. J Biomol Struct Dyn., 23(1), 13-27.

13. Svozil, D., Kalina, J., Omelka, M. & Schneider, B. (2008) DNA conformations and their sequence preferences. Nucleic Acids Res., 36(11), 3690-3706.

14. Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., & Dickerson, R. E. (1981). Structure of a B-DNA dodecamer: conformation and dynamics. Proceedings of the National Academy of Sciences of the United States of America, 78(4), 2179–2183.

15. Schneider, B., Neidle, S., & Berman, H. M. (1997). Conformations of the sugar-phosphate backbone in helical DNA crystal structures. Biopolymers, 42(1), 113–124.

16. Pichler, A., Rüdisser, S., Winger, R. H., Liedl, K. R., Hallbrucker, A., & Mayer, E. (2000). The role of water in B-DNAs BI to BII conformer substates interconversion: a combined study by calorimetry, FT-IR spectroscopy and computer simulation. Chemical Physics, 258(2), 391–404.

17. Rudolf H. Winger, Klaus R. Liedl, Simon Rüdisser, Arthur Pichler, Andreas Hallbrucker, A., & Mayer, E. (1998). B-DNA's BI → BII Conformer Substate Dynamics Is Coupled with Water Migration. J. Phys. Chem. B, 102(44), 8934–8940.
18. Grzeskowiak, K., Yanagi, K., Privé, G. G., & Dickerson, R. E. (1991). The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-G. The effect of base pair reversals. The Journal of Biological Chemistry, 266(14), 8861–8883.
19. Heddi, B., Foloppe, N., Oguey, C. & Hartmann, B. (2008). Importance of accurate DNA structures in solution: the Jun-Fos model. J. Mol. Biol., 382, 956–970.
20. Heddi, B., Oguey, C., Lavelle, C., Foloppe, N. & Hartmann, B. (2010). Intrinsic flexibility of B-DNA: the experimental TRX scale. Nucleic Acids Res., 38(3), 1034-1047.
21. Imeddourene, A. B., Elbahnsi, A., Guéroult, M., Oguey, C., Foloppe, N., & Hartmann, B. (2015). Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. PLOS Comput. Biol., 11(12), e1004631.
22. Dans, P. D., Pérez, A., Faustino, I., Lavery, R., & Orozco, M. (2012). Exploring polymorphisms in B-DNA helical conformations. Nucleic Acids Research, 40(21), 10668–78.
23. Dans, P. D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R., & Orozco, M. (2014). Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. Nucleic Acids Research, 42(18), 11304–20.
24. Bader, R. F. W. (1994). Atoms in Molecules: A Quantum Theory. Oxford University Press, New York.
25. Dickerson, R. E., Grzeskowiak, K., Grzeskowiak, M., Kopkam, M. L., Larsen, T., Lipanov, A., Prive, G. G., Quintana, J., Schultz, P., Yanagi, K., et al. (1991). Polymorphism, packing, resolution, and reliability in single-crystal DNA oligomer analyses. Nucl. Nucl., 10, 3-24.
26. Jain, S., Richardson, D. C. & Richardson, J. S. (2015). Computational Methods for RNA Structure Validation and Improvement (chapter 7). In Methods in Enzymology: Structures of Large RNA Molecules and Their Complexes. Woodson, S. A. & Allain, F. H. T. (Eds). Volume 558, 181–212. Elsevier, USA.
27. Heddi, B., Foloppe, N., Oguey, C. & Hartmann, B. (2008). Importance of accurate DNA structures in solution: the Jun-Fos model. J. Mol. Biol., 382, 956–970.
28. Dans, P. D., Ivani, I., Hospital, A. Portella, G., González, C. & Orozco, M. (2016). How accurate are accurate force-fields for DNA. Nucleic Acids Res. In press.

29. Pérez, A., Luque, F. J. & Orozco, M. (2012) Frontiers in molecular dynamics simulations of DNA. Acc. Chem. Res., 45, 196–205.

30. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A. & Orozco, M. (2007). Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophysical Journal, 92(11), 3817–29.

31. 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. Nature Methods,13, 55-58.

32. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. & Kollman, P.A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc., 117, 5179–5197.

33. Grabowski, S. J., Sokalski, W. A., & Leszczynski, J. (1996).The Possible Covalent Nature of N−H···O Hydrogen Bonds in Formamide Dimer and Related Systems: An Ab Initio Study. J. Phys. Chem. A., 110(14), 4772-79.

34. Koch, U., & Popelier, P. L. A. (1995). Characterization of C-H-O Hydrogen Bonds on the Basis of the Charge Density. The Journal of Phys. Chem., 99(24), 9747–9754.

35. Cubero, E., Orozco, M., Hobza, P. & Luque, F. J. (1999).Hydrogen bond versus anti-hydrogen bond: a comparative analysis based on the electron density topology. J. Phys. Chem. A, 103, 6394–6401.

36. Dans, P. D., Danilāne, L., Ivani, I., Dršata, T., Lankaš, F., Walther, J., Illa Pujagut, R., Battistini, F., Gelpí, J. L., Lavery, R., & Orozco, M. (2016). Long-timescale dynamics of the Drew–Dickerson dodecamer. Nucleic Acids. Res., 44(9), 4052-4066.

37. Case, D.A., Babin, V., Berryman, J.T., Betz, R.M., Cai, Q., Cerutti, D.S., Cheatham, T.E. III, Darden, T.A., Duke, R.E., Gohlke, H. et al.(2014). AMBER. University of California, San Francisco.

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

39. Hospital, A., Faustino, I., Collepardo-Guevara, R., González, C., Gelpí, J. L., &Orozco, M. (2013). NAFlex: A web server for the study of nucleic acid flexibility. Nucleic Acids Res., 41(W1), W47-W55.

40. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M., *et al.*(2015) BigNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. Nucleic Acids Res., 44(D1), D272-D278.

41. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G.A. et al. (2009). Gaussian 09, Revision D.01., Gaussian Inc., Wallingford CT.

42. Vega,D. & Almeida,D. (2014). AIM-UC: an application for QTAIManalysis. Journal of Comput. Methods in Sci. and Engineering, 14, 131–136.

43. Martin-Pintado, N., Deleavey, G.F., Portella, G., Campos-Olivas, R., Orozco, M., Damha, M.J. & Gonzalez, C. (2013). Backbone FC-H···O hydrogen bonds in 2F-substituted nucleic acids. Angew. Chem. Int. Ed. Engl., 52, 12065–12068.

TABLES

Table 1. C-H···O hydrogen bond parameters (average distance, average angle, charges, and energy) and electron density at the bond critical point was computed at the MP2(FC)/6-31G(d,p) level.

Seq / dinuc.	Distance (Å)	Angle (º)	ρ (a.u.)	∇^2 (a.u.)	E_{Hbond}[a] (kcal/mol)	q_C[b] (e)	q_H (e)	q_O (e)	U_{Coulomb} (kcal/mol)
AGCA (GpC)	3.30 ± 0.02	139	0.011 ± 0.002	0.037 ± 0.004	-3.51	-0.018	0.229	-0.523	-11.6
AGTA (GpT)	3.19 ± 0.06	129	0.011 ± 0.001	0.037 ± 0.005	-3.75	-0.221	0.261	-0.523	-13.8
TACT (ApC)	3.34 ± 0.06	142	0.011 ± 0.001	0.037 ± 0.004	-3.44	-0.018	0.229	-0.523	-11.6
CCGG (GpG)[c]	3.2	138	0.013	0.043	-4.52	0.074	0.200	-0.523	-10.8
TCGA (GpA)[c]	3.0 ± 0.1	136	0.018 ± 0.001	0.059 ± 0.004	-7.19	0.161	0.188	-0.523	-10.0
A-U bp[d]	3.6	---	0.006	0.021	-0.85	0.572	0.060	-0.548	-3.0
U-U bp[e]	3.3	---	0.016	0.047	-5.19	-0.364	0.181	-0.548	-9.7

[a] Computed from the Laplacian values using the linear regression from Cubero *et al*[35]. [b] Taken directly from the AMBER 14 libraries (as reported in the original parm94 article[32]). [c] Adapted with permission from Dans *et al*[23]. [d] We reproduced the values for the C2-H2···O2 H-bond as reported in the work of Martin-Pintado *et al*[43]. [e] Idem than (d) for the C5-H5···O2 hydrogen bond.

FIGURES

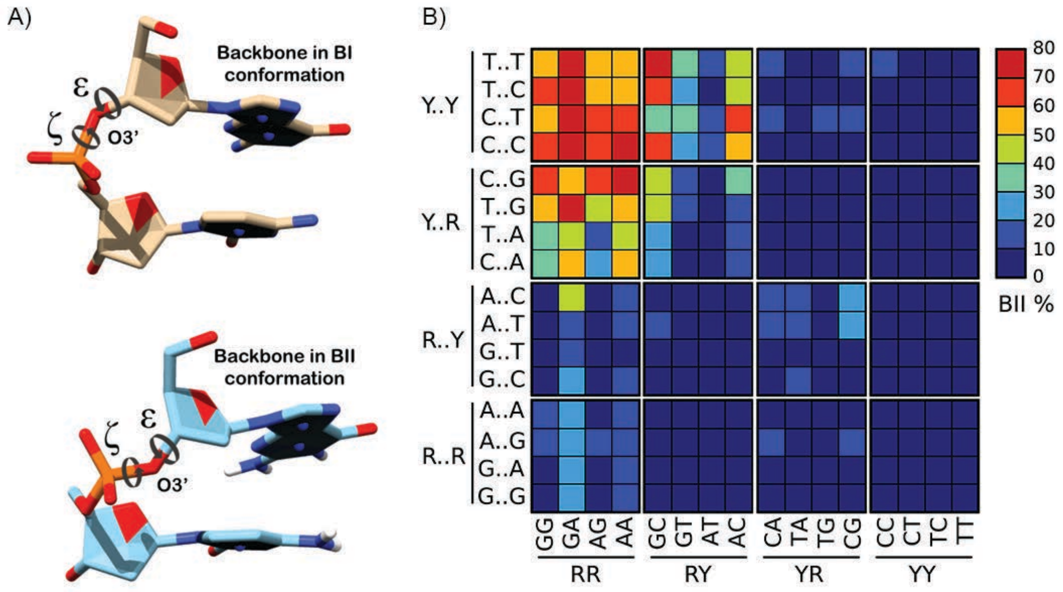


Figure 1. A) Depiction of B-DNA BI and BII conformers resulting from rotations around the ζ and ε torsions. B) Sequence dependence of BII backbone conformations. The percentage occurrence of BII backbone states for the phosphodiester junction at the central base step of each of the 256 possible tetranucleotide 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).

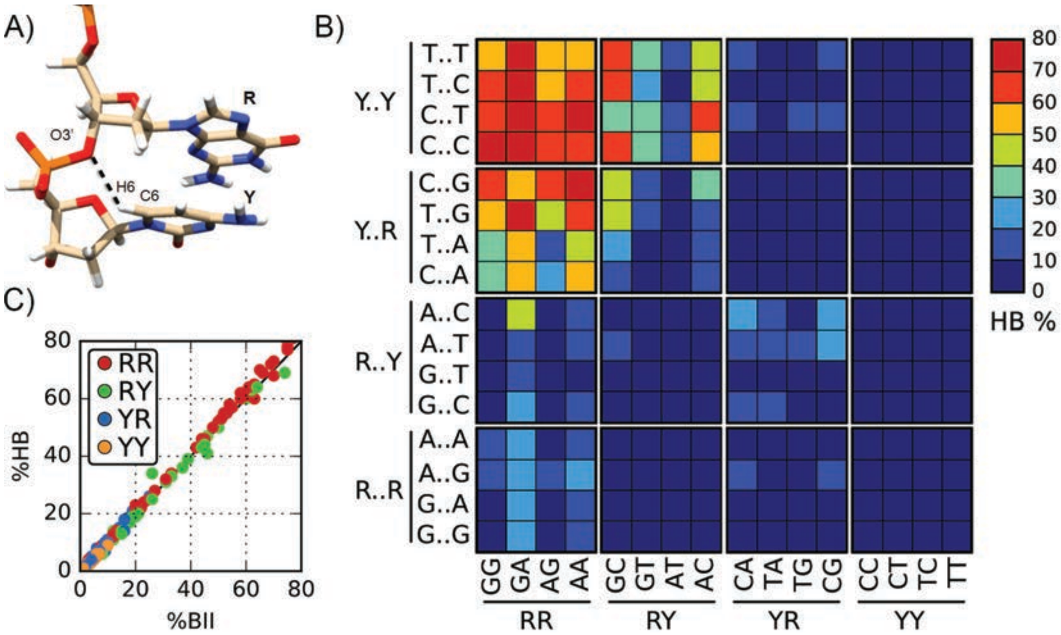


Figure 2. A) Representation of the C6-H6···O3' interaction in an RpY step showing the atoms involved. B) Sequence dependence of C-H···OH-bond formation. The percentage occurrence of either the C6-H6···O3' or the C8-H8···O3' H-bond at the central base step of each of the 256 possible tetranucleotide sequences is shown, 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). C) Correlation between the percentage of BII (%BII, horizontal axis) and of occurrence of formation of the C-H···O H-bonds at the central base step of each of the 256 possible tetranucleotide sequences, color-coded according to base type of the central base step; the correlation coefficient is 0.998.

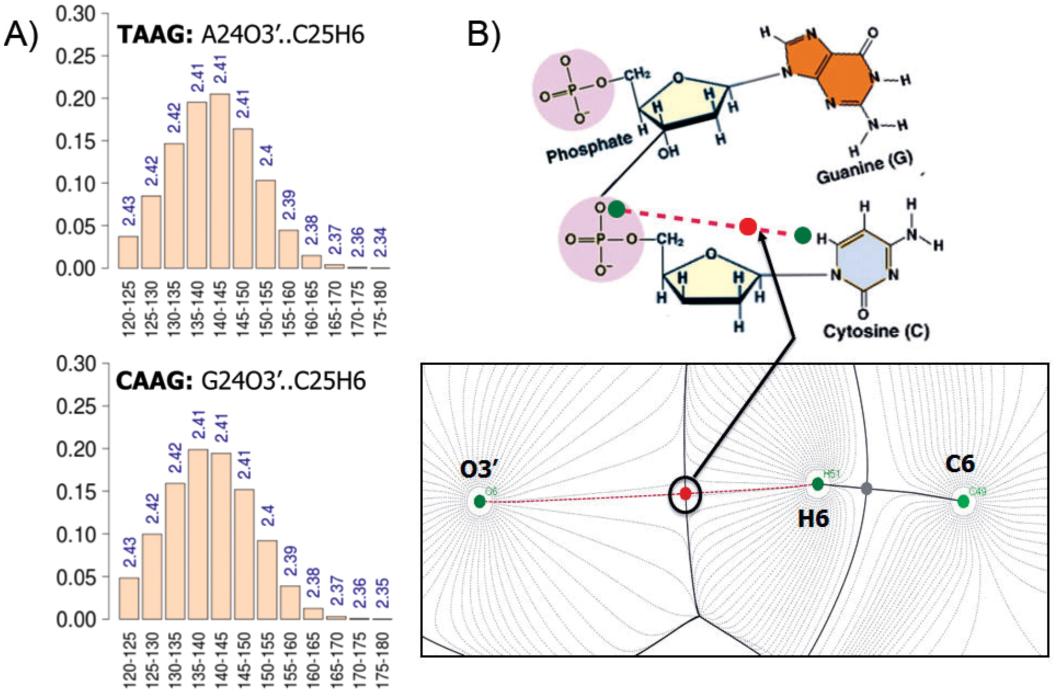


Figure 3. A) Distribution of C6-H6···O3' angles in RpY steps. Structures with H6···O3' bond distances < 2.5 Å and C6-H6···O3' angles > 120° were selected and a histogram was built. The mean distance of the corresponding set is given above each bin bar. The equivalent distributions obtained with the parmBSC1 force field are reported in Figure S1. B) Hydrogen bond AIM analysis for the GpC dinucleotide in the BII conformation. The bond critical point is indicated by a red dot. The nuclear critical points (located at the position of the nuclei) are indicated by green dots, while the basin paths and the gradient field are shown with grey lines. The bond paths, defined by the chosen two-dimensional projection (plane), are shown with red dotted lines.

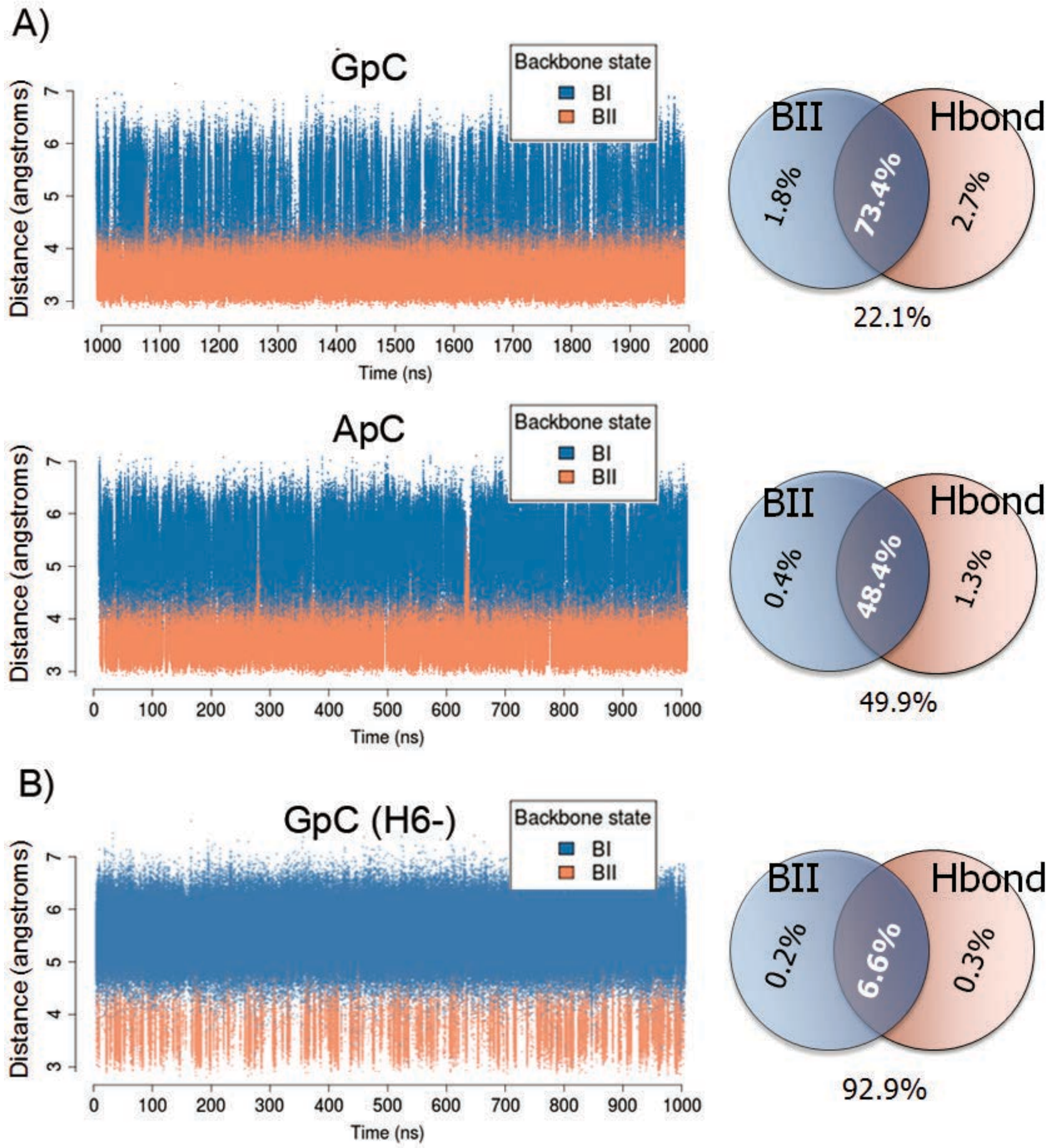
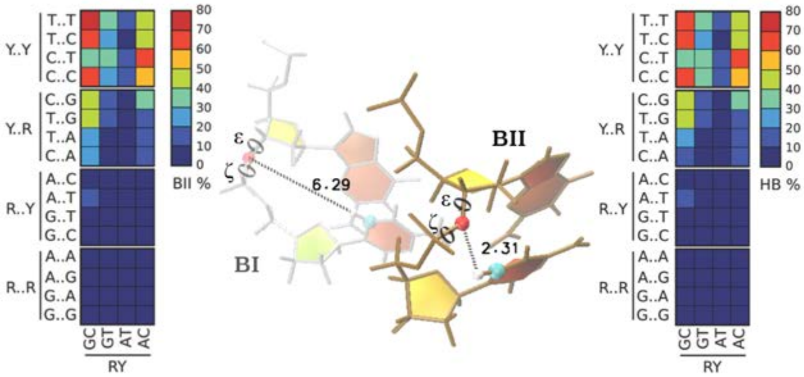


Figure 4. A) Left: Time evolution of C6···O3' distance in two RpY steps (GpC and ApC) colored by the backbone conformation of the step. Right: Venn diagrams of occurrences of BII state and C6-H6···O3' hydrogen bonds at the same RpY steps. An equivalent figure but showing the results obtained with parmBSC1 force field is shown in Figure S2 in the Supporting Data. B) Same than (A) for the simulation without the H6 atom in GpC, labeled CAAG(H6-). The results for TAAG(H6-) are presented in Figure S3.

TOC GRAPHIC


Y..Y
T..T
T..C
C..T
C..C
Y..R
C..G
T..G
T..A
C..A
R..Y
A..C
A..T
G..T
G..C
R..R
A..A
A..G
G..A
G..G
GC
GT
AT
AC
RY
80
70
60
50
40
30
20
10
0
BII %
ε
ζ
6.29
BI
BII
ε
ζ
2.31
HB %

2 The Physical Properties of B-DNA beyond Calladine's rules.

This work is based on the in-depth analysis of the miniABC sequence library using the newest parmbsc1 force-filed, which corrects many known caveats of the former parmbsc0 force-field. The new library was designed to optimize the number of relatively short oligomers needed to cover the complete tetranucleotide space and partly the hexanucleotide space.

Our results determine that helical parameters are transferable (with few exceptions) at the tetranucleotide level and encourage us to make qualitative observations of their variability and inter-dependence that would prove reliable (see Figure 4.1). We focus on uni- versus bi-modality of helical parameters and explain the distinction in relationship with sequence, backbone state and ion environment.

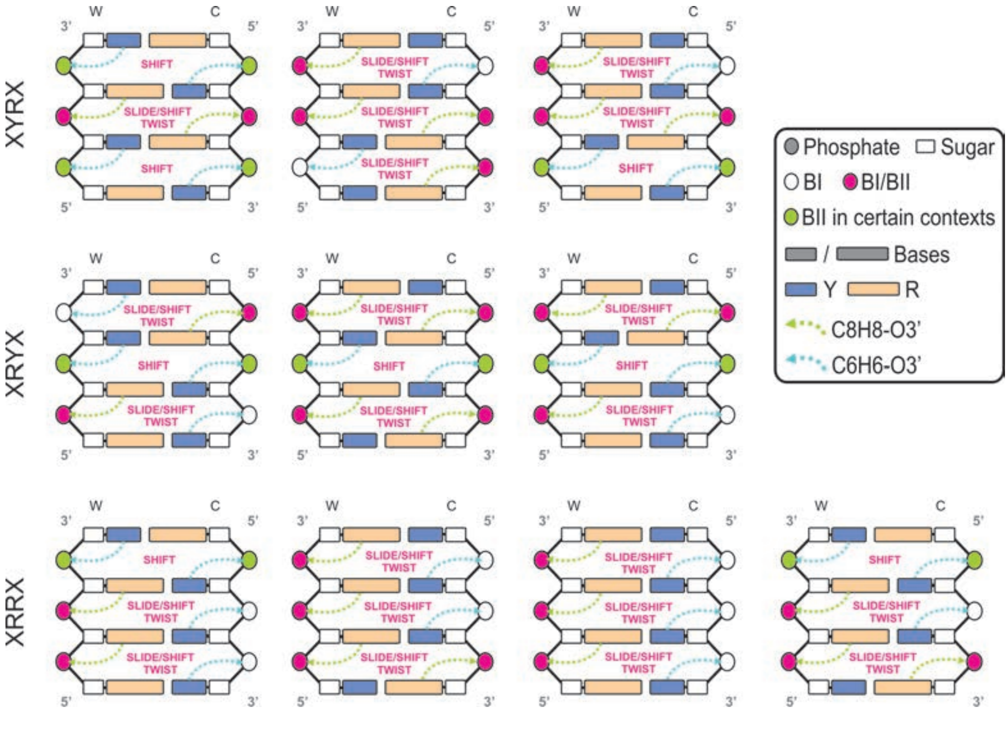


Figure 1 Scheme of the polymorphic landscape of B-DNA at the tetranucleotide level. The 136 unique tetranucleotides were grouped according to purines (R) and pyrimidines (Y), for which only 10 unique combinations exist.

Analysis of data show that B-DNA samples its internal coordinates in a concerted way, generating a complex choreography of conformational transitions that modulates DNA polymorphisms. Therefore many helical parameters and backbone torsions show consistent sequence-specific correlation patterns among the 3 bps of a tetramer. Cations represent an additional player in this negotiation, having the ability to subtly modify the polymorphic landscape of the DNA particularly at the bps level.

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. (submitted)

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.

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's 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 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's earlier qualitative 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 rules presented herein can be transformed into quantitative predictions of the structural features of any canonical 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(13) developed a reduced set of empirical rules, 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, the rules suggested that clashes between bases are avoided by a combination of concerted changes in twist, roll, and slide, as the base pair propeller increases to improve stacking(13). Unfortunately, the accuracy and predictive power of these rules, even in the most recent versions, is limited(1, 14). Attempts to gain more quantitative information were based on the analysis of the variability in local helical parameters in structural databases(15, 16), but to date[1], 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, 17, 18).

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, 19–21). The second major requirement is the availability of accurate force fields, such as the recently developed PARMBSC1(22), which has been shown to represent DNA with a quality indistinguishable from experimental measurements(23). 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(24) 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 geometry of B-DNA, but also those determining coordinated backbone conformational changes, and the correlations between various helical deformations. A new, extended, and comprehensive reformulation of

[1] 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.

Calladine's 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(25–27), 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(22) and PARMBSC0(28) 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(29) and simulated using the *pmemd.cuda* code(30) from AMBER14(29), following the standard ABC protocol(10). Additional details are provided in Suppl. Material. Trajectories are accessible at the BigNAsim server: http://www.multiscalegenomics.eu/MuGVRE/modules/BigNASimMuG/.(31)

Analysis. Trajectories were processed with the *cpptraj*(32) module of the AMBERTOOLS 15 package(29), and the NAFlex server(33) for standard analysis. DNA helical parameters and backbone torsion angles were measured and analyzed with the CURVES+ and CANAL programs(34), 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), while X:X and XX represent a base pair and base-pair step respectively. Bayesian Information Criterion (or BIC)(35, 36) 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(37, 38) 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*. averages and covariances) 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(39–41) model parameters were computed from each of the four MD training libraries used in this work (*i.e*. μABC$_{\mathrm{BSC0}}$-K, miniABC$_{\mathrm{BSC0}}$-K, miniABC$_{\mathrm{BSC1}}$-K, miniABC$_{\mathrm{BSC1}}$-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- 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 Tables A1-A6 in the Appendix), 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 means and covariances for each of the 13 miniABC library sequences for the four different parameter sets extracted from the μABC$_{\mathrm{BSC0}}$-K, miniABC$_{\mathrm{BSC0}}$-K, miniABC$_{\mathrm{BSC1}}$-K, and miniABC$_{\mathrm{BSC1}}$-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 elsewere(9, 23)), 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(43)) 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(44), but most probably too short to lead to isotope exchange signals in NMR experiments(45). 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(46) at an AA step 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(47, 48), but to our knowledge, has not been previously observed in atomistic simulations.

Equilibrium distributions of intra base-pair deformations are close to Harmonic. A BIC analysis was carried out for the distributions of all six of the helical intra base pair parameters at the central base pair in all 32 possible distinct 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 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 pairs, 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 twist also shows clear sequence rules, with A:T pairs having a sizeable negative value when there is an R 5' to the A, while propeller is close to zero for G:C pairs within YGR trinucleotides.

Equilibrium distributions of inter base-pair deformations are frequently strongly anharmonic. Bi-normality (*i.e*. deviation from Gaussianity) in the equilibrium distributions of the inter base-pair 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 base-pair helical distributions (Figure 3 and Suppl. Fig. 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(49)) 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 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-base pair, 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) YR base-pair steps typically have higher than normal slide and roll, ii) RY base-pair steps typically have lower than normal slide and roll, and iii) YY and RR steps have lower than normal tilt values. Any further rules concerning the average values of helical inter base-pair 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(50, 51) and are also detected in ^{31}P NMR spectra(52, 53). 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(54). 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 steps: in general, Na^+ increases the total percentage of the BII state (Figure 4), but reduces its sequence-dependence, in perfect agreement with experimental data(55). 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, 56). 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(57).

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 base-pair 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-helical (base pair) 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 base-pair 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 steps, weaker correlations can be found (depending on the tetranucleotide sequence-environment) in shift-tilt, slide-rise and roll-twist. Finally, YR steps 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 central step (70% of the bimodal inter-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-base pair 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 base-pair step *i+1* and *i-1* (Suppl. Figures S18 and S19), highlighting the subtle mechanical coupling between backbone and base conformations within DNA(57).

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.

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's 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 (averages and covariances) 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 base pair coordinates, and the tetranucleotide level for inter base pair 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 base pair 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. 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 base pair step 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 base pair step, but distant and more complex correlations exist between BI/BII conformational states and the helical conformation of neighbouring steps.
- Helical parameters at a given base pair 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.
- All of these qualitatively extended Calladine rules can now be transformed into quantitative predictions of the structural features of canonical DNA sequences. These 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 http://mmb.pcb.ub.es/webdev/slim/miniABC/public/).
- 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(58) 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 base-pair step MC_DNA coarse grain model, which is coupled to a Monte Carlo algorithm that sample the conformational space (https://mmb.irbbarcelona.org/MCDNA/). Using the extended Calladine's rules presented herein, the backbone and sugar conformational sub-states are predicted and rebuild at atomic resolution, based only on the spontaneous values of inter helical parameters.

ACKNOWLEDGMENTS

2.1 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). The authors thank the Ascona B-DNA Consortium for the trajectories in the standard set (µABC). 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. Neidle S (2008) *Principles of nucleic acid structure* (Elsevier).
2. Fuller W, Forsyth T, Mahendrasingam A (2004) Water-DNA interactions as studied by X-ray and neutron fibre diffraction. *Philos Trans R Soc B Biol Sci* 359(1448):1237–1248.
3. Hartmann B, Piazzola D, Lavery R (1993) BI-BII transitions in B-DNA. *Nucleic Acids Res* 21(3):561–8.

4. Balaceanu A, et al. (2017) The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J Phys Chem Lett* 8(1):21–28.
5. Dans PD, et al. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 42(18):11304–11320.
6. Zgarbová M, et al. (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(2):275–287.
7. Várnai P, Djuranovic D, Lavery R, Hartmann B (2002) Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res* 30(24):5398–406.
8. Pérez A, Luque FJ, Orozco M (2007) Dynamics of B-DNA on the Microsecond Time Scale. *J Am Chem Soc* 129(47):14739–14745.
9. Dans PD, et al. (2016) Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res* 44(9):4052–4066.
10. Pasi M, et al. (2014) μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 42(19):12272–12283.
11. Rohs R, et al. (2010) Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem* 79:233–69.
12. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* 5(11):789–796.
13. Calladine CR (1982) Mechanics of sequence-dependent stacking of bases in B-DNA. *J Mol Biol* 161(2):343–52.
14. Cheatham TE, Brooks BR, Kollman PA, III (2001) Molecular modeling of nucleic acid structure. *Curr Protoc nucleic acid Chem* Chapter 7:Unit 7.5.
15. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 95(19):11163–8.
16. Dans PD, Pérez A, Faustino I, Lavery R, Orozco M (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 40(21):10668–10678.
17. Pérez A, Luque FJ, Orozco M (2012) Frontiers in Molecular Dynamics Simulations of DNA. *Acc Chem Res* 45(2):196–205.
18. Dans PD, Walther J, Gómez H, Orozco M (2016) Multiscale simulation of DNA. *Curr Opin Struct Biol* 37:29–45.
19. Beveridge DL, 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(6):3799–3813.
20. Dixit SB, 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(6):3721–3740.
21. Lavery R, 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(1):299–313.
22. Ivani I, et al. (2015) Parmbsc1: A refined force field for DNA simulations. *Nat Methods* 13(1):55–58.

23. Dans PD, et al. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 45(7):4217–4230.
24. Pérez A, et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 92(11):3817–29.
25. Mintseris J, Eisen MB (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC Bioinformatics* 7(1):429.
26. Medvedev P, Georgiou K, Myers G, Brudno M Computability of Models for Sequence Assembly. *Algorithms in Bioinformatics* (Springer Berlin Heidelberg, Berlin, Heidelberg), pp 289–301.
27. Orenstein Y, Shamir R (2013) Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers. *Bioinformatics* 29(13):i71-9.
28. Pérez A, et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 92(11):3817–29.
29. Case D, et al. (2014) AMBER14. Available at: ambermd.org.
30. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 9(9):3878–3888.
31. Hospital A, et al. (2016) BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res* 44(D1):D272–D278.
32. Roe DR, Cheatham TE (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 9(7):3084–3095.
33. Hospital A, et al. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res* 41(W1):W47–W55.
34. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* 37(17):5917–5929.
35. Schwarz G (1978) Estimating the Dimension of a Model. *Ann Stat* 6(2):461–464.
36. Kass RE, Raftery AE (1995) Bayes Factors. *J Am Stat Assoc* 90(430):773–795.
37. Schilling MF, Watkins AE, Watkins W (2002) Is Human Height Bimodal? *Am Stat* 56(3):223–229.
38. de Helguero F (1904) Sui Massimi Delle Curve Dimorfiche. *Biometrika* 3(1):84.
39. Petkevičiūtė D, Pasi M, Gonzalez O, Maddocks JH (2014) cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res* 42(20):e153.
40. Gonzalez O, Pasi M, Petkevičiūtė D, Glowacki J, Maddocks JH (2017) Absolute versus Relative Entropy Parameter Estimation in a Coarse-Grain Model of DNA. *Multiscale Model Simul* 15(3):1073–1107.
41. Gonzalez O, Petkevičiūtė D, Maddocks JH (2013) A sequence-dependent rigid-base model of DNA. *J Chem Phys* 138(5):055102.
42. Glass G V, Hopkins KD (2008) *Statistical methods in education and*

psychology (Boston : Allyn & Bacon). 3rd ed.

43. Mitchell JS, Glowacki J, Grandchamp AE, Manning RS, Maddocks JH (2017) Sequence-Dependent Persistence Lengths of DNA. *J Chem Theory Comput* 13(4):1539–1555.
44. Andreatta D, Sen S, Pérez Lustres JL, Kovalenko SA, Ernsting NP, Murphy CJ, Coleman RS, Berg MA (2006) Ultrafast Dynamics in DNA: "Fraying" at the End of the Helix. *J Am Chem Soc* 128:6885-6892.
45. Priyakumar UD, MacKerell J. AD (2005) NMR Imino Proton Exchange Experiments on Duplex DNA Primarily Monitor the Opening of Purine Bases. *J Am Chem Soc* 128:678-679.
46. Pasi M, Lavery R (2016) Structure and dynamics of DNA loops on nucleosomes studied with atomistic, microsecond-scale molecular dynamics. *Nucleic Acids Res* 44(11):5450–5456.
47. Altan-Bonnet G, Libchaber A, Krichevsky O (2003) Bubble Dynamics in Double-Stranded DNA. *Phys Rev Lett* 90(13):138101.
48. Zeida A, MacHado MR, Dans PD, Pantano S (2012) Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* 86(2):1–7.
49. Maehigashi T, et al. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res* 40(8):3714–3722.
50. Djuranovic D, Hartmann B (2003) Conformational Characteristics and Correlations in Crystal Structures of Nucleic Acid Oligonucleotides: Evidence for Sub-states. *J Biomol Struct Dyn* 20(6):771–788.
51. Madhumalar A, Bansal M (2005) Sequence Preference for BI/BII Conformations in DNA: MD and Crystal Structure Data Analysis. *J Biomol Struct Dyn* 23(1):13–27.
52. Heddi B, Foloppe N, Bouchemal N, Hantz E, Hartmann B (2006) Quantification of DNA BI/BII Backbone States in Solution. Implications for DNA Overall Structure and Recognition. *J Am Chem Soc* 128(28):9170–9177.
53. Tian Y, et al. (2009) ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J Phys Chem B* 113(9):2596–2603.
54. Ben Imeddourene A, et al. (2015) Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLOS Comput Biol* 11(12):e1004631.
55. Heddi B, Foloppe N, Oguey C, Hartmann B (2008) Importance of Accurate DNA Structures in Solution: The Jun–Fos Model. *J Mol Biol* 382(4):956–970.
56. Dršata T, et al. (2013) Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J Chem Theory Comput* 9(1):707–721.
57. Svozil D, Kalina J, Omelka M, Schneider B (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res* 36(11):3690–3706.
58. De Bruin L, Maddocks JH (2018) cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res*. doi:10.1093/nar/gky351.

FIGURES

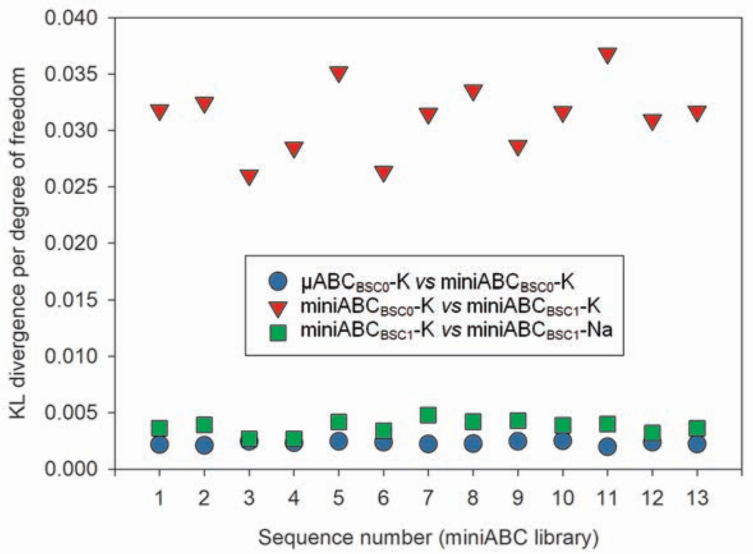


Figure 1. Symmetric Kullback-Leibler divergence per degree of freedom between Gaussian distributions, which is a combined measure of differences in values of first and second moments, for each of the thirteen oligomers in the miniABC training library, but for cgDNA model parameter sets fit to different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

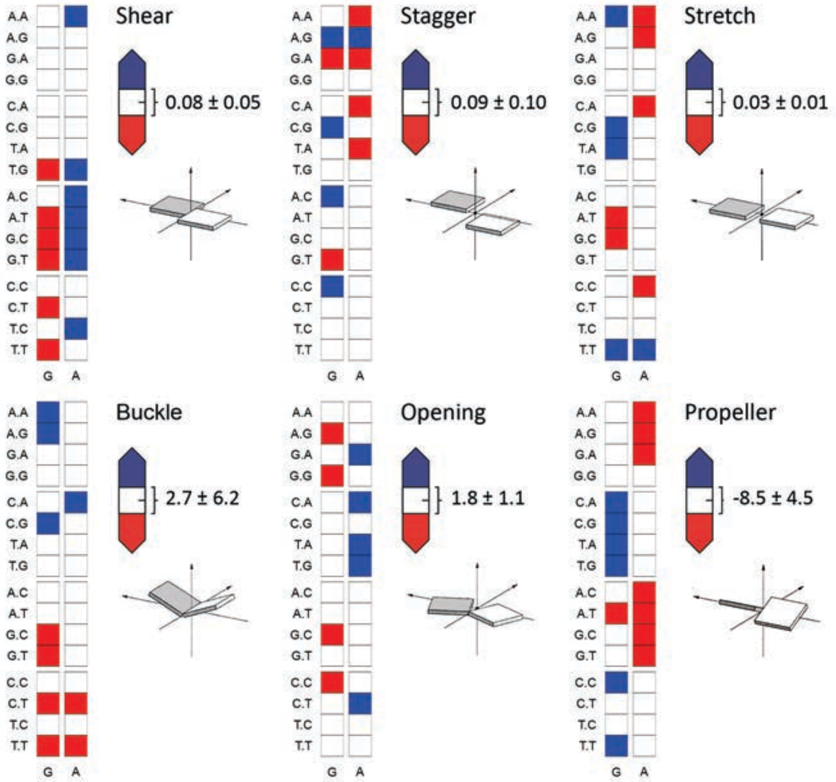


Figure 2. Average values of intra base-pair helical coordinates of the central base-pair in all 32 distinct trinucleotide sequence contexts. Results obtained from the miniABC$_{BSC1}$-K simulations. The global averages are over all sequence contexts and standard deviations reflect the variation among trinucleotide contexts.

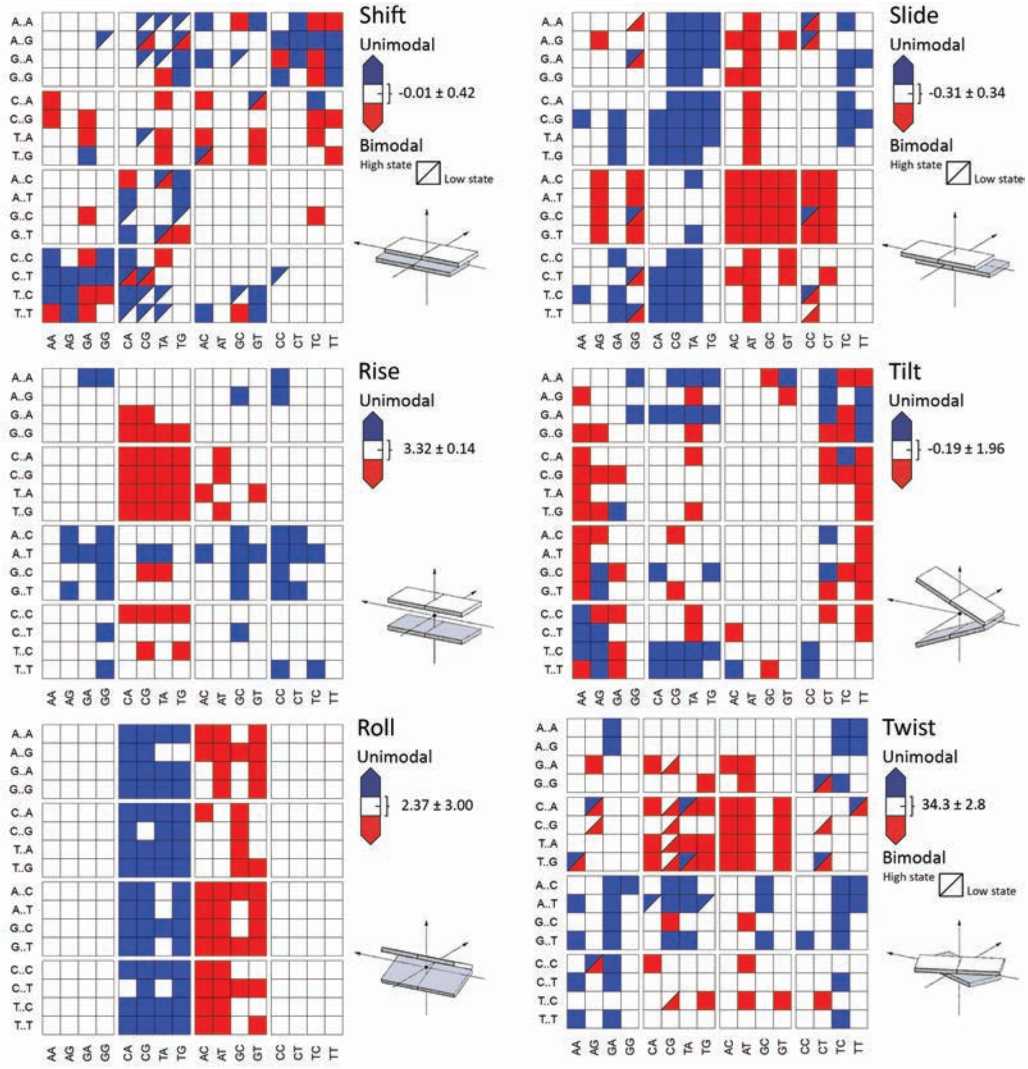


Figure 3. Average values of inter base-pair, or junction or step, helical coordinates for the central junction set in all possible 256 tetranucleotide contexts. Results obtained from the miniABC$_{BSC1}$-K simulations. Tetranucleotides classified as bimodal (half-square) are polymorphic (*i.e.* they sample two clear conformational sub-states). The global averages, exhibited 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.

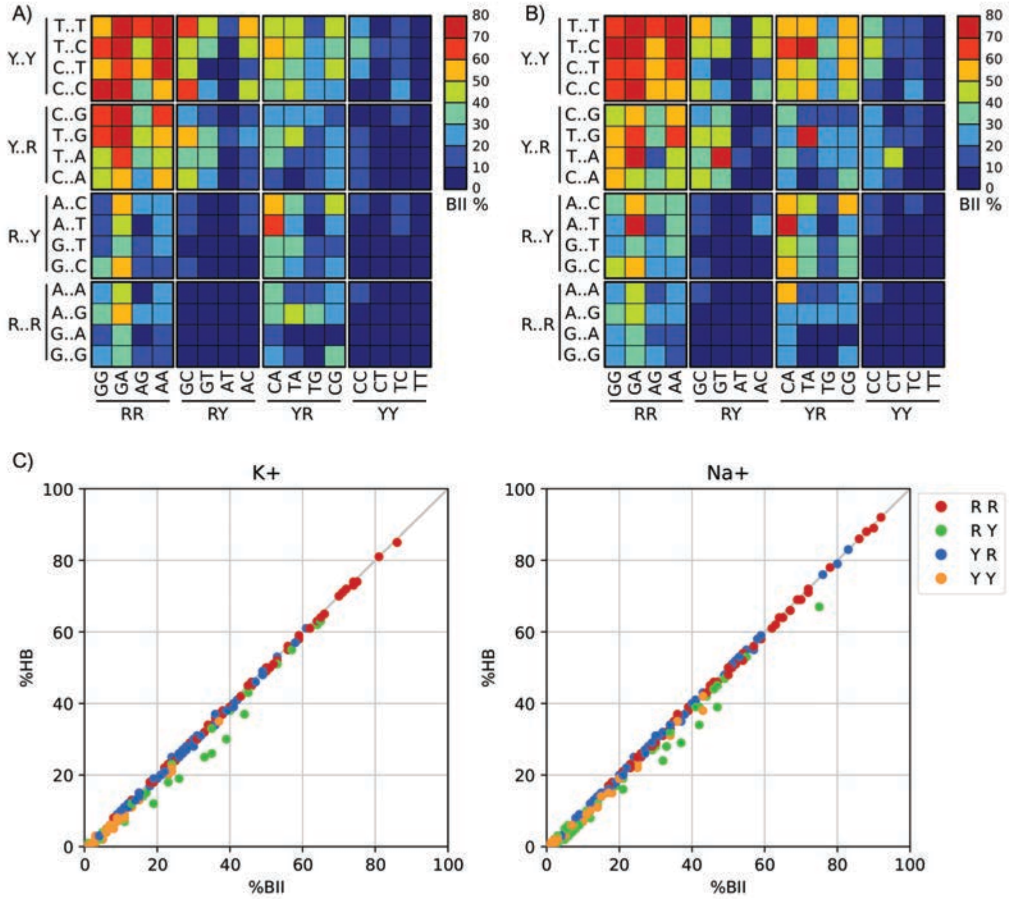


Figure 4. Sequence dependence of BII backbone conformations comparing K$^+$ and Na$^+$. A) miniABC$_{BSC1}$-K BII percentages. B) miniABC$_{BSC1}$-Na BII percentages. C) Correlation between the percentage of BII (%BII, horizontal axis) and of occurrence of formation of the C–H···O H-bonds (%HB, vertical 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.

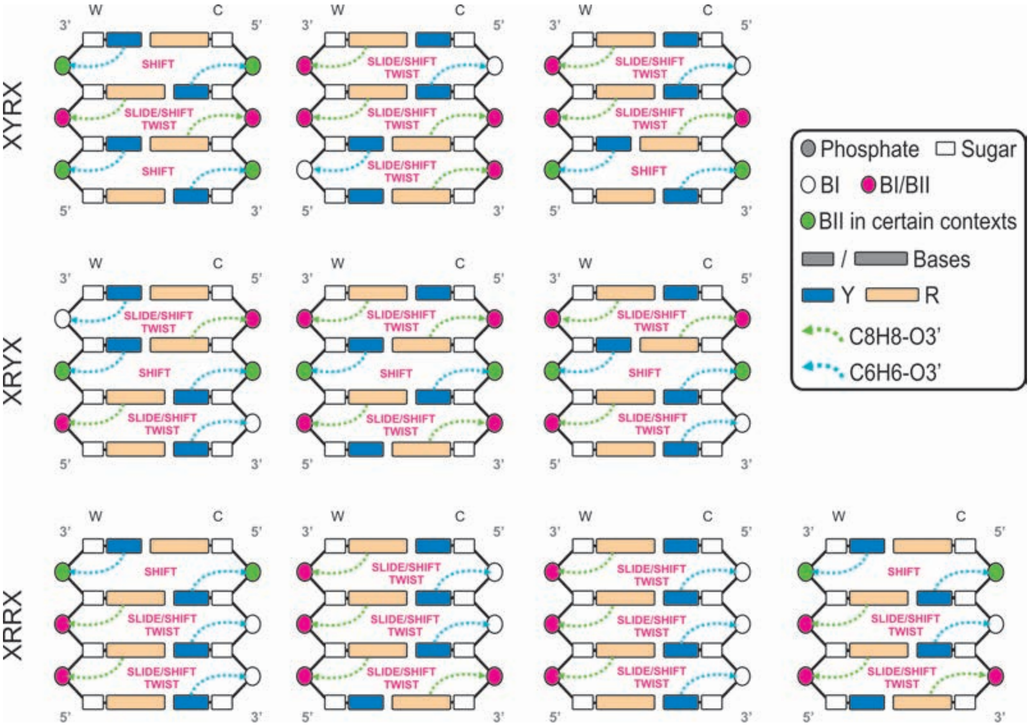


Figure 5. Schema of the polymorphic, or multi-well, landscape exhibited by B-DNA at the tetranucleotide level expressed in the purine (R)/pyrimidine (Y) alphabet, for which only 10 distinct combinations exist, but which still distinguish all possible behaviours. The only helical coordinates that can exhibit 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).

Bibliography for Chapter IV

1. Galindo-Murillo R, Robertson JC, Zgarbová M *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* 2016;**12**:4114–27.

2. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:gkw1355.

3. Galindo-Murillo R, Roe DR, Cheatham TE. On the absence of intrahelical DNA dynamics on the μs to ms timescale. *Nat Commun* 2014;**5**:5152.

4. Pérez A, Luque FJ, Orozco M. Frontiers in Molecular Dynamics Simulations of DNA. *Acc Chem Res* 2012;**45**:196–205.

5. Dršata T, Pérez A, Orozco M *et al.* Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J Chem Theory Comput* 2013;**9**:707–21.

6. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.

7. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.

8. Ben Imeddourene A, Elbahnsi A, Guéroult M *et al.* Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. MacKerell A (ed.). *PLOS Comput Biol* 2015;**11**:e1004631.

9. Whelan DR, Hiscox TJ, Rood JI *et al.* Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation. *J R Soc Interface* 2014;**11**:20140454.

10. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 1981;**52**:7182–90.

11. Rohs R, Jin X, West SM *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem* 2010;**79**:233–69.

12. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

13. Ivani I, Dans PD, Noy A *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat Methods* 2016;**13**:55–8.

14. Beveridge DL, Barreiro G, Suzie Byun K *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* 2004;**87**:3799–813.

15. Lavery R, Zakrzewska K, Beveridge D *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* 2010;**38**:299–313.

16. Dixit SB, Beveridge DL, Case DA *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* 2005;**89**:3721–40.

CHAPTER V | Information Transfer Through the DNA

Previous chapters show that DNA sequence determines its overall structural properties and that sequence-dependent effects are of great importance, being exploited by the cell to regulate DNA function [1–7]. Therefore, the structural diversity of DNA needs to be described considering at least the first neighbor bases of each bps. Generally, the description of sequence effects to this level (the tetranucleotide) is sufficient, since sensitivity to the context diminishes rapidly with sequence length. Almost all mixed-sequence DNA tends to B-DNA and small distortions at a bps are quickly corrected at adjacent steps. However, there are certain sequence elements that have an increased predisposition to respond to long-range sequence effects in a significant way. Typically, bimodality and low stability of certain steps are closely related to sequence context dependence. Additionally, local distortions of different magnitudes imposed to the DNA by the binding of ligands have sometimes a long range of compensatory structural responses that should also be quantified and clarified.

Although some general observations can be made from inspecting the database of crystal and NMR DNA structures, the experimental data is fragmentary and inappropriate for exploring the DNA sequence space, even at the tetramer level, and much more so beyond it. Current, state-of-the-art MD simulations are particularly suitable for this task and have already been shown to provide a plausible description of sequence effects up to the tetramer level [2,8–11]. Interestingly, the results obtained in such studies, including our own analysis of the miniABC set (see previous Chapter 4.2), allow us to further pinpoint several cases of tetramers with a unusual behavior, such as low stability/high flexibility, predisposition to bimodality and high sensitivity to sequence context. These cases are the exception and not the rule, with only a small percentage of 3-5 % showing clear bimodal behavior (see Chapter 5.1), but might be important to explain local flexibility of certain DNA motifs.

We decided to investigate higher-than-tetramer sequence effects in the particular case of the d(CpTpApG)$_2$ tetranucleotide (from here on CTAG) a sequence showing unusual flexibility in simulations deposited in our BigNAsim database [12] and in miniABC trajectories. Initially it was unclear whether the unusual behavior of this sequence was an artifact of the limited sampling (μs) or highlighted a more complex long-range effect on the properties of DNA. Systematic analysis and study of many replicas of the same tetramer in different context convinced us that the strange behavior of this sequence detected in simulations was not an equilibration artifact. We established non-negligible effects of hexameric and even octameric sequence context. Exploring the details of the intricate sequence-dependent mechanisms that account for this long-range sequence modulation of base-pair dynamics, we connect it to a sequential domino effect of backbone equilibrium, electrostatics and solvent influence. Based on such findings, which have been assembled into the manuscript *Long-Range Effects Modulate Helical Properties of some DNA Dinucleotide Pairs*, we were able to envision and formulate a complex mechanism of information transfer across DNA through coordinated backbone movements. The backbone then modulates base step geometry affecting several helical parameters along the sequence.

As an extension of the knowledge obtained on sequence context effects and long-range information transfer through the DNA, I chose to focus in a separate work on the cooperative protein binding on the DNA. At a conceptual level, this type of cooperativity means that the binding affinity of two different proteins is enhanced and stabilized by DNA in the ternary complex [3,13–16]. Needless to say, cooperativity is a key component of binding specificity [3], by modulating the efficiency of binding even at low concentrations of the protein partners.

In most case cooperativity involves a direct interaction between the two proteins, but there are cases where interacting proteins are too far to establish any energetic interactions. One of this systems for which long-range cooperative binding has been well studied is the ternary complex comprising of the effector protein BAMHI type II Endonuclease [17], the secondary binder glucocorticoid receptor DNA-binding domain (GRDBD) [18] and DNA. This is the system that I focused my attention on as well in the study *Allosterism and signal transfer in DNA*.

In the BAMHI-DNA-GRDBD ternary system the binding sites of the two molecules are physically separated. This means that cooperativity implies that the DNA is able to transfer information from one binding site to the other. The structural parameter of most relevance to the binding of these two proteins is the major groove width. We confirm that cooperative protein binding is related to protein-induced changes in the flexibility of the major groove [16,19]. The distance and relative orientation of the two binding sites can lead to differences in conformational response upon binding of the effector protein. Interestingly, our findings additionally suggest that the origin of cooperativity in this system differs from conventional allosteric interactions in that the binding of the first protein (BAMHI) predisposes the dynamics rather than structure of the molecule to accommodating the second (GRDBD). It is actually the entropic part of the free energy that plays a dominant role in the cooperative nature of the binding process. This is an example of “allosteric communication without conformational change” originally suggested by Cooper and Dryden, previously demonstrated for small ligands by the group and coworkers [20].

1 Long-Range Effects Modulate Helical Properties of some DNA Dinucleotide Pairs

In this work we carry out a detailed analysis of CTAG in 40 different sequence contexts. We focus on this specific tetranucleotide sequence based on evidence of its high sequence dependence and polymorphism from a number of trajectories run in the group and in our miniABC dataset. We find evidence of intrinsic multi-modality of the individual trajectories in three helical bps parameters (shift, slide and twist). Shift distribution is tri-modal, while twist and slide distributions are bi-modal but only 4 specific combinations of substates are possible. Additionally, large differences in the distributions of these internal coordinates of the d(TpA) step are seen with sequence variation.

We then go on to explain how information travels to allow the d(TpA) step to "feel" its sequence environment based on the concerted movements of the backbone and bases, which are also coupled to the formation of the unconventional h-bonds described in Section 4.1 of Chapter IV and a small contribution from the known sugar puckering flexibility. We detect this communication of mechanical information up to the octamer level, which means over almost one helix turn. We further examine the remote effects (beyond hexamer) in more detail, pointing out which types of sequences are more susceptible to transmitting information and at which steps communication vanishes.

Definite experimental validation is impossible from a mere analysis of resolved structures in the database because of the averaging out of such subtle dynamic effects. However, distributions of helical parameters observed in structures containing the CTAG tetramer provide an indirect, but strong support to the 4-state model of TpA dynamic in CTAG. Lastly, we perform a genomic analysis of the occurrences of this tetramer in different organisms and observe a clear depletion compared to other tetramers, which might be connected to its high flexibility.

Publication:

Alexandra Balaceanu, Diana Buitrago, Jurgen Walther, Pablo D. Dans and Modesto Orozco. Long-Range Effects Modulate Helical Properties of some DNA Dinucleotide Pairs (in preparation).

LONG RANGE EFFECTS MODULATE HELICAL PROPERTIES OF SOME DNA DINUCLEOTIDE PAIRS

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.

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) tetramer. Contrary to the assumption of near neighbors (dimer-model) and next-to-nearest neighbors (tetramer-model) the properties of the central d(TpA) step change quite significantly dependent on the hexamer context and in a few cases are modulated by remote neighbors (beyond the hexamer level). Our results highlight the existence of previously undescribed mechanisms for the long-range transmission of structural information into the DNA.

INTRODUCTION

Early structural models of DNA derived from fiber 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 DNA as a flexible and polymorphic molecule, 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 [11–18].

Attempts to determine the principles relating sequence and structure originated in the eighties when by processing the scarce experimental data available Chris R. Calladine [19] developed a series of heuristic rules relating sequence with some structural characteristics of DNA. In the late nineties [20] Olson and Zhurkin 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 nearest neighbor model must be assumed, i.e. the helical geometry can be decomposed at the bps level; iii) structural 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, 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 multi-nanosecond trajectories led to artefactual results due to a previously unknown error of the most used force-field at that time [21]. A newer force-field [22] and higher computational capabilities provided descriptions of DNA properties that were more reasonable, but still far from the required accuracy [12,23,24]. The availability of the highly-accurate PARMBSC1 force-field [25,26] and the development of new MD codes taking advantage of a new generation of computers [27–30] 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 [31–34] revealed two major findings which challenged current models of DNA flexibility. First, the nearest neighbor model is insufficient to describe DNA flexibility, as the variability in bps parameters depending on tetramer environment can be more pronounced than the variability found when comparing different bps for a given tetramer context. Second, several distributions of tetrameric helical parameters are not normal and a part of such non-normal distributions are in fact multi-modal, which means that the physical properties of such tetramers cannot be represented by a single set of elastic parameters (equilibrium values and associated stiffness). Analysis of data revealed that the changes between sub-states happen towards a series of coordinated changes along the backbone [17,34,35], where unusual H-bond interactions and subtle changes in the solvent environment play a key role [18,36]. The analysis of ABC data and of additional trajectories stored in our BigNaSim database [21] suggested that a tetramer-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) tetramer (in the following CTAG): a very polymorphic state for which results were significantly different depending on the simulation.

We present here a detailed analysis of CTAG in different sequence contexts. Results demonstrate that long-range effects modulate the geometrical properties of the central d(TpA) step. Such long-range effects are very visible at the hexamer level, but quite surprisingly extend beyond this level, indicating the existence of a complex mechanism of information transfer across DNA through coordinated backbone movements.

METHODS

The choice of sequences and the simulation details. The systematic study of sequence dependent effects beyond the tetramer level have been to date impossible, due to the huge number of sequences that need to be considered. For example, the study of all hexamers would require the analysis of 2,080 sequences, and to consider all octamers 32,826 sequence combinations are needed. Fortunately, the analysis of ABC- simulations where tetramers appear in different molecular environments suggests that sequences effects beyond the tetramer are rare, and if they exist, are localized in certain ultra-flexible sequences. We focuss our interest here in one of the most flexible tetramer CTAG, for which comparison of ABC trajectories and those found in BigNAsim suggest the existence of potential hexamer dependences. Thus, we built a library of 40 different sequences covering the entire hexamer space (XpCpTpApGpX) as well as all possible pyrimidine(Y)/purine(R) combinations at the octamer level in several repeats (see

Supp. Methods). All the sequences were prepared using the leap module of AMBERTOOLS 16 [37] and standard ABC protocol [34]. Accordingly, systems were built from Arnott's parameters, neutralizing the DNA with monovalent 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 [31,32]. Water was represented with the SCP/E model [38], Smith-Dang parameters were used for ions [39–41] and the recent PARMBSC1 force-field was considered to represent nucleic acids interactions [25]. Trajectories (collected in the NPT ensemble T=298 K, P=1 atm.) were extended from 0.5 μs to up to 10 μs. All simulations were performed with the pmemd.cuda code using periodic boundary conditions and Particle Mesh Ewald [42,43]. Movements of hydrogen atoms were annihilated using SHAKE [44], which allowed us the use of a 2 fs integration step. All trajectories collected here are accessible through the MuG BigNAsim portal [44]: http://www.multiscalegenomics.eu/MuGVRE/modules/BigNASimMuG/.

Analysis. Standard analysis were done using *cpptraj* module of the AMBERTOOLS 16 package [37], the NAFlex server [44] CURVES+ and CANAL programs [45], following the standard ABC-conventions [34]. 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 [46,47] and Helguerro's theorem [48] as described elsewhere [12]. Classification of the torsional states of the different rotatable bonds in the DNA backbone was done using standard criteria [49]. Correlations between different torsions were determined by circular correlation analysis (see Suppl. Methods for additional details). 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 [50,51] in the helical space of the central d(TpA) step after Lankas' normalization of the different rotational and translational degrees of freedom [52]. The essential dynamics of the central d(TpA) step is defined by three eigenvectors (explaining 60% of variance), which are defined by four bps deformations (shift, tilt, roll and twist) and four deformations at the pairing (buckle and propeller twist) of the two bases (dT and dA) composing the central d(TpA) step. The distributions of the four informative bps deformations were subjected to detailed analysis (see Suppl. Method 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 tetramer) were done using symmetrized Kullback-Leibler (KL) divergences [52] followed by hierarchical cluster analysis using Ward's clustering criterion [53], where the dissimilarities are squared before cluster updating [54], using as descriptive variable the 8 distinguished helical variables detected by the PCA of the meta-trajectory (see Suppl. Methods for additional details). Molecular interaction potentials of the different duplexes were computed using our MIP program [54] implementing linearized solutions to the Poisson Boltzmann equation and Na^+ as a probe particle (see Suppl. Methods for additional details). Stacking and hydrogen bonding were followed by geometrical and energetic criteria for both the dimer and the tetramer, as described in detail in Supp. Methods. Structural database analysis was done using all DNA structures containing the CTAG tetramer. Genomic analysis was done to determine the prevalence of the CTAG tetramer in different wild type genomes and its resilience to mutation. Genomes of *H. sapiens* (hg19), *E. coli* (NC_000913.3) and *S. cerevisiae* (sacCer3) were analysed. Occurrences of this tetramer were then mapped, using Homer software [54], 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 tetramer along the genome in 30 different cancer types (data from [54]) was determined.

RESULTS AND DISCUSSION

The CTAG shows a dramatic and complex polymorphism. We collected trajectories for 40 oligonucleotides all of them containing the CTAG tetramer in a central position (see Methods and Supp. Table S1), all of them were stable, sampling structures that fit well in the B-like double helical conformation. As suggested by the analysis of ABC-simulations and of trajectories deposited in BigNAsim, CTAG is highly polymorphic as shown in clear bimodal distributions of some helical parameters. To check that the bimodalities are not artefacts of limited sampling we extended trajectories for selected tetramers to 10 μs regime, tracing the changes in the distribution of helical parameters. The good convergence found in Supp. Figure 1 support the robustness of our results and suggest a fast dynamic of interchange of the different states (see below).

In order to obtain a global average picture of the conformational space accessible to the CTAG tetramer 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 emerged as determinant to explain 60% of the variance in the meta-trajectory: roll, twist, shift and slide. As seen in the BIC analysis summarized in Figure 1, deviations from Gaussian distribution are the main responsible for the polymorphism 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 are normal, 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. Roll distributions were unimodal in all cases, but the position of the peak were displaced towards slightly higher values when the central tetramer is surrounded by R at 5' and Y at 3' (i.e. RCTAGY hexamers), 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 is clear for individual sequences (see Figure 2 and Suppl. Figure S2), with the different substates being sampled in a fast equilibrium along the time scale of the simulations (see examples in Supp. 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 4 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 (Figure 4) by 3 main states: 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 polymorphism is nearly impossible as experimental structures are always averages (i.e. assume a normal unimodal distribution). However, plotting the scarce experimental data available for the CTAG tetramer on the two-dimensional population plots (shift-twist, shift-slide and twist-slide) derived from meta-trajectories provide 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 high twist. Finally the twist-slide plot show only two regions of high probability consistent with the same slide/twist correlation found experimentally (see Figure 3).

Hexamer dependence in central d(TpA) conformation. All the sequences studied here correspond to the same tetramer, so we could expect a similar distribution of helical parameters at the central d(TpA) step. However, this is not the case as shown in selected examples in Supp. Figure S2, where large differences in the distributions of helical coordinates for the d(TpA) step appears. To analyse this in more detail we perform KL analysis of the 40 trajectories in the 6-dimensional 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 relational dendogram (Figure 5), which clearly show the presence of at least two major clusters. The first one is populated mainly by sequences where the central tetramer 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 in three different sub-clusters, two of which are formed almost exclusively of sequences where the central tetrad 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 Supp. Figure S4, which demonstrate that the hexamer content has a non-negligible role in defining the properties of the central d(TpA) step in the CTAG tetramer, a clear exception of the next-to-nearest-neighbour model.

The existence of long-range effects imply that the motion of the central TA step must be somehow connected to the distant base pairs. Mechanical information should travel from one site to the other to allow the TA 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 TA was behaving as the better well-known YR step: d(CpG) [18,34,36,55]. The two possible twist substates (HT/LT) at the AT step, were connected to the backbone BI/BII polymorphism at the next GA junction (note that BI/BII inter-conversion is mainly governed by the ζ torsion). Furthermore, the BI/BII polymorphism at GA is possible due to the formation of the intra C8H8-O3' hbond and the shift polymorphism in the same junction (Figure 6A, and B) [36]. Similar results were found if looking to the correlation of twist at the central TA step with the bps at the 5'-side (CT). It is then clear that the main backbone polymorphism (BI/BII) is linked to the base polymorphisms, mainly to shift and twist (Supp. Table 2) up to the hexamer level. The information travels through successive backbone and base polymorphisms which are limited to some specific substates due to DNA's crankshaft motion (Supp. Table 2). This concerted movement of some 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 hbonds are the main responsible for the information transfer between the backbone and the base [36] (with perhaps a small contribution from the known sugar puckering flexibility, see Supp. Table S2), both backbone and base polymorphisms can be followed by looking only to the formation of those C8H8-O3' hbonds in RR and YR steps, or C6H6-O3' hbonds in RY and YY steps. The correlated/anti-correlated formation of these hbonds away from the central TA step clearly explains the transfer of mechanical information up to the hexamer level, and also up to the octamer level depending on the sequence (Figure 6C)

Structural information travel beyond the hexamer level. Sequences created here cover all the hexamer space with some redundancy that allowed us to check for some remote effects beyond the hexamer. Quite surprisingly, such effects are clearly visible already in dendogram shown in Figure 5, where sequences showing the same hexamer sequence appear in two very different branches. This is the case of RCTAGY hexamers which populate two distinct clusters, with YR..YR octamers having the tendency to make an exception and display remarkably different behaviour than other R..Y sequences. In addition, although there is a clear cluster containing R..R hexamers, there are a few exceptions where an R..R in specific octamers leads to significant changes in the population of the preferred regions of the helical space at the central d(TpA) step, namely TA..GG, GG..AC and GG..GG. In contrast, when flanked by Y..R, the central TpA step seem to maintain a very consistent and stable behaviour (see Figure 7).

In particular, and focusing on selected cases, YpCpTpApGpR sequences (Y..R) are very stable at the hexamer level where the sequence effects seem to be completely dampened down. They are all in the same cluster in the dendogram of Figure 5 and display consistent distributions in all multimodal 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. R..Y hexamers have two very distinct types of behavior, depending quite clearly on the flanking base of the octamer. RR..YY octamers tend to populate zero shift states and have equal populations of high/low twist as well as of negative/positive slide. The YR..YR octamers are strikingly different. They have a strong preference for positive shift and do not visit low twist or negative slide very often. This is probably due to a domino effect of hbond proclivity that does not allow a BII backbone at the 3’ side and compensates by shifting towards the major groove. Finally, R..R hexamers can also show variability, but only in some particular cases, when instead of shifting towards the minor groove as typical, they shift towards the major, maintaining similar distributions of twist and slide. (It is hard to justify why this happens based on our data, but 2 of the 3 cases where an R..R hexamer is assigned to a different group than its own have XR..GG and none of the R..R hexamers in the fourth group has a GG at the 3’ side).

Data mining of structural databases and genomic implications. We analyze the experimentally obtained structures of DNA stored in the protein database that contain the CTAG tetranucleotide sequence in order to qualitatively validate our results. Although the experimental data is scarce, with only a fraction of the tetramer sequence space covered and barely none of the hexamer space, the analysis of experimental structural parameters of TpA steps flanked by 5’C-3’G confirms that multimodality is not a force field artifact. We observe correlations between twist, slide and shift very consistent with our results as shown in Figure 3. Indeed, the experimental structures tend to crowd the middle of the plot, surely an effect of the averaging of structures inherent to the structure determination technique that assumes unimodality. However, whenever a particular structure deviates from zero shift conformation, the twist and slide are necessarily in high and positive states, respectively, and they are also highly correlated with each other, in perfect agreement with our results (Figure 3). PDB structures containing the CTAG have values for the shift, slide, roll and twist helical parameters that cover the multimodal distributions obtained in our trajectories. There are only 106 naked DNA structures (some with small ligands or metal ions) and 160 structures of protein-bound DNA containing CTAG. Slide and twist are clearly bimodal, with peaks that fit well to our results (Figure 8). TpA shifts 2 Å towards both the minor or major groove in several protein-bound DNA structures, but the data on naked DNA seems to be insufficient to cover these deformations: there is a small peak at +2 Å, but highly underestimated compared to our results. Roll has a broad distribution, similar to what we obtain, but again low sequence coverage might be to blame for non-uniformity and bias towards several discrete values.

Previous analysis suggests that CTAG has really unique physical properties which should provide the genome a point of high flexibility and polymorphism. Very interestingly, CTAG is one of the lowest populated tetramer in all species (see Supp. Figure S5) appearing mainly on intergenic regions and very rarely on genes. Interestingly, rare CTAG tetramers are well conserved with an unusually low rate of stable SNPs mapping on them (Supp. Figure S6), 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. The same conclusion can be reached from the analysis of cancer genomic data which show that again CTAG is very rarely mutated in cancer (Supp. Figure S7). The unusual physical properties of the CTAG tetramer matches it unusual prevalence and distribution in the genome and its extreme resilience to either germinal (SNPs) o somatic (cancer) mutations. It is tempting to believe that cell takes advantage of the unusual properties of CTAG as points of high flexibility which might help to fold chromatin.

CONCLUSIONS

We present here an in-depth study of one of the most “structurally speaking” polymorphic tetranucleotide found in B-DNA. The complete helical space of the CTAG tetramer has been analyzed by means of extensive molecular dynamics simulations, and by data mining the Protein Data Bank, confirming its highly polymorphic behavior 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 consequence its dynamics, are sequence dependent, and fine-tuned by long-range sequence effects that goes beyond the hexamer context. Based on the concerted and correlated movements of bases and backbone torsions for the described multimodal 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 long-range structural “connection” allows the TpA step to “feel” its sequence environment up to the octamer level, and eventually adopt a different substate if needed. Moreover, we found that previously described unconventional intra-molecular Hydrogen bonds of the type C8H8-O3’ and C6H6-O3’ which link the movements of the bases with the torsions in the backbone, could be used as descriptors of such correlated motions. Finally, we found that although this highly flexible tetramer is extremely underrepresented in several genomes along the animal Kingdome, being mostly present in intergenic sequences, it has been preserved with a low rate of mutation in normal and cancer cell lines implying a possible physical role for CTAG at genomic level.

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.

FUNDING

Spanish Ministry of Science [BFU2014-61670-EXP, BFU2014-52864-R], Catalan SGR, the 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) (all awarded to M.O.); MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona).

REFERENCES

1. Wilkins MHF, Stokes AR, Wilson HR. Molecular Structure of Nucleic Acids: Molecular Structure Of Deoxypentose Nucleic Acids. *Nature* 1953;**171**:738–40.
2. Franklin RE, Gosling RG. Molecular Configuration in Sodium Thymonucleate. *Nature* 1953;**171**:740–1.
3. Lucas AA, Lambin P, Mairesse R *et al.* Revealing the Backbone Structure of B-DNA from Laser Optical Simulations of Its X-ray Diffraction Diagram. *J Chem Educ* 1999;**76**:378.
4. Kypr J, Kejnovská I, Renciuk D *et al.* Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res* 2009;**37**:1713–25.
5. Kato M. Structural bistability of repetitive DNA elements featuring CA/TG dinucleotide steps and mode of evolution of satellite DNA. *Eur J Biochem* 1999;**265**:204–9.
6. Kielkopf CL, Ding S, Kuhn P *et al.* Conformational flexibility of B-DNA at 0.74 å resolution: d(CCAGTACTGG)2. *J Mol Biol* 2000;**296**:787–801.
7. Maehigashi T, Hsiao C, Woods KK *et al.* B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res* 2012;**40**:3714–22.
8. Monchaud D, Allain C, Bertrand H *et al.* Ligands playing musical chairs with G-quadruplex DNA: A rapid and simple displacement assay for identifying selective G-quadruplex binders. *Biochimie* 2008;**90**:1207–23.
9. Radhakrishnan I, Patel DJ. DNA Triplexes: Solution Structures, Hydration Sites, Energetics, Interactions, and Function. *Biochemistry* 1994;**33**:11405–16.
10. Kaushik M, Kaushik S, Bansal A *et al.* Structural diversity and specific recognition of four stranded G-quadruplex DNA. *Curr Mol Med* 2011;**11**:744–69.
11. Dai J, Carver M, Yang D. Polymorphism of human telomeric quadruplex structures. *Biochimie* 2008;**90**:1172–83.
12. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.
13. Dans PD, Danilāne L, Ivani I *et al.* Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res* 2016;**44**:4052–66.
14. Imeddourene A Ben, Xu X, Zargarian L *et al.* The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res* 2016;**44**:3432–47.
15. Ben Imeddourene A, Elbahnsi A, Guéroult M *et al.* Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. MacKerell A (ed.). *PLOS Comput Biol* 2015;**11**:e1004631.
16. Tian Y, Kayatta M, Shultis K *et al.* ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J Phys Chem B* 2009;**113**:2596–603.
17. Zgarbová M, Jurečka P, Lankaš F *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* 2017;**57**:275–87.

18. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.

19. Calladine CR, Drew HR, Luisi BF *et al. Understanding DNA : The Molecule and How It Works*. Elsevier Academic Press, 2004.

20. Olson WK, Gorin AA, Lu XJ *et al.* DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 1998;**95**:11163–8.

21. Cheatham TE, Cieplak P, Kollman PA. A Modified Version of the Cornell *et al.* Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J Biomol Struct Dyn* 1999;**16**:845–62.

22. Pérez A, Marchán I, Svozil D *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 2007;**92**:3817–29.

23. Dršata T, Lankaš F. Multiscale modelling of DNA mechanics. *J Phys Condens Matter* 2015;**27**:323102.

24. Dršata T, Pérez A, Orozco M *et al.* Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J Chem Theory Comput* 2013;**9**:707–21.

25. Ivani I, Dans PD, Noy A *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat Methods* 2016;**13**:55–8.

26. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:gkw1355.

27. Jiang W, Phillips JC, Huang L *et al.* Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. *Comput Phys Commun* 2014;**185**:908–16.

28. Salomon-Ferrer R, Götz AW, Poole D *et al.* Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 2013;**9**:3878–88.

29. Lee J, Cheng X, Swails JM *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput* 2016;**12**:405–13.

30. Páll S, Abraham MJ, Kutzner C *et al.* Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. Springer, Cham, 2015, 3–27.

31. Beveridge DL, Barreiro G, Suzie Byun K *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* 2004;**87**:3799–813.

32. Dixit SB, Beveridge DL, Case DA *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* 2005;**89**:3721–40.

33. Lavery R, Zakrzewska K, Beveridge D *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* 2010;**38**:299–313.

34. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

35. Pasi M, Maddocks JH, Lavery R. Analyzing ion distributions around DNA:

sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res* 2015;**43**:2412–23.

36. Balaceanu A, Pasi M, Dans PD *et al.* The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J Phys Chem Lett* 2017;**8**, DOI: 10.1021/acs.jpclett.6b02451.

37. 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, LX and PAK. AMBER 2016. 2016.

38. Berendsen HJC, Grigera JR, Straatsma TP *et al.* The missing term in effective pair potentials. *J Phys Chem* 1987;**91**:6269–71.

39. Smith DE, Dang LX. Computer simulations of NaCl association in polarizable water. *J Chem Phys* 1994;**100**:3757–66.

40. Dang LX. Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J Am Chem Soc* 1995;**117**:6954–60.

41. Dang LX, Kollman PA. Free Energy of Association of the K+:18-Crown-6 Complex in Water: A New Molecular Dynamics Study. *J Phys Chem* 1995;**99**:55–8.

42. Salomon-Ferrer R, Götz AW, Poole D *et al.* Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 2013;**9**:3878–88.

43. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N* ·log(*N*) method for Ewald sums in large systems. *J Chem Phys* 1993;**98**:10089–92.

44. Ryckaert J-P, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 1977;**23**:327–41.

45. Lavery R, Moakher M, Maddocks JH *et al.* Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* 2009;**37**:5917–29.

46. Schwarz G. Estimating the Dimension of a Model. *Ann Stat* 1978;**6**:461–4.

47. Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* 1995;**90**:773–95.

48. Schilling MF, Watkins AE, Watkins W. Is Human Height Bimodal? *Am Stat* 2002;**56**:223–9.

49. Ghosh A, Bansal M. A glossary of DNA structures from A to Z. *Acta Crystallogr Sect D Biol Crystallogr* 2003;**59**:620–6.

50. Jolliffe IT. *Principal Component Analysis*. Springer-Verlag, 1986.

51. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933;**24**:417–41.

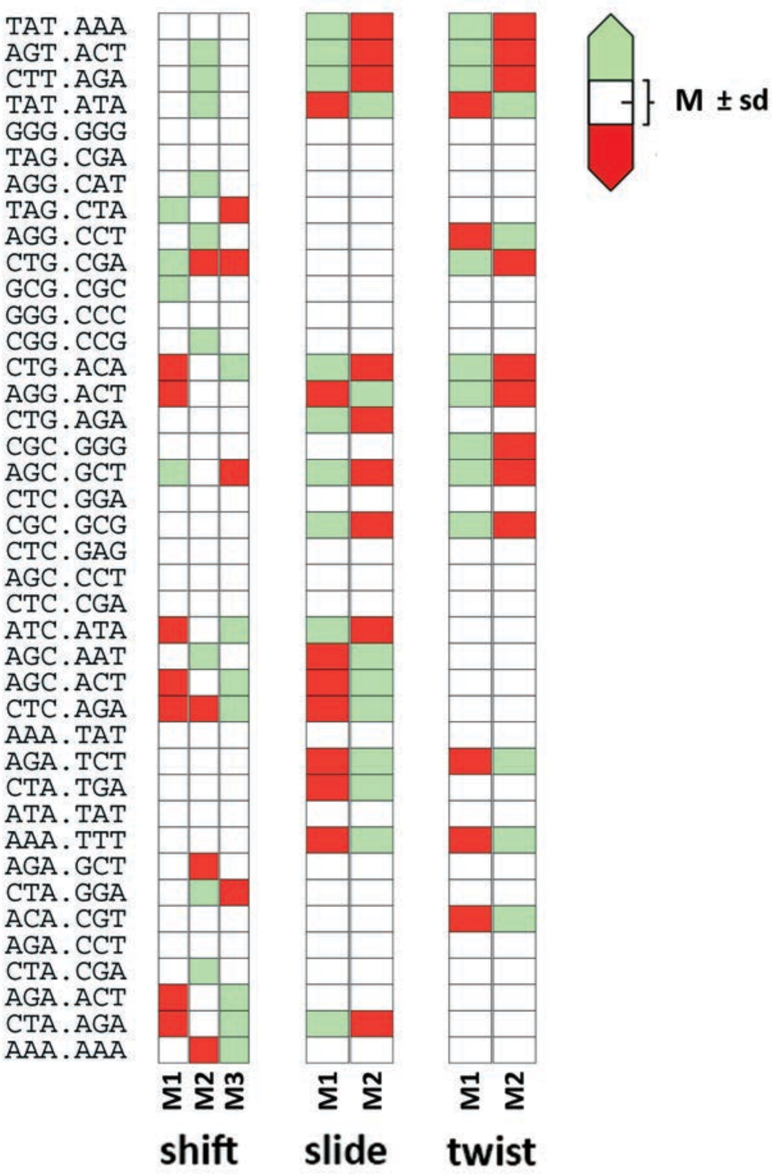
52. Dršata T, Lankaš F. Theoretical models of DNA flexibility. *Wiley Interdiscip Rev Comput Mol Sci* 2013;**3**:355–63.

53. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* 1963;**58**:236–44.

54. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* 2014;**31**:274–95.

55. Beveridge DL, Barreiro G, Suzie Byun K *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* 2004;**87**:3799–813.

FIGURES



	shift			slide		twist	
	M1	M2	M3	M1	M2	M1	M2
TAT.AAA				🟩	🟥	🟩	🟥
AGT.ACT		🟩		🟩	🟥	🟩	🟥
CTT.AGA		🟩		🟩	🟥	🟩	🟥
TAT.ATA		🟩		🟥	🟩	🟥	🟩
GGG.GGG							
TAG.CGA							
AGG.CAT		🟩					
TAG.CTA	🟩		🟥				
AGG.CCT		🟩				🟥	🟩
CTG.CGA	🟩	🟥	🟥			🟩	🟥
GCG.CGC	🟩						
GGG.CCC							
CGG.CCG		🟩					
CTG.ACA	🟥		🟩	🟩	🟥	🟩	🟥
AGG.ACT	🟥			🟥	🟩	🟩	🟥
CTG.AGA				🟩	🟥		
CGC.GGG						🟩	🟥
AGC.GCT	🟩		🟥	🟩	🟥	🟩	🟥
CTC.GGA							
CGC.GCG				🟩	🟥	🟩	🟥
CTC.GAG							
AGC.CCT							
CTC.CGA							
ATC.ATA	🟥		🟩	🟩	🟥		
AGC.AAT		🟩		🟥	🟩		
AGC.ACT	🟥		🟩	🟥	🟩		
CTC.AGA	🟥	🟥	🟩	🟥	🟩		
AAA.TAT							
AGA.TCT				🟥	🟩	🟥	🟩
CTA.TGA				🟥	🟩		
ATA.TAT							
AAA.TTT				🟥	🟩	🟥	🟩
AGA.GCT		🟥					
CTA.GGA		🟩	🟥				
ACA.CGT						🟥	🟩
AGA.CCT							
CTA.CGA		🟩					
AGA.ACT	🟥		🟩				
CTA.AGA	🟥		🟩	🟩	🟥		
AAA.AAA		🟥	🟩				

Figure 1. Relative propensities of the multimodal 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 multimodal distributions with standard deviations that reflect the variation of propensity of each component among sequences. The propensity values were computed BIC analysis (see Methods and Suppl. Methods).

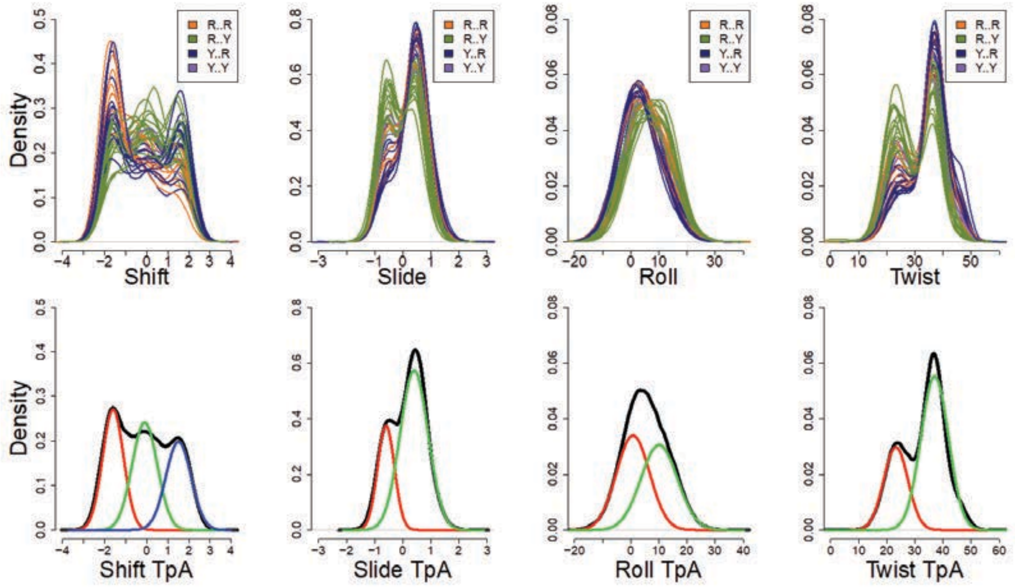


Figure 2. Normalized frequencies of those bps helical parameters found to be bi-normal and tri-normal according to the BIC analysis. FIRST ROW: Overlapped density of the shift, slide, roll, and twist parameters at the central TpA step of the 40 sequences studied (see Supp. Table S1). SECOND ROW: Density 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).

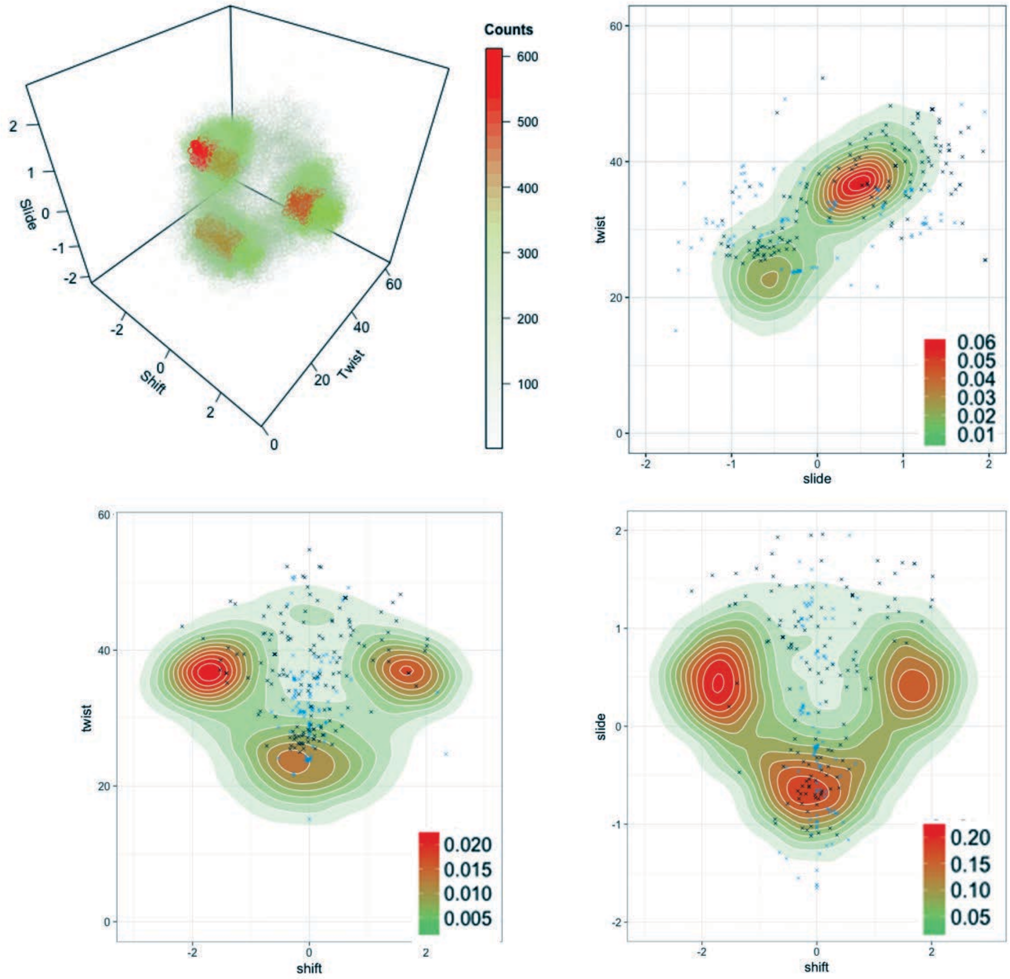


Figure 3. 3D and 2D counts in the shift, slide, and twist planes from MD simulations at the central TpA step. In the 2D density plots experimental structures from the PDB (see Supp. Methods) were added as black crosses (Protein-DNA complexes), or blue crosses (isolated DNA).

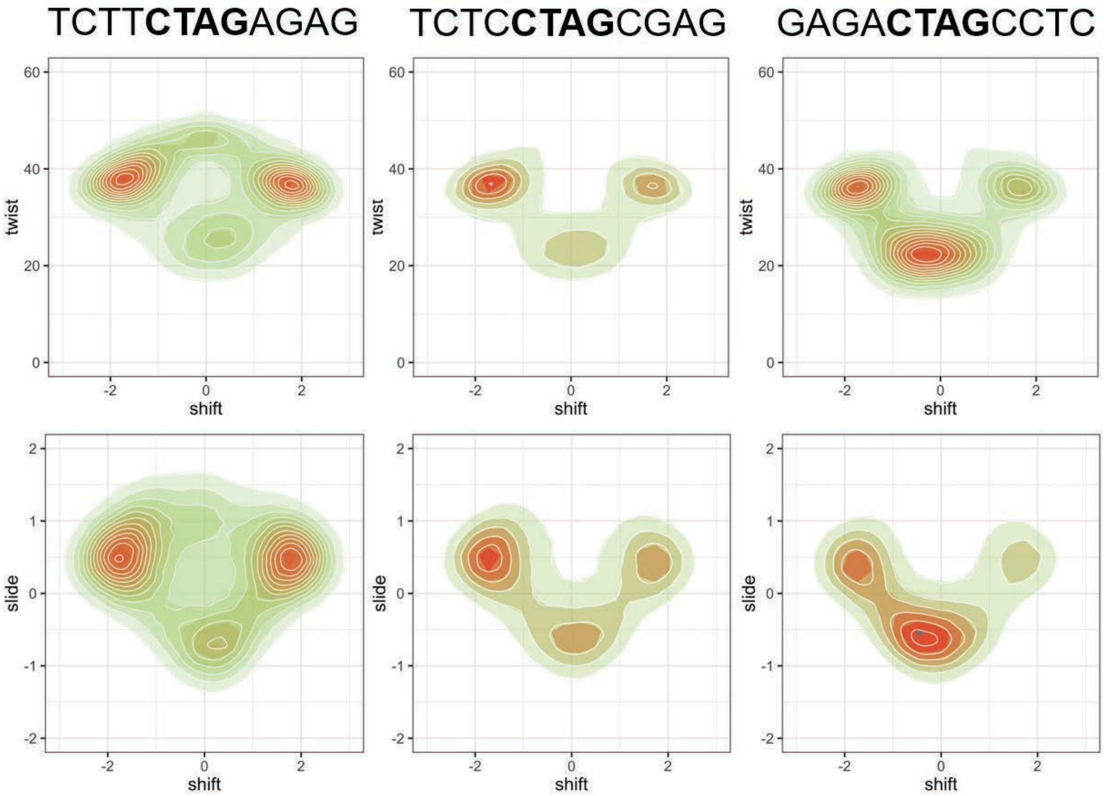


Figure 4. 2D density plots in the shift/twist and shift/slide planes at the central TpA step for 3 selected sequences.

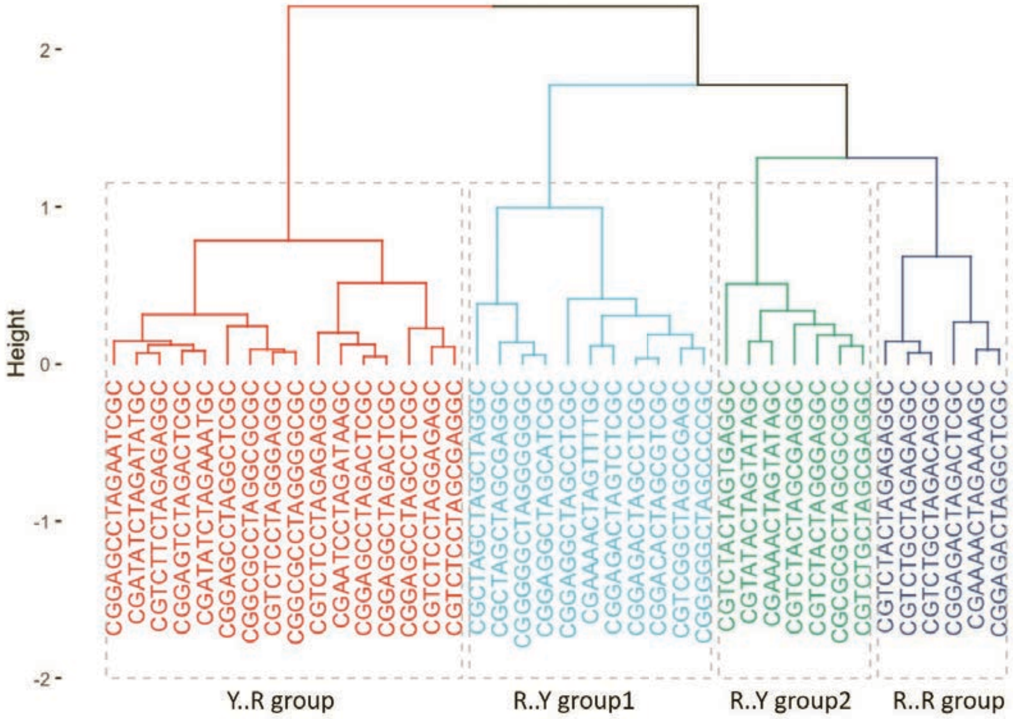


Figure 5. Dendogram obtained from a hierarchical clustering method using Ward's criterion to classify the sequences. The distances were obtained from the symmetric Kullback-Leibler (KL) divergence in the space of 6 helical parameters: shift, slide and twist of TpA step, buckle and propeller of dT, and the buckle of dA (see Supp. Methods).

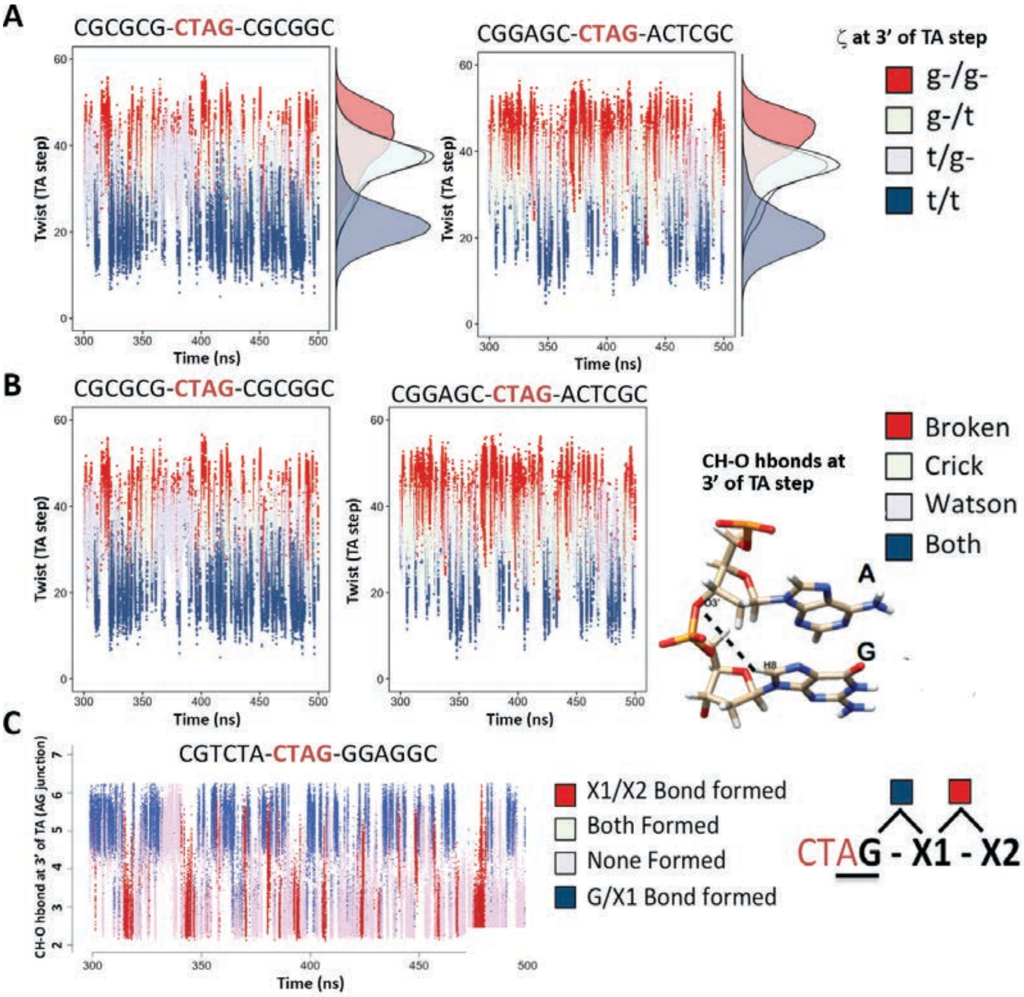


Figure 6. Concerted movements along the backbone and the bases explain the flow of structural information from CTAG tetramer to the octamer level. A) Correlation between twist and the BI/BII population (reduced to the ζ torsion at the 3'-side of TA) at the TA junction. B) Correlation between twist at TA and the CH-O hbond formed at the AG junction (bps +1). C) Correlation between the CH-O hbond at the AG junction with the CH-O hbond at bps+1 (hexamer level), and bps+2 (octamer level). Note that the CH-O hbonds are always coupled to BII propensities, stabilizing the BII substate.

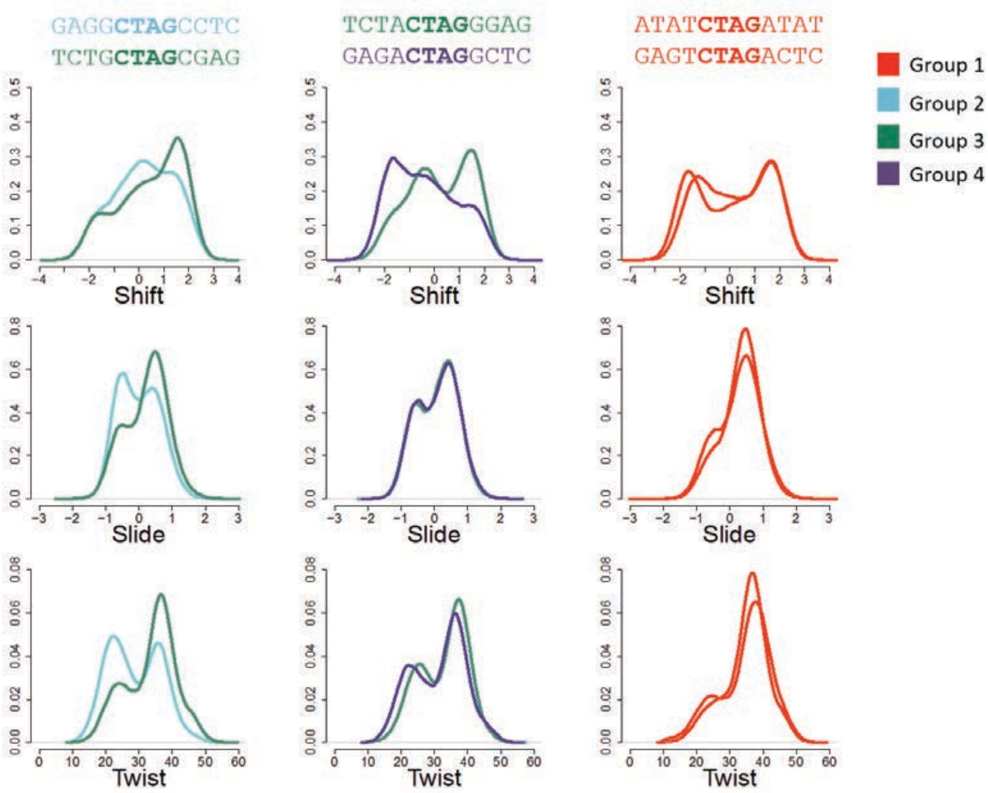


Figure 7. Normalized frequencies of shift, slide and twist at the central TpA step for 3 pairs of selected sequences showing non-negligible octamer effects. The colors used correspond to the groups found in the clustering analysis.

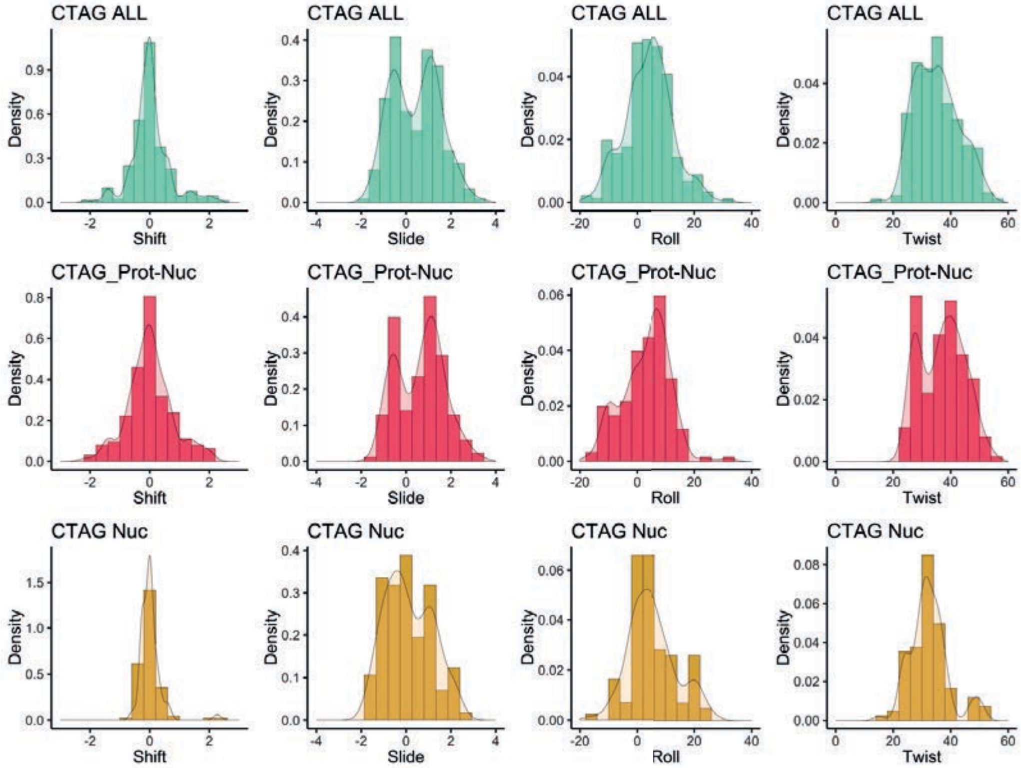


Figure 8. Normalized frequencies of shift, slide, roll and twist obtained from the data mining of the PDB for all structures containing DNA (FIRST ROW), for Protein-DNA complexes (SECOND ROW), and for isolated DNA structures (THIRD ROW).

2 Allosterism and signal transfer in DNA

In 2013, an extensive study both *in vivo* and *in vitro* on the allosteric coupling between proteins bound to the DNA [16] was published and proposed a mechanism for cooperativity based on explicit solvent unrestrained MD simulations. Their model has later been challenged when another group ran significantly longer MD simulations and found significant noise in the data, concluding that an atomistic representation is entirely unfit for addressing this problem.

This motivated us to clarify the issue, based on the our experience that sufficient sampling and the use of the latest generation force fields for DNA [21] (parmbsc1, see Chapter 4) are very successful in deciphering subtle modulation in the conformational landscape, even when structural effects are absent or very mild. We first discuss the structural response, looking at correlations and causality in the geometrical descriptors that would account for a site-to-site information transfer in the DNA. We find that the presence of BAMHI enriches the coupling between the degrees of freedom of the two binding sites.

From a thermodynamic perspective, we eliminate the possibility of a predominantly enthalpic explanation and find that the mechanism is entropy-mediated. In a nutshell, the way this takes place is through the inhibition of the flexibility in the secondary binding site by the restriction in conformational freedom upon effector binding at its own site. This happens because the two binding sites are dynamically coupled. In terms of free energy changes, the formation of the ternary complex is cooperative because in addition to paying an entropy penalty for binding to its own site, the first protein also does some of the unfavorable thermodynamic "work" required to stiffen the secondary binding site.

Encouraged to study the effect in more depth, we adapt a methodology of computing and breaking down transfer entropy from information theory (previously used on proteins) for the study of our system. From the entropy transfer point of view, allosteric communication may be a general property of DNA that should be taken into consideration.

Publication:

Balaceanu A, Pérez A, Dans PD and Orozco M. Allosterism and signal transfer in DNA. Nucleic Acids Res 2018, DOI: 10.1093/nar/gky549.

Allosterism and Signal Transfer in DNA

Alexandra Balaceanu[1], Alberto Pérez[2], Pablo D. Dans[1] and Modesto Orozco[1,3,*]

[1]Joint IRB-BSC Program on Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain.

[2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States.

[3]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.

ABSTRACT

We analyzed the basic mechanisms of signal transmission in DNA and the origins of the allostery exhibited by systems such as the ternary complex BAMHI-DNA-GRDBD. We found that perturbation information generated by a primary protein binding event travels as a wave to distant regions of DNA following a hopping mechanism. However, such a structural perturbation is transient and does not lead to permanent changes in the DNA geometry and interaction properties at the secondary binding site. The BAMHI-DNA-GRDBD allosteric mechanism does not occur through any traditional models: direct (protein-protein), indirect (reorganization of the secondary site) readout, or solvent-release. On the contrary, it is generated by a subtle and less common entropy-mediated mechanism, which might have an important role to explain other DNA-mediated cooperative effects.

INTRODUCTION

Macromolecules are capable to transport information signals emerging from, for example ligand binding, to distant regions activating a variety of secondary effects. One of them is allostery, which implies changes in the binding of a second ligand due to prior interaction of the allosteric effector [1–5]. Allostery has been deeply studied for proteins, where it has been characterized as one of the main mechanisms of control of their activities [5–7]. Much less studied, but equally important, DNA-mediated allostery has been attributed a crucial role in the control of DNA-protein interactions [8–17]. Most cases of cooperativity in DNA can be explained by a "direct read-out" mechanism. Accordingly, a primary protein recognizes a DNA sequence by means of specific interactions with one of its domains, while other(s) domain(s) make(s) specific interactions with a secondary protein, positioning it near the second DNA sequence to be recognized [9–11]. In addition to the "direct readout" model, two other allosteric mechanisms have been suggested: the "indirect read-out", where the primary protein distorts the structure of DNA improving the binding characteristics of the secondary site [12,13], and the "solvent release mechanism" that assumes that primary binding induces changes in water or ion distribution reducing the desolvation cost required for the secondary binding [14]. Recent experiments [15,18,19], and theoretical models [17,20] have shown cases of cooperativity in DNA-protein binding that apparently do not fit within any of the traditional paradigms, as binding sites are distant and binding-induced structural changes in DNA are absent or very mild. Although in proteins a model of "allostery without conformational change" [21] has been described, very few studies have focused on similar processes in DNA. In fact, cooperative binding of BAMHI type II Endonuclease [22] and the glucocorticoid receptor DNA-binding domain (GRDBD) [23] to the DNA is (to our knowledge) the first described example of "allostery without conformational changes" involving DNA [15,24]. The difficulties in understanding the origins of the BAMHI-DNA-GRDBD allostery highlight our limited knowledge on the mechanisms in which information is transferred along DNA [18,25–27].

In this contribution, we propose a model of DNA allostery based on communication from site-to-site by entropy transfer using correlated motions to transmit information through the system. Moreover, we show that the source of allostery is the directionality of time-delayed correlations between the internal degrees of freedom of DNA, which accounts for causality and explains the thermodynamics of complex formation.

MATERIAL AND METHODS

Simulated systems: We explore allosteric effects in duplex DNA containing the canonical BAMHI binding site (d(GGATCC)) and the canonical GRDBD binding site (d(AGAACATGATGTTCT) separated by linkers of increasing length (4, 7, 11 and 15). In all cases four systems were simulated: the naked DNA, the BAMHI-DNA complex, the GRDBD-DNA complex and the BAMHI-DNA-GRDBD trimer (see Supplementary Figure S1). The 4x4 systems were created using Nucleic Acid Builder and standard B-DNA geometrical parameters [28,29], except for the binding region where the geometries were transferred from the respective crystal structures (PDB codes 2BAM and 1R4R). Note that the GRDBD binding site used (the one coming from the X-ray structure), correspond to what authors from ref. [15] labeled as the "reverse" GRDBD sequence.

System preparation: The systems were immersed in octahedral boxes of water, which were defined to guarantee no DNA atom was placed at less than 10 Å from any face of the periodic cell. Hydrated systems were then neutralized by adding Na^+ ions and extra 100 mM NaCl and subjected to a standard minimization/thermalization/pre-equilibration [30,31] procedure followed by a 50 ns equilibration prior to production runs. All the topologies and coordinate files were build with the *leap* program of AmberTools 15 [32] and/or with the utilities provided by GROMACS [33], and run with GROMACS machinery.

Production runs. MD trajectories were collected in the isothermal (T= 298 K), isobaric (P=1 atm) ensemble, using a Langevin thermostat [34] and Andersen-Parrinello barostat [35,36]. The parmbsc1 force-field was used for DNA [37], coupled with ff99SB-ILDN for proteins [38], Dang's parameters for ions [39], and TIP3P waters [40]. Periodic boundary conditions and Particle Mesh Ewald (real space cutoff 12 Å and grid spacing 1.2 Å) were used to account for remote electrostatic interactions [41]. Van der Waals contacts were truncated at the real space cutoff. All bonds containing hydrogen were constrained using LINCS [42], which allowed us to use an integration step of 2 fs. Trajectories were extended for 1 μs each, and system sizes range from 91,164 to 263,834 atoms. Additional trajectories where BAMHI was instantaneously bound in a 1 μs equilibrated naked DNA (linker size=7) were collected to test the mechanism in which perturbation is transferred along DNA. As the direct contacts between the two proteins are possible for the 4-nt linker, most of the discussion is limited to the 7-, 11- and 15-nt linkers. All collected trajectories are available through our BigNasim database [43].

Analysis of the trajectories. Conformational analysis was performed with the use of Curves+ and Canal programs [44] to obtain the DNA internal coordinates at each time frame such as helical parameters and backbone dihedrals. Standard analysis on those helical parameters were performed using tools implemented in our NaFleX server [45], and further analysis was done with a combination of in-house tools.

Analysis of the structural response. Correlations between geometrical variables were evaluated taking into account the nature of the coordinates involved, either linear or circular (backbone torsions). Correlation between two directional variables was assessed with the use of Jammalamadaka formula [46]. To compute time-delayed correlations we used the cross-correlation [47] between the time series of major groove widths at two positions on the DNA with a maximum lag of 5 ns (see the Suppl. Data for a more comprehensive description of these techniques). Correlations network analysis of backbone torsions was performed by computing all circular-circular correlation (edges) between backbone angles (nodes), and using them to build an interaction network represented as a descriptive network graph using the R package igraph [48]. Classical molecular interaction potentials were computed (using our CMIP code [49]) to determine the changes in recognition properties from an enthalpic point of view (only considering Coulomb and VdW interactions), induced by BAMHI binding, on the region of DNA that binds GRDBD. The ionic strength and the reaction-field dielectric constant were set to 0.15 and 78.4 M, respectively, while the dielectric constant for DNA was set to 8 [50]. A protonated methylamine was used as probe particle. Cation analysis was performed by determining the cation distribution in curvilinear cylindrical coordinates. The distribution of sodium cations around the DNA was determined from the last 200 ns of each MD trajectory, and analyzed using Canion [51]. The limits of the grooves were defined as reported elsewhere [52,53]. In order to analyze the effect of protein interaction with DNA while reducing the thermal noise, we defined as protein "sensing contacts" all pairs of amino-acid/nucleotide that when coming in close proximity to each other (distances between centers of mass below 7 Å and at least one atom pair distance below 3 Å) produce the most significant perturbations in the DNA. Selecting structures from the trajectories using these criteria yielded meta-trajectories consisting of at least 10,000 disperse frames that were analyzed together.

Analysis of thermodynamical properties. Entropy calculations were performed using both the Schlitter and Andricioaei/Karplus methods [54,55] employing two separate alignment methods. Absolute entropies of the DNA heavy atoms in the secondary binding region (GRDBD recognition site) were calculated for naked DNA, and all complexes at increasing time windows (50 ns to 1 μs). From these time dependent values we used the Harris' extrapolation scheme [56] to obtain converged absolute entropies at infinite simulation time, and used them to calculate the entropy changes upon protein binding. The dihedral entropy for DNA backbone torsions was computed as a function of the Kullback–Leibler divergence [57] of real dihedral state populations from the assumed independent populations, as described by Cukier [58]. The set of dihedrals was chosen at the interface between linker and secondary binding regions, and it includes torsions on both Watson and Crick strands (α, β, ε, γ and ζ). Using major groove widths fluctuations along the DNA we calculate the transfer entropy (TE) between two base pairs following the method proposed by Schreiber [59], computing the TE as a summation of Shannon entropy terms [57], which stems from the calculation of conditional entropies between time series separated in time by τ (chosen here to be 2 ns). A rough estimate of the difference in free energy of binding GRDBD to the free and protein bound DNA was obtained by following an adaptation of the Confine-Convert-Release (CCR) method described by Roy *et al.* [60], based itself on previously described confinement methods [61,62]. We calculate the energy of confining each structure (naked DNA, BAMHI-DNA, GRDBD-DNA and BAMHI-DNA-GRDBD complexes) to its energy minimum by thermodynamic integration with increasing restraints. The negative of this energy is the release term. In the convert step that completes the thermodynamic cycle, we calculate the energy difference of the DNA atoms between the two highly restrained complexes. Finally, the total binding free energy is calculated from the sum of these individual contributions.

More details on all analyses and methods used are given in the Supplementary Data.

RESULTS AND DISCUSSION

The structural response. We explore here BAMHI-DNA-GRDB allostery [15] using a variety of theoretical approaches. We first investigated the possibility that direct protein-protein interactions can justify the observed cooperative binding of BAMHI and GRDBD to DNA (the list of simulated systems is shown in Suppl. Fig. S1). To this end, we computed the protein-protein interaction energy during the last 100 ns of the 1 μs molecular dynamics (MD) trajectories of DNA bound to BAMHI and GRDBD (10^4 snapshots for each case). While for a short (4 bp) linker protein-protein interaction is sizeable (-17.7±2.0 kcal/mol), for longer linkers it is negligible (< 0.5 kcal/mol in all cases), precluding a "direct read-out" mechanism. Furthermore, analysis of helical parameters and groove dimensions (Figure 1 and Suppl. Fig. 2) demonstrates that (when well equilibrated trajectories are used) the interaction of DNA with BAMHI does not significantly alter the helical geometry at the secondary (GRDBD) binding site, arguing against an "indirect readout" model. Additionally, analysis of the MD trajectories clearly shows that BAMHI-induced changes in water and ion environment are restricted to the BAMHI region (see Suppl. Figure S3 A), also arguing against the prevalence of an ion- or water- release mechanism. The ability of the DNA at the secondary binding region to recognize charged aminoacids in the presence or absence of the effector protein (BAMHI) was assessed with our classical molecular interaction potential (CMIP [49]). We used a protonated methylamine probe to simulate the presence of a charged amino acid sidechain(see Suppl. Fig. S3 B) and found essentially no difference in the electrostatic or van der Waals terms between the bound and free DNA for any of the linker region sizes (see Suppl. Table S1). The cooperative effect detected experimentally should then be explained by a less common alternative mechanism, which implies a subtle flow of information between primary and secondary binding sites.

To investigate the mechanism of the information transfer along DNA we explored the correlation between the movements in primary and secondary binding sites for naked and BAMHI bound DNAs. Early studies on this system suggested the existence of strong correlations in the movements at BAMHI and GRDBD binding sites [15,16], while more refined calculations showed that such correlations may emerge from equilibration artefacts [17]. Indeed (as seen in Suppl. Fig. S4), short simulations lead to an overestimation of the correlation between the two binding sites.But when long equilibration windows are considered, correlations between the two binding sites are still clearly larger for the DNA-BAMHI complex than for the naked DNA (Suppl. Fig. S4 and Fig. 2), suggesting that the GRDBD binding site feels in a dynamic way the presence of the BAMHI, even for the longest linker. Depending on the selected time window and the linker size, the differences in correlation strength between naked and bound DNA can vary,but they are always larger in the presence of the protein, most visibly so in the case of the 15 bp linker system, which also shows the higher cooperativity experimentally.

To establish if the structural correlations observed imply causation, we computed a time-delayed correlation between DNA residues [63], which account for the time lag that might appear as the signal travels from one binding site to the other (see Supplementary Data). The expected time lag at each position was calculated based on the linear progression of correlation maxima at the first 3 base pairs away from the source (bp 5), which show the least amount of noise. The assumption made here is that the signal travels at a constant speed through the sequence.Figure 3 shows correlation coefficients of the major groove widths between base pair 5 and all subsequent base pairs with their corresponding time delay, either in the forward (5'→3' on the Watson strand – from bp 5 to each other bp) or reverse direction (3'→5' – from all further base pairs to bp 5). The top half of Figure 3 depicts a comparison of such time-delayed major groove width correlations between BAMHI bound DNA of the different liker sizes.In the systems with cooperativity-favorable linker sizes (7 bp and 15 bp), the delayed correlation in the 5'→3' direction (forward, orange line) is out of phase from the correlation in the 3'→5' direction (reverse, blue line), so that at key positions, the correlation is significantly stronger in the forward direction. This is due to the fact that the forward correlation at such positions decays significantly slower than in the reverse case. These results indicate that the effect of the major groove fluctuations at BAMHI site (when this protein is bound), on later fluctuations at the secondary binding site, persist for longer times in the case of 7 and 15 bp linkers. In contrast, the 11 bp linker system has practically symmetric responses at these key positions. This suggest that the major groove width fluctuations in the BAMHI binding region drive the motions in the GRDBD binding region using the specific 5'-3' direction for linkers 7 and 15, whereas the nature of the structural correlation between binding sites for the 11 bp linker system is small and non-directional. The bottom half of the figure shows the corresponding plots for the naked DNA of different linker sizes. In the absence of the perturbation induced by the binding of the effector protein, the phase shift does not appear. To further confirm that this behavior is due to the introduction of a perturbation at the effector binding site, we ran a separate simulation where we introduced a gradual harmonic tear opening the major groove width at base pair 5 (bound region) and looked on the effect of this perturbation to the forward and reverse delayed correlations (see Suppl. Fig. 5). This result shows that a perturbation introduced in the binding region of BAMHI will indeed affect the symmetry of the cross-correlation. There is, additionally, significant difference (p-value <0.01 for the cumulative r-square in the GRDBD binding site of favorable linker systems, obtained through pair-wise t-test) between correlation coefficients at the secondary binding region between free and bound DNA when taking the time delay into consideration. This suggests that correlations in the naked DNA might be intrinsic and determined by simply the shape of the double helix that synchronizes motions instantaneously, whereas in the protein-bound duplex the response to perturbation takes a certain amount of time to travel through the sequence.

The changes in the connection pattern between recognition sites in DNA due to BAMHI binding became even more evident when network analysis tools (see Suppl. Data) are used to find connectivity maps between the different backbone torsions (Figure4; see results for a broader correlation range in Suppl. Fig. S6).Clearly, the presence of BAMHI enriches the connectivity between the different torsional degrees of freedom. Interestingly, while such connections are mostly local and sequential for the naked DNA, BAMHI binding triggers crosstalk between the backbones of the two strands, through the space, and from the linker region to the secondary binding site, as expected from a “hopping” information transfer mechanism (Figure 4, and Suppl. Fig. S6), *i.e*. where the information flows by giving hops between the dihedrals backbone of non-sequential residues.

As noted above, the crosstalk between the two binding sites detected upon BAMHI binding does not lead in average to dramatic geometrical changes at distant regions (Figure 1 and Suppl Fig. 3), but generates pulses of distortion that can travel quite long distances generating non-negligible temporary geometrical distortions in the duplex. Although differences in major groove width between naked and BAMHI-bound DNAs are in average rather small out of the BAMHI binding site (Figure 5), they increase dramatically for those selected structures where we detected the strongest DNA-protein contacts (around 10^4 frames, see Methods). This suggests that “protein-sensing” leads to a sizeable distortion pulse at quite long distances thanks to a “hopping” mechanism with ½ and full turn periodicities (Figure 5, left). During a protein-sensing event the signal can be transferred to remote regions of the DNA, but once the contact is released, the structure relaxes and the signal starts to dissipate, as shown by the evolution in time of major groove width correlations along the sequence during and after protein clenching (Suppl. Fig. S7). Very interestingly, the distortion pattern introduced by “protein sensing” is enhanced at the GRDBD binding site for linkers 7 and 15, *i.e*. those showing experimentally strong cooperativity, while for linker 11, where cooperativity is not experimentally detected, the peak of the perturbation wave is displaced with respect to the secondary recognition site. Therefore, the linker size modulates the impact of the distortion signal at the secondary binding site, suggesting thatprotein contacts affect not only the major groove width at the secondary binding site, but also the cross-talk between the two binding sites, as itcan be observed from the time evolution of structural correlations during and after protein sensing (Suppl. Fig. S7).

To further validate the idea that “protein-sensing” generates a wave of distortion transmitting a pulse of information to distant regions of DNA, we analyzed the correlation between the groove width after 50, 100 and 200 ns of the instantaneous insertion of BAMHI in its binding site in equilibrated DNAs. Average results in Figure 5 illustrate the generation of a perturbation wave by "protein-sensing" that travels with a ½ and 1 turn periodicity to reach the GRDBD binding site, confirming previous suggestions on the transient nature of the structural distortion.

The entropic origin of cooperativity. The analysis of the structural response presented above suggests that DNA acts as a wire transmitting pulses of information originated at the primary binding site that travel to distant regions. The existence of such a mechanism of information transfer is a necessary, but not sufficient condition for the appearance of cooperativity. So, the question is now how these changes in the dynamics of DNA affect binding thermodynamics. To answer this question, we first evaluate the impact that backbone correlations (shown in Figure 4) have in the DNA entropy. With this purpose we computed the dihedral entropy at the interface of the linker region and the secondary binding site for all linker sizes following the method described by R.I. Cukier [58], which measures the decrease of entropy arising from the dependence among the dihedrals (see Suppl. Data). The results (Figure 6 top) strongly suggest that the entropic change associated with the network of correlations at these positions depend on the linker size, in good qualitative agreement with experimental data (Figure 6, see Suppl. Data for details).

We further processed our equilibrated MD trajectories (see Suppl. Data) of the naked DNA, BAMHI-DNA, GRDBD-DNA and BAMHI-DNA-GRDBD complexes to determine the (DNA) entropy change arising from GRDBD binding in naked DNA and when DNA is already bound to BAMHI. From these values we can define the entropy cooperativity as $\Delta\Delta S(coop)=\Delta S(B)-\Delta S(A)$, where $\Delta S(A)$ and $\Delta S(B)$ are the entropic variation associated to the binding of the GRDBD protein to naked or BAMHI-bound DNA respectively. Entropies were computed from the analysis of the mass-weighted covariance matrix as described by Andricioaiei-Karplus [55], and to have an additional estimate from Schlitter's formulation [54]. To gain extra confidence on the robustness of the results two alignment methods were used to define the average structure, and estimates were obtained for different time-windows, which were then combined using Harris' extrapolation scheme [56] to obtain values extrapolated at infinite simulation time (see Suppl. Data for details). An example of the obtained results is summarized in Figure 6 bottom (the results are quite robust to the approach used to align the duplexes, to the procedure followed to transform oscillations into entropy measures, to the extension of the trajectory, or even through the simulation of replicas; see Suppl. Fig. S8). Thus, as suggested by dihedral entropy measures above, the cooperativity studied here has an entropic origin. Very interestingly, for linkers of 7 and 15 bp, where large cooperative effects were detected [15], we observe that the entropy change associated to the binding of GRDBD is significantly reduced when the DNA is previously interacting with BAMHI (leading to a positive cooperative entropy term). On the contrary, for duplexes with an 11 bp linker no significant entropy differences are found when binding happens in naked or BAMHI-bound DNA, suggesting no sizeable cooperativity, in perfect qualitative agreement with experimental findings [15].

It is worth noting that the entropy-mediated mechanism of cooperative binding observed herein, is also supported by relative changes in the effective temperature computed from atomic oscillations in the presence/absence of the first protein (see Suppl. Fig. S9). Additionally, the Confine-Convert-Release (CCR) calculations [61,60,62] (see Suppl. Data. and Suppl. Fig. S10) further confirm the expected free-energy change associated to cooperativity for linkers 7 and 15, in agreement with the relative k_{off} measured experimentally [15].

We went one step further and examined the information transfer landscape of the naked and BAMHI-bound DNA using Schreiber's formulation of entropy transfer [59]. This approach allows us to find entropy sinks and sources upon the binding of the protein to DNA, and explains how given pairs of nucleotides from one binding site to the other communicate with each other using entropy transfer [64–67]. Thus, based on the Shannon formulation of entropy [57], but taking into account the time delayed conditional probabilities of time series [59], we quantify the allosteric communication through the DNA (for details see Suppl. Data). Results are summarized in Figure 7 and Supp. Fig. S11 and show the entropy transfer landscape between residues of the DNA when BAMHI is bound

(Supp. Fig. S11, right), and the quite uniform landscape of the naked DNA (Supp. Fig. S11, left). Without the protein, only few residues in the diagonal of the entropy map display net entropy ($T^{NET}{}_{i\rightarrow j}$) transfer (the communication is local in nature), while for most of the residues the information flowing in and out to the rest of the sequence is basically the same. The binding of BAMHI produces a drastic change, dominated by a sizeable net entropy transfer from the bp in the BAMHI bound region to the secondary binding region (thus, in the 5'→3' direction), which involves several bp and is clearly non-local (Figure 7).The results show that in the presence of the effector protein (BAMHI), the base pairs of its binding site are major entropy sources for several base pairs along the sequence, whereas base pairs in the secondary binding region specifically change their entropy transfer characteristics, becoming notable acceptors of entropy (Figure 7). Analyzing the provenance of these changes as shown in the entropy transfer landscape of Supp Fig S11, the BAMHI binding region seems to be an exceptionally strong entropy source for the bases that bind to the GRDBD protein (Supp. Fig. S11), displaying directionality in the interactions of the two binding regions, which could be considered as an entropic switch that controls the binding of GRDBD.

CONCLUSIONS

Results reported here suggest that BAMHI binding to DNA generates a perturbation wave that travels to quite distant regions, and if the linker length is suitable, produces a change in structural correlations between residues in the secondary binding site. This change reduces the entropy cost associated to the second binding. We are pointing then to protein-induced changes in DNA-entropy as the origin of cooperativity in the explanation for BAMHI-DNA-GRDBD binding cooperativity. This type of entropy-mediated allostery was previously suggested for protein-protein interactions [21,68,69], and for the binding of small minor groove binders to DNA [27,56], but to our knowledge, it has not been previously described at the molecular level for DNA-protein binding. Our work also highlightsthe significant information transfer between base pairs in these systems. From the entropy transfer point of view, allosteric communication may be a general property of DNAthat should be taken into consideration. Furthermore, we demonstrate that the knowledge of time delayed correlations and entropy transfer is needed to quantify allosteric cross-talk through the DNA, as an alternative to the established paradigms of cooperativity and allosterism. Time delayed events and causality analyses have only recently started to be viewed as crucial tools for studying allosteric communication in proteins. We now show that information transfer through DNA merits the same attention as a mechanism to explain cooperativity. We speculate that this entropy-mediated cooperativity can be quite general, considering that many proteins involved in DNA recognition are too small to make significant protein-protein contacts to account for the direct readout mechanism, that in many cases proteinsdo not introduce large remote structural distortions in DNAupon binding making the indirect readout also unlikely, and that the rearrangement of solvent molecules is usually quite local in nature precluding for most of the cases the solvent-release paradigm [70–72].

AVAILABILITY

The trajectories and the associated analyses are accessible from the MuGBigNAsim portal: http://www.multiscalegenomics.eu/MuGVRE/modules/BigNASimMuG/.

SUPPLEMENTARY DATA

Analysis of the trajectories and additional results are included in the Supplementary Data. Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank Prof. R. Lavery for very useful scientific discussions, and Prof. J.L. Gelpi and Dr. A. Hospital for help in the calculations of protein-protein interactions.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.

FUNDING

Spanish Ministry of Science [BFU2014-61670-EXP, BFU2014-52864-R]; Catalan SGR, the 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); MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). The project was awarded computer resources from the PRACE 9th call. Funding for open access charge: European Research Council (ERC SimDNA).

REFERENCES

1. Changeux J-P, Edelstein SJ. Allosteric Mechanisms of Signal Transduction. *Science (80-)* 2005;**308**:1424–8.

2. Liu J, Nussinov R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLOS Comput Biol* 2016;**12**:e1004966.

3. Cui Q, Karplus M. Allostery and cooperativity revisited. *Protein Sci* 2008;**17**:1295–307.

4. Motlagh HN, Wrabl JO, Li J *et al.* The ensemble nature of allostery. *Nature* 2014;**508**:331–9.

5. Hilser VJ, Wrabl JO, Motlagh HN. Structural and Energetic Basis of Allostery. *Annu Rev Biophys* 2012;**41**:585–609.

6. Monod J, Wyman J, Changeux Jp. On The Nature Of Allosteric Transitions: A Plausible Model. *J Mol Biol* 1965;**12**:88–118.

7. Koshland DE, Némethy G, Filmer D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry* 1966;**5**:365–85.

8. Chaires JB. Allostery: DNA Does It, Too. *ACS Chem Biol* 2008;**3**:207–9.

9. Georges AB, Benayoun BA, Caburet S *et al.* Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? *FASEB J* 2010;**24**:346–56.

10. Lelli KM, Slattery M, Mann RS. Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Annu Rev Genet* 2012;**46**:43–68.

11. Slattery M, Riley T, Liu P *et al.* Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* 2011;**147**:1270–82.

12. Moretti R, Donato LJ, Brezinski ML *et al.* Targeted Chemical Wedges Reveal the Role of Allosteric DNA Modulation in Protein−DNA Assembly. *ACS Chem Biol* 2008;**3**:220–9.

13. Chenoweth DM, Dervan PB. Allosteric modulation of DNA by small molecules. *Proc Natl Acad Sci U S A* 2009;**106**:13175–9.

14. Harris L-A, Williams LD, Koudelka GB. Specific minor groove solvation is a crucial determinant of DNA binding site recognition. *Nucleic Acids Res* 2014;**42**:14053–9.

15. Kim S, Brostromer E, Xing D *et al.* Probing Allostery Through DNA. *Science (80-)* 2013;**339**:816–9.

16. Xu X, Ge H, Gu C *et al.* Modeling Spatial Correlation of DNA Deformation: DNA Allostery in Protein Binding. *J Phys Chem B* 2013;**117**:13378–87.

17. Dršata T, Zgarbová M, Jurečka P *et al.* On the Use of Molecular Dynamics Simulations for Probing Allostery through DNA. *Biophys J* 2016;**110**:874–6.

18. Lesne A, Foray N, Cathala G *et al.* Chromatin fiber allostery and the epigenetic code. *J Phys Condens Matter* 2015;**27**:64114.

19. Camunas-Soler J, Alemany A, Ritort F. Experimental measurement of binding energy, selectivity, and allostery using fluctuation theorems. *Science* 2017;**355**:412–5.

20. Xu X, Ge H, Gu C *et al.* Modeling spatial correlation of DNA deformation: DNA allostery in protein binding. *J Phys Chem B* 2013;**117**:13378–87.

21. Cooper A, Dryden DT. Allostery without conformational change. A plausible

model. *Eur Biophys J* 1984;**11**:103–9.

22. Newman M, Strzelecka T, Dorner LF *et al.* Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science* 1995;**269**:656–63.

23. Luisi BF, Xu WX, Otwinowski Z *et al.* Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* 1991;**352**:497–505.

24. Narasimhan K, Pillay S, Huang Y-H *et al.* DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Res* 2015;**43**:1513–28.

25. Rohs R, Jin X, West SM *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem* 2010;**79**:233–69.

26. Panne D. The enhanceosome. *Curr Opin Struct Biol* 2008;**18**:236–42.

27. Tevis DS, Kumar A, Stephens CE *et al.* Large, sequence-dependent effects on DNA conformation by minor groove binding compounds. *Nucleic Acids Res* 2009;**37**:5550–8.

28. Cheatham TE, Brooks BR, Kollman PA *et al.* Molecular modeling of nucleic acid structure. *Curr Protoc nucleic acid Chem* 2001;**Chapter 7**:Unit 7.5.

29. Neidle S. *Oxford Handbook of Nucleic Acid Structure*. Oxford University Press, 1999.

30. Dans PD, Danilāne L, Ivani I *et al.* Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res* 2016;**44**:4052–66.

31. Pérez A, Luque FJ, Orozco M. Dynamics of B-DNA on the Microsecond Time Scale. *J Am Chem Soc* 2007;**129**:14739–45.

32. Case DA, Babin V, Berryman J *et al.* Amber 14. 2014.

33. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model* 2001;**7**:306–17.

34. Loncharich RJ, Brooks BR, Pastor RW. Langevin dynamics of peptides: The frictional dependence of isomerization rates ofN-acetylalanyl-N?-methylamide. *Biopolymers* 1992;**32**:523–35.

35. Andersen HC, C. H. Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys* 1980;**72**:2384–93.

36. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 1981;**52**:7182–90.

37. Ivani I, Dans PD, Noy A *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat Methods* 2015;**13**:55–8.

38. Lindorff-Larsen K, Piana S, Palmo K *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 2010;**78**:1950–8.

39. Smith DE, Dang LX. Computer simulations of NaCl association in polarizable water. *J Chem Phys* 1994;**100**:3757–66.

40. Jorgensen WL, Chandrasekhar J, Madura JD *et al.* Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;**79**:926–35.

41. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N*·log(*N*) method for Ewald sums in large systems. *J Chem Phys* 1993;**98**:10089–92.

42. Hess B, Hess B, Bekker H *et al.* LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput CHEM* 1997;**18**:18--1463.

43. Hospital A, Andrio P, Cugnasco C *et al.* BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res* 2016;**44**:D272–8.

44. Lavery R, Moakher M, Maddocks JH *et al.* Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* 2009;**37**:5917–29.

45. Hospital A, Faustino I, Collepardo-Guevara R *et al.* NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res* 2013;**41**:W47–55.

46. Jammalamadaka SR, SenGupta A. *Topics in Circular Statistics*. WORLD SCIENTIFIC, 2001.

47. Bolinder EF. The Fourier integral and its applications. *Proc IEEE* 1963;**51**:244–245, 252–3.

48. Csardi G, Nepusz T. The igraph software package for complex network research | BibSonomy. *InterJournal, Complex Syst* 2006;**1695**:1–9.

49. Gelpí JL, Kalko SG, Barril X *et al.* Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins* 2001;**45**:428–37.

50. Cuervo A, Dans PD, Carrascosa JL *et al.* Direct measurement of the dielectric polarization properties of DNA. *Proc Natl Acad Sci* 2014;**111**:E3624–30.

51. Lavery R, Maddocks JH, Pasi M *et al.* Analyzing ion distributions around DNA. *Nucleic Acids Res* 2014;**42**:8138–49.

52. Dans PD, Faustino I, Battistini F *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* 2014;**42**:11304–20.

53. Pasi M, Maddocks JH, Lavery R. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res* 2015;**43**:2412–23.

54. Schlitter J, Jürgen. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem Phys Lett* 1993;**215**:617–21.

55. Andricioaei I, Karplus M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J Chem Phys* 2001;**115**:6289–92.

56. Harris SA, Gavathiotis E, Searle MS *et al.* Cooperativity in drug-DNA recognition: a molecular dynamics study. *J Am Chem Soc* 2001;**123**:12658–63.

57. Cover TM, Thomas JA. *Elements of Information Theory*. Wiley-Interscience, 2006.

58. Cukier RI. Dihedral Angle Entropy Measures for Intrinsically Disordered Proteins. *J Phys Chem B* 2015;**119**:3621–34.

59. Schreiber T. Measuring Information Transfer. *Phys Rev Lett* 2000;**85**:461–4.

60. Roy A, Perez A, Dill KA *et al.* Computing the Relative Stabilities and the Per-Residue Components in Protein Conformational Changes. *Structure* 2014;**22**:168–75.

61. Michael D. Tyka, Anthony R. Clarke and, Sessions RB. An Efficient, Path-Independent Method for Free-Energy Calculations. 2006, DOI:

10.1021/JP060734J.

62. Cecchini M, Krivov S V., Spichty M *et al.* Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers. *J Phys Chem B* 2009;**113**:9728–40.

63. Vatansever S, Gümüş ZH, Erman B. Intrinsic K-Ras dynamics: A novel molecular dynamics data analysis method shows causality between residue pair motions. *Sci Rep* 2016;**6**:37012.

64. Gourévitch B, Eggermont JJ. Evaluating Information Transfer Between Auditory Cortical Neurons. *J Neurophysiol* 2007;**97**:2533–43.

65. Staniek M, Lehnertz K. Symbolic Transfer Entropy. *Phys Rev Lett* 2008;**100**:158101.

66. Hacisuleyman A, Erman B. Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin. Briggs JM (ed.). *PLOS Comput Biol* 2017;**13**:e1005319.

67. Kamberaj H, van der Vaart A. Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations. *Biophys J* 2009;**97**:1747–55.

68. Capdevila DA, Braymer JJ, Edmonds KA *et al.* Entropy redistribution controls allostery in a metalloregulatory protein. *Proc Natl Acad Sci* 2017;**114**:4424–9.

69. Guo J, Zhou H-X. Protein Allostery and Conformational Dynamics. *Chem Rev* 2016;**116**:6503–15.

70. Remenyi A, Lins K, Nissen LJ *et al.* Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev* 2003;**17**:2048–59.

71. Tahirov TH, Sato K, Ichikawa-Iwata E *et al.* Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter. *Cell* 2002;**108**:57–70.

72. Shiina M, Hamada K, Inoue-Bungo T *et al.* A Novel Allosteric Mechanism on Protein–DNA Interactions underlying the Phosphorylation-Dependent Regulation of Ets1 Target Gene Expressions. *J Mol Biol* 2015;**427**:1655–69.

.

FIGURES

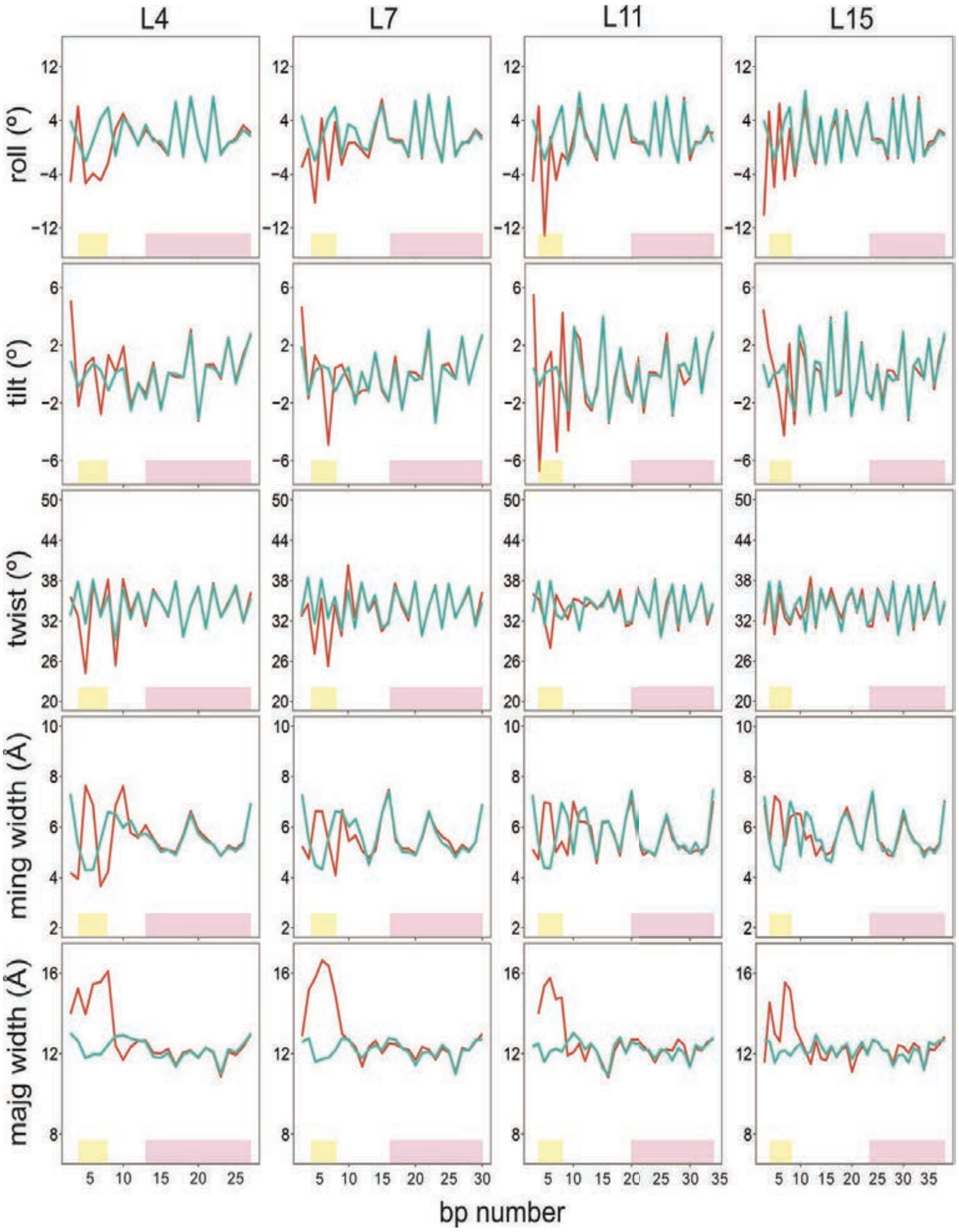


Figure 1. Average (0.2-1 μs) rotational inter base-pair parameters (in degrees) and groove widths (Major: majg and Minor: ming in Å) for the DNA duplexes with linkers of different lengths. Values for free DNA (cyan) versus those for BAMHI-bound DNA (red), where the binding region of BAMHI is highlighted in yellow and the recognition site for the GRDBD in magenta.

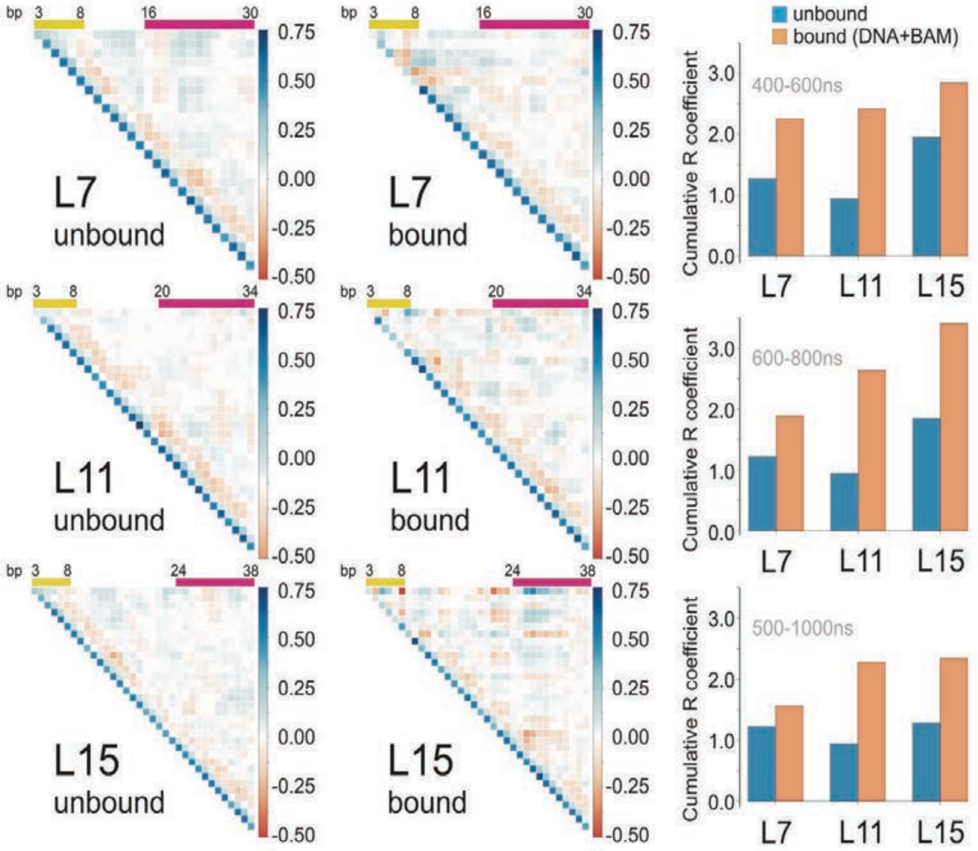


Figure 2. Left: Upper diagonal matrix of major groove width correlation coefficients along the sequence for the unbound and BAMHI-bound DNA. The regions where the two proteins bind to the DNA are highlighted in yellow and magenta. Correlations have been calculated for the fully equilibrated last 200 ns of each trajectory. Right: Cumulative R-coefficients between the nucleotides involved in BAMHI binding and the ones belonging to the GRDBD recognition site for different time windows of the simulation (p-value <0.01 between replicas, see main text).

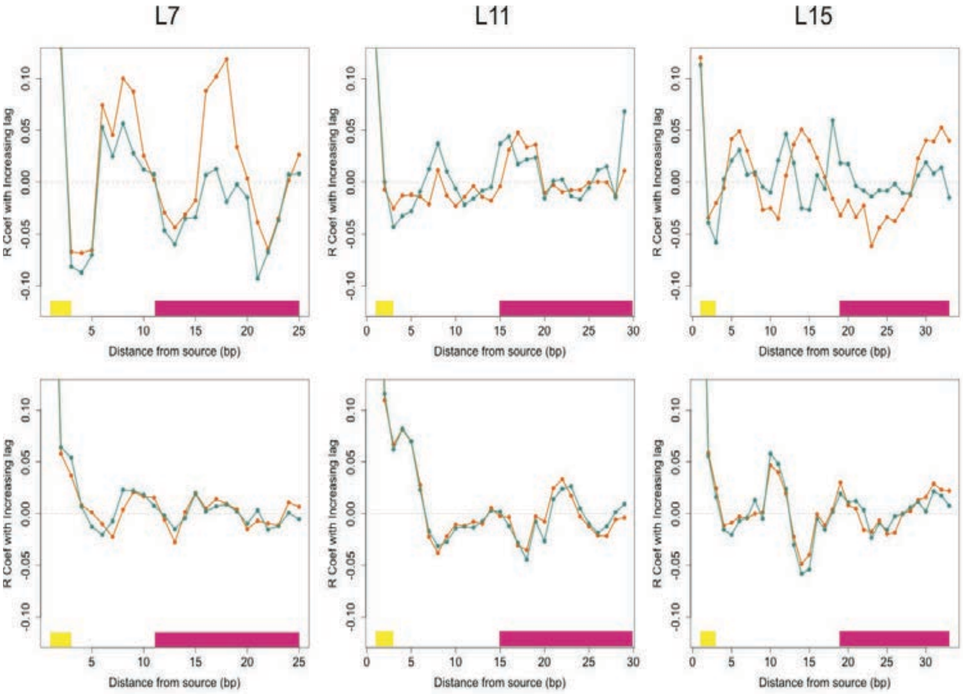


Figure 3. Time-delayed cross correlation of major groove widths between base pair5 belonging to the BAMHI binding regions and all subsequent bases in all linker size system. Top: bound complexes; Bottom: Corresponding naked DNAs. The abscissa denotes the distance in base pairs from the perturbation source (bp 5). "Forward" correlations (5' to 3' on Watson strand, corresponding to cross-correlations from bp 5 to all other bases) are depicted in orange, whilst "reverse" correlations (3' to 5' direction, referring to cross-correlations from each bp to bp 5) appear in blue. The curves are calculated as described in the Supplementary Data for the DNA+BAMHI complexes of each linker size.

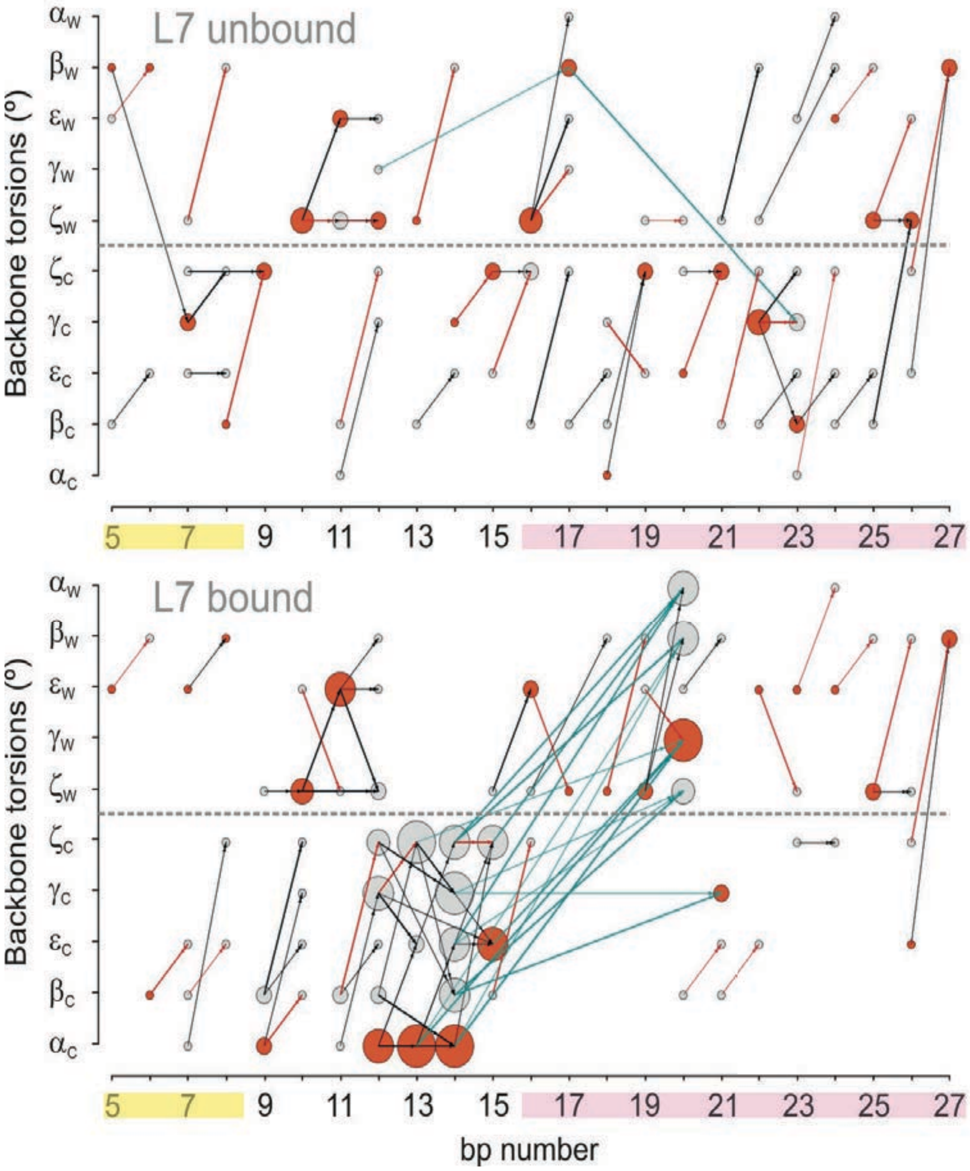


Figure 4. Correlations network analysis showing through space propagation of correlated motions in the DNA backbone (both the Watson (W) and Crick (C) strand torsions) with 0.5>|R|>0.4 for naked and BAMHI-bound DNA (case of 7 bp linker) from the center of the BAMHI binding site along the sequence. Vertex size is proportional to their degree (the number of connections) and, for each level, the vertex with the highest degree is coloured in orange. Correlations between levels more than 4 base pairs apart are depicted as cyan arrows. From each level, the strongest correlation with a starting point on that level is shown as an orange

arrow, unless it is already coloured in blue. Protein binding sites are highlighted in yellow and magenta.

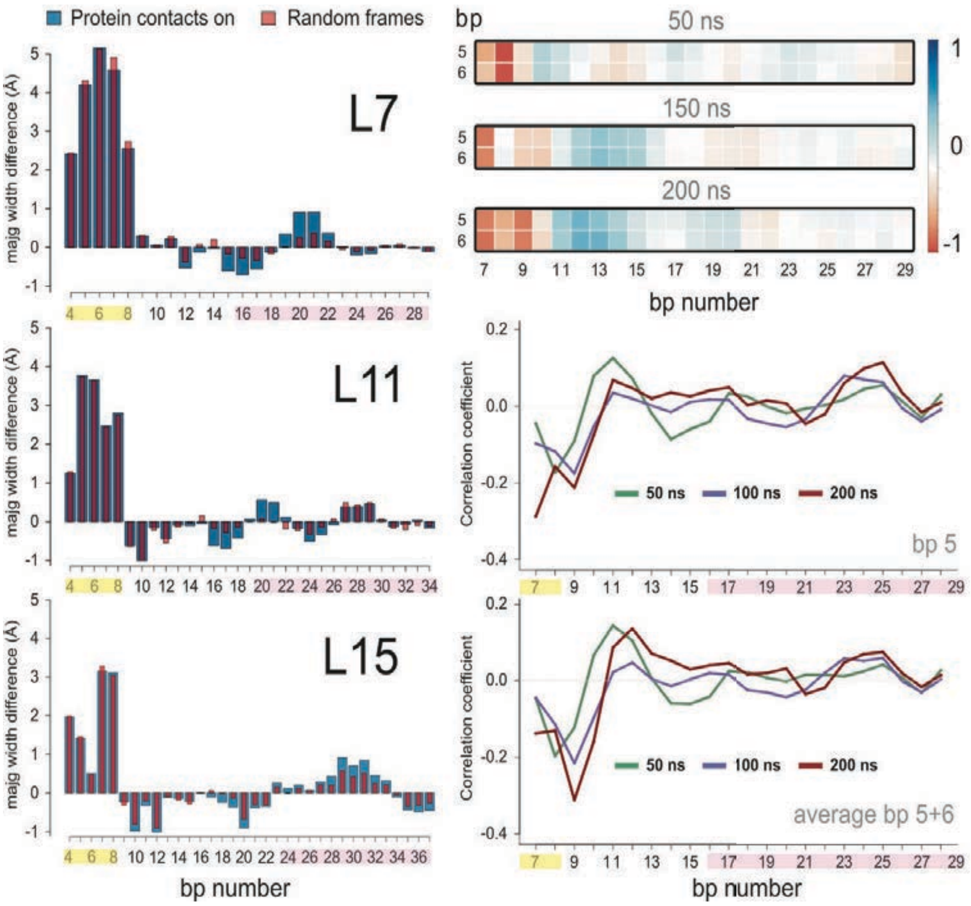


Figure 5. Left: Differences in major groove width between naked and BAMHI-bound DNAs (7 bp linker) for those frames (around 10,000) where there are strong DNA-protein contacts (bars in blue). In each plot there is also an additional control obtained by choosing random snapshots of the same trajectories (i.e. ensembles not enriched in strong-contacts; bars in red). The protein binding sites are highlighted in yellow and magenta. Right: Correlation coefficients between the groove width after 50, 100 and 200 ns of simulation time after the instantaneous insertion of BAMHI in its binding site in equilibrated DNAs for the 7-bp liker DNA. Average results over 5 replicas illustrate a large initial perturbation at the BAMHI binding site that over time is propagated and gains amplitude at the second binding site.

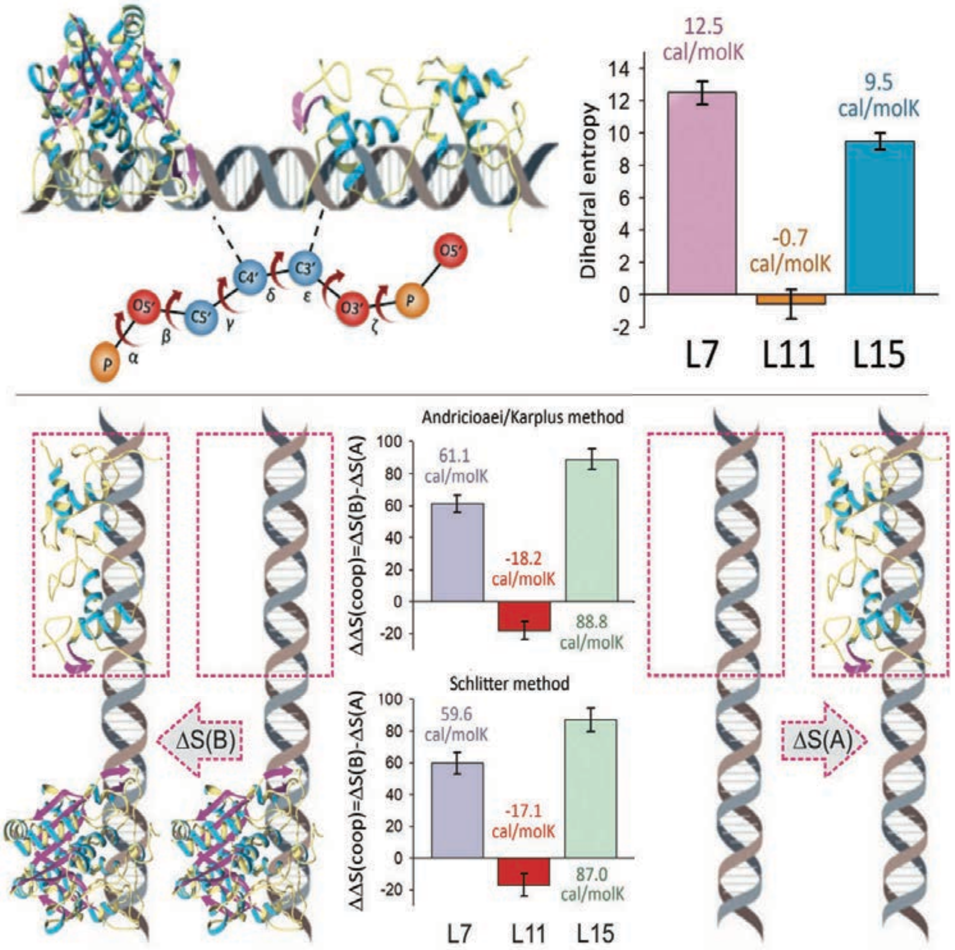


Figure 6. Top: Dihedral entropy change at the interface between linker and secondary binding regions for linkers of 7, 11 and 15 bp. Bottom: DNA entropy variation ΔS(B)-ΔS(A) induced by the binding of GRDBD at the secondary binding region for DNA duplexes containing 7, 11 and 15 bp linkers. Positive values mean that entropy contributes favourably to the cooperative binding of BAMHI and GRDBD.

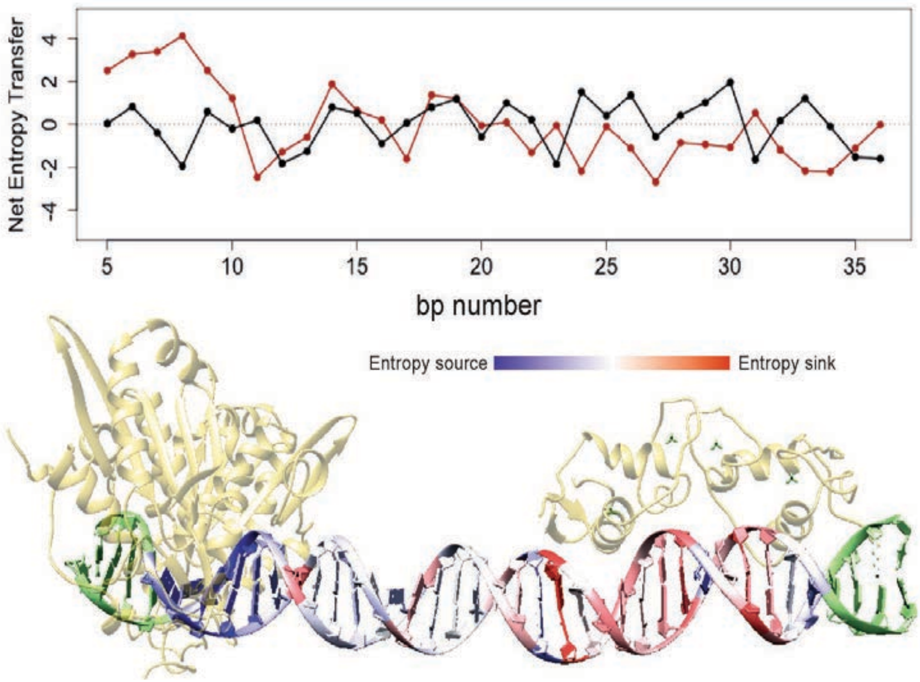


Figure 7. Top: Net entropy transfer from each residue to the rest of the sequence, calculated for free (black line) and BAMHI-DNA (red line). Residues with positive values of net entropy transfer are entropy sources, while residues with negative values are entropy sinks. Bottom: Values of net entropy transfer difference between bound and naked DNA are mapped onto the three-dimensional structure of the double helix (blue is for base pairs that are stronger entropy sources in the presence of BAMHI and red is for base pairs that behave more like entropy acceptors when the protein is bound compared to the free DNA).

Bibliography for Chapter V

1. Olson WK, Gorin AA, Lu XJ *et al.* DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 1998;**95**:11163–8.

2. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.

3. Rohs R, Jin X, West SM *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem* 2010;**79**:233–69.

4. Wecker K, Bonnet MC, Meurs EF *et al.* The role of the phosphorus BI-BII transition in protein-DNA recognition: the NF-kappaB complex. *Nucleic Acids Res* 2002;**30**:4452–9.

5. Battistini F, Hunter CA, Gardiner EJ *et al.* Structural Mechanics of DNA Wrapping in the Nucleosome. *J Mol Biol* 2010;**396**:264–79.

6. Klimasauskas S, Kumar S, Roberts RJ *et al.* HhaI methyltransferase flips its target base out of the DNA helix. *Cell* 1994;**76**:357–69.

7. Tian Y, Kayatta M, Shultis K *et al.* ^{31} P NMR Investigation of Backbone Dynamics in DNA Binding Sites †. *J Phys Chem B* 2009;**113**:2596–603.

8. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:gkw1355.

9. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

10. Lavery R, Zakrzewska K, Beveridge D *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* 2010;**38**:299–313.

11. Zgarbová M, Jurečka P, Lankaš F *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* 2017;**57**:275–87.

12. Hospital A, Andrio P, Cugnasco C *et al.* BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res* 2016;**44**:D272–8.

13. Jolma A, Yin Y, Nitta KR *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 2015;**527**:384–8.

14. Chaires JB. Allostery: DNA Does It, Too. *ACS Chem Biol* 2008;**3**:207–9.

15. Liu J, Nussinov R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLOS Comput Biol* 2016;**12**:e1004966.

16. Kim S, Brostromer E, Xing D *et al.* Probing Allostery Through DNA. *Science (80-)* 2013;**339**:816–9.

17. Newman M, Strzelecka T, Dorner LF *et al.* Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science* 1995;**269**:656–63.

18. Luisi BF, Xu WX, Otwinowski Z *et al.* Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* 1991;**352**:497–505.

19. Xu X, Ge H, Gu C *et al.* Modeling Spatial Correlation of DNA Deformation: DNA Allostery in Protein Binding. *J Phys Chem B* 2013;**117**:13378–87.

20. Sarah A. Harris †, Evripidis Gavathiotis ‡, Mark S. Searle ‡ *et al.* Cooperativity in Drug−DNA Recognition: A Molecular Dynamics Study. 2001, DOI: 10.1021/JA016233N.

21. Ivani I, Dans PD, Noy A *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat Methods* 2016;**13**:55–8.

RNAs are the latest challenge of structural biology. Single-stranded RNA has the ability to fold onto itself and form highly complex structures that are diverse both in conformational and functional spaces.

Because of their particular chemistry, size and flexibility, determining and understanding their 3D architecture has been difficult for both experimental and theoretical researchers. We thought it a very adequate time to review the most notable directions that computer models of RNAs are taking and to predict the impact that new trends might have on the determination of the physical properties of RNA. We addressed individually recent computational approaches and point out their strengths and weaknesses as well as the lessons from the past that prompted their development. We followed a systematic description from highly accurate QM models specifically applicable to small systems, to classical atomistic representations of MD, CG models, less accurate, but able to deal with large models and finally the trending bioinformatics approaches.

Ab initio QM approaches offer a very accurate description of small RNA systems, but their scope is limited, due to their high computational cost, which constrains them typically to the development and validation of MD RNA force fields. Such force-fields are at the core of molecular dynamics simulations that have been widely used to study RNA properties. However, our detailed analysis highlight major shortcomings of all current RNA force-fields, which might be related to its over-training to reproduce duplex A-RNA or even small inaccuracies in the QM models to which they are fitted. More effort seems necessary to develop new atomistic RNA force-fields able to describe the entire flexibility space that can be covered by this molecule.

All-atom MD simulations with explicit solvent are still very computationally expensive for managing extensive sampling of large and flexible RNAs. For that reason, a common approach is to reduce the resolution of the particle representation from atoms to beads, together with a potential function that either retains the essential physics or is based on a simplified formulation. It is not as straightforward as in MD to develop potentials that are accurate in describing interactions between beads, but simple harmonic terms, sometimes coupled with empirical terms work sufficiently well at this scale. The bottom line is that CG models are a trade-off between accuracy and calculation speed.

Taking advantage of the fast expanding database of experimental structural and sequence information, together with other observables determined from experiment, there is a big trend of trying to extract convoluted relationships between all these data that will ultimately produce dependable predictions for 3D assembly of RNA, regardless of the physical basis of its folding. Sequence information can be used based on evolutionary constraints to retain similar structures of functionally related RNAs. Additionally, recurrent structural information can be implemented to train algorithms to recognize structural patterns (of various scales) commonly seen in experimental structures of RNA.

Publication:

Pablo D. Dans, Diego Gallego, Alexandra Balaceanu, Leonardo Darré, Hansel Gómez, and Modesto Orozco; Modeling, Simulations, and Bioinformatics in the service of RNA Structure. (accepted to Chem (Cell-Group)

MODELING, SIMULATIONS, AND BIOINFORMATICS AT THE SERVICE OF RNA STRUCTURE

Pablo D. Dans,[1,2] Diego Gallego,[1,2,3] Alexandra Balaceanu,[1,2] Leonardo Darré,[1,2,4]
Hansel Gómez,[1,2] and Modesto Orozco[1,2,3,]*

[1] Institute for Research in Biomedicine (IRB Barcelona), the Barcelona Institute of Science and Technology, 08028 Barcelona, Spain.
[2] Joint BSC-IRB Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Spain.
[3] Department of Biochemistry and Biomedicine, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain.
[4] Functional Genomics Lab, Institut Pasteur of Montevideo, Montevideo, Uruguay.

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

SUMMARY

While chemically close to DNA, the RNAs can adopt a wide range of structures, from regular helices to complex globular conformations showing a complexity similar to that of proteins. The determination of the structure of RNA molecules, crucial for function understanding, is severely handicapped by their size and flexibility, which makes difficult the systematic use of experimental approaches. Simulation techniques are suffering also of very severe problems, related to the accuracy of the methods and their ability to sample a large and complex conformational landscape. Recent approaches created to reduce the shortcoming of the current generation of simulation methods will be reviewed here, following a systematic description from highly accurate models able to deal with small systems, to coarse grained approaches, less accurate, but applicable to deal with large models.

INTRODUCTION

What have we learned about RNA structure from QM and QM/MM methods?

Physics teaches as that *ab initio* quantum mechanics (QM) can represent with high accuracy any biomolecular system, among them RNA. Unfortunately, due to their computational cost, the practical application of *ab initio* QM formalisms to large systems, such as RNA, is often impossible. In fact, even simpler QM methods, like those based on the density functional theorem (DFT) fail to treat systems larger than 10^2 atoms, several orders of magnitude less than the size required to study RNAs in solution.[1] Further simplifications of the basic QM formalism, like those implicit in semiempirical (SE) methods can extend the range of applicability of QM theory, but at the expense of an expected loss of accuracy.[2]

High level QM and DFT calculations have had a central role in the development and validation of recent RNA force fields (FFs; see below). A recent example is the B97D3/AUG-CC-PVTZ study of the backbone and glycosidic torsions by Aytenfisu *et al.*[3] which highlights systematic errors in current RNA FFs which might lead to incorrect molecular dynamics (MD) trajectories. Different conclusions were reached by the group led by Šponer´s group using again DFT calculations as reference conclude that the errors in current RNA FFs are related to imbalanced hydration and not to intrinsic errors in the classical gas phase Hamiltonian.[4] The same group has recently studied 46 different backbone conformations of the UpU dinucleotide step (see Figure 1) using a variety of QM methods from the state-of-the-art CCSD(T) to the last-generation SE algorithms,[5] providing the community with an invaluable dataset for refinement of RNA FFs. Very recently our group has used for the first time DFT/MM (density functional theory/molecular mechanics) calculations to fit some dihedrals directly for QM calculations in solution, opening a new approach to use DFT calculations in the refinement of RNA FF (see next section).[6]

Hobza and coworkers have led the use of high level QM methods for the description of nonconvalent interactions in biomolecules, including nucleobases. The latest contributions from the group include the construction of reference databases of interaction energies computed at very high QM level. These databases are very useful for the parametrization and validation of lower level methods, including force fields. To get a deeper look into this benchmark calculations we recommend the last review published by this group.[7,8] On the other hand focusing our in the last couple of years, we should cite the DFT study by Rypniewski *et al.*[9] of the C-U and U-U pairings, where the authors suggest that unusual tautomeric forms, or even anionic states of pyrimidines can play a role in stabilizing certain forms of RNA. Similarly, Preethi*et al.*[10,11] used high level DFT methods to study the impact of post-transcriptionally modifications in the base pairings occurring in different RNA motives (interface, rRNA, and the intron-exon complexes). In another nice study Wilson *et al.*[12] analysed the 154 non-redundant RNA-protein π interactions observed in PDBwith the M06-2X[13] functional, finding that these π-π interactions provide a huge stabilization to the protein-RNA complex. Unusual interactions affecting RNA nucleobases have been also studied by means of high-level QM theory. For example, Chawla *et al.*[14] used high-level QM theory (up to CCSD(T)) to explore the interactions between the O4' atom and the π cloud of the nucleobase, finding that this apparently irrelevant interaction can, in fact, significantly affect the packing of RNA. Cation-RNA interaction has been another traditional field for the use QM theory. A recent example of this family of studies was published by Casalino *et al.*[15] who used DFT calculations to study typical Mg^{2+}-RNA binding motifs, providing a useful benchmark set to develop next-generation of Mg^{2+}-adapted FFs.

Catalysis is another field where the use of QM is strictly necessary since it typically implies the restructuration of bonds and electronic effects that classical FFs cannot handle. In that sense, QM/MM calculations have become the standard approach for the study of ribozymes.[16-27] Several of these studies focused on the role of metallic ions or cofactors in catalysis.[21-27] Other studies focused on the use of QM methods to understand complex experiments like those based on the measurement of kinetic isotope effects (KIEs).[28] Very interestingly, other non-ribozyme RNAs can also exhibit catalytic properties, as it has been descrived for unpaired nucleotides in non-catalytic RNAs.[29] Mlýnský and Bussi[29] published a nice study where by using QM(DFTB3[30]) / MM calculations in the context of enhanced sampling methods the characterize the pattern of reactivity of specific RNA motifs like the uGAAAg tetraloop.

It is difficult to predict the impact of QM calculations on RNA in the next decade, but expectation exist that a new generation of QM methods using new partitioning schemes would allow the representation of more realistic segments of RNA. For example, Jin *et al*.[31] have shown good representations of 15-mer RNAs using an electrostatically embedded generalized molecular fractionation with the conjugated caps (*i.e.* EE-GMFCC) method. Additional expectations arise from the generation of simplified QM approaches, such as the density functional tight binding (DFTB) or RNA-adapted SE approaches. As an example, Huang *et al*.[32] have recently introduced a multidimensional B-spline correction map (BMAP) to the sugar puckering in the AM1/d-PhoT[33] semiempirical Hamiltonian, which was successfully applied to reproduce different RNA transesterification reactions. Similarly, York's group has presented exciting results on RNA systems using quantum mechanical force fields (QMFFs),[34] which scale linearly with system size and are then much faster than fully-coupled QM methods.

Structure and dynamics of RNA molecules as seen from the classical world: What is in charge?

Despite recent advances in computers and in simulation tools, there is no expectation that QM methods will be able to deal (at least in the next decades) with even medium-size (10^2 nucleotides) RNAs in solution. This has fueled the development of MM-methods (like molecular dynamics, MD) based on atomistic classical FFs. The most popular RNA FFs are those originated from the AMBER (*Assisted Model Building with Energy Refinement*) and the CHARMM (*Chemistry at Harvard Macromolecular Mechanics*) communities. Although both rely grossly on the same formulation of the potential energy functional, they differ in the parameterization strategy (see Vangaveti *et al*.[35] and references therein). The four decades of healthy competition between CHARMM- and AMBER- developers have promoted a dramatic advance in the field. Nevertheless, it would be highly desirable that other communities join the race too, which seems to be the case of the OPLS (*Optimized Potentials for Liquid Simulations*)- one which has recently published a careful calibration of the torsions of nucleosides and nucleotides from high level DFT calculations and NMR observables.[36]

During the last decades, the evolution of both AMBER and CHARMM RNA force fields has been fueled by the ever-growing computational power, pushing the time-length boundaries of Molecular Dynamics (MD) trajectories. The extension of trajectories has made evident errors not visible in shorter simulations. For example, in the case of the AMBER-community (Figure 2) the 94 and 99's force fields (AMBER-ff94 and AMBER-ff99) were used for almost 2 decades, until significant errors emerged in long-scale simulations. This forced the development of new parameters aimed mainly at refining specific torsion angles (AMBER-ff99-BSC0-χ_{OL3},[37,38] AMBER-ff99-χ_{YIL},[39] and AMBER-ff99-TOR[40]), and certain non-bonded terms[41]. The CHARMM-community has mainly focused on the representation of proteins and lipids for many years and the evolution of the nucleic acids version has been slower (Figure 2). CHARMM36 RNA parameters[42] were a major advance for this community, as it corrected major caveats of previous versions (CHARMM22[43] and CHARMM27[44]), allowing then for more reliable simulations of different types of RNAs. Worth mentioning is that both AMBER and CHARMM FFs have been extended to account for non-coding nucleotides (more than 100 variants are available),[45,46] opening the possibility to study epigenetic changes in RNA and extending FF-based calculations to the study of non-natural nucleic acids.

Despite the titanic efforts of the CHARMM- and AMBER- communities, several errors in FF-based simulations still persist. For example, high populations of non-native stacking conformations,[47] or significant thermodynamic unbalances between the folded and unfolded states.[48] These difficulties might point toward fundamental problems in current RNA FFs. For example, unbalanced π-stacking[49] and/or hydrogen bond interactions,[47,50] improper hydration of RNA functional groups,[50] or even fundamental problems in the pair-wise additive potential formalism. This has boosted yet another round of parameterization efforts (highlighted in gray on the timeline of Figure 2), using in some cases renewed methodological approaches. The field is now in an exciting, but also turbulent phase, and it is not trivial, even for an expert, to decide the combination of patches to add to the default FFs. We will use the next lines to provide the readers with some clues for selecting the best FF for his/her particular problem.

The water model. Improving the water model is one possible way to change the balance between hydrogen bonding and stacking of the bases and correct some of the known caveats of current RNA FFs. Advances in this direction were presented by Bergonzo and Cheatham in 2015,[50] where the AMBER-ff99-BSC0-χ_{OL3} FF was combined with four different water models (TIP3P,[51] SPC/E,[52] TIP4P-Ew[53] and OPC[54]) to study the conformational landscape of the rGACC tetramer. The standard water model in CHARMM is TIP3P$_{CHARMM}$ (TIP3P modified to include non-zero LJ terms for hydrogen atoms), which has been explicitly used in the force field charge parameterization. As a consequence, CHARMM-based simulations are, in principle, restricted to the TIP3P$_{CHARMM}$ model. However, recent efforts in protein force field development,[55] suggest that changing off-diagonal LJ interactions (mainly the dispersion component) between water hydrogen atoms and the protein (without affecting water-water interactions) improves the protein compaction and folding compared to experimental data. These results suggest a possible pathway for RNA force field improvement focused on the water model.

In any case, the modification of the water model is always a risky decision as current water models have been validated in thousands of studies (just TIP3P model collects more than 24,000 citations). A RNA-tuned water model might lead to very strong links between RNA FF and water model, and to potential problems in the transferability of the resulting water model.

Scaling the phosphate LJ parameters. A further step to improve nucleic acid interactions was the modification of the LJ parameters of charged phosphates. Based on the work of Steinbrecher *et al.*,[56] on bio-organic phosphates, a ~5% increase of the van der Waals radii in RNA phosphate oxygen atoms (OP1/2, O5' and O3') was proposed , which (when combined with the OPC model) improved the representation of the conformational ensemble of the rGACC tetramer.[50] Unfortunately the improvement was not transferable to rCCCC.[50] Pak and coworkers suggested that the previous correction (named vdW$_{bb}$) could artefactually weaken phosphate hydration, developing an alternative LJ correction (vdW$_{YP}$), based on differential pair-dependent Lorentz-Berthelot combination rules.[57] More precisely, the vdW radii of the OP1/2 phosphate atoms and the O2' were scaled up by 5%, but only for intra-molecular interactions, whereas the original unscaled vdW radii were used for the interaction with water. This parameterization, combined with the OPC water model, was tested in four tetramers: rGACC, rCCCC, rAAAA, and rCAAU, showing good results for rCCCC and rAAAA, while for rGACC the minor conformation reported in NMR was not reproduced and artefactual non-native structures were obtained for rCAAU.[57] The FF variation suggested by Pak and coworkers failed also to correct the problem of the low melting temperature of the UUCG tetra loop hairpin.[56] Altogether, it is clear that the correction of phosphate non-bonded potential is useful, but it is unable to correct all problems of last generation RNA FFs.

Scaling the nucleobases LJ parameters. A recent work from Pak and coworkers[58] suggests a correction to the LJ parameters of the nucleobase nitrogen and oxygen atoms: the vdW radii are scaled down by 2.5% for intra-molecular interactions while keeping the unscaled parameters for nucleobase-solvent interactions. This approach, which reinforces base pairing was successfully applied to fold the thrombin-binding DNA aptamer G-quadruplex,[58] and we might expect that the patch would be transferable to RNA. Focusing in the re-calibration of π-stacking interactions, Chen and Garcia,[41] modified heavy atom LJ parameters by a slight (5%) reduction in the vdW radii combined with a 20%/10% reduction in the vdW well for nucleobase/nucleobase and nucleobase/water respectively. By using this scaling strategy, they reported reasonable agreements to CCSD(T) stacking energies in the gas phase, and with aggregation constants and CD spectra in solution. This LJ modification, combined with a re-parameterized glycosidic torsion, allowed the authors to fold, for the first time, two hyper stable tetraloops (rUUCG and rGCAA). However, it failed to fold the rCUUG tetraloop,[41] and presents problems in tetra loop/ tetramer conformational preferences,[47] and kissing loop structural stability,[59] showing again that RNA FF re-calibration is more complex than anticipated.

Refitting RNA backbone dihedral angles. As an extension to AMBER-ff99-χ_{YIL}[39] D. Wales and I. Yildirim recently focused on re-fitting the α/γ torsion pair, generating an upgraded version called AMBER-ff99-χ+α/γ,[60] (homologous to AMBER-ff99-bsc0-χ_{OL3} regarding the identity of the refined torsions). A substantial difference in the parameterization approach compared to BSC0 resides in the molecular model used to generate the Quantum Mechanics (QM) and MM potential energy profiles. Such model consisted in an RNA di-nucleotide where the nucleobase is substituted with a methyl group, and in two possible conformations, one where both sugar rings are in 3'-endo, and the other where the 5' sugar is in 3'-endo while the 3' sugar is in 2'-endo (the rest of the torsions being kept in canonical A-form). This parameterization improved substantially the conformational preferences of several tetramers by strongly suppressing the occurrence of non-native stacked conformations. However, major errors are visible for other tetramers such as rCAAU, and rAAAA, as well as for the UUCG loop, where sampled structures are incompatible with NMR data.[60] Following a more disruptive approach, Aytenfisu *et al.*[3] simultaneously refined α, β, γ, ε, ζ and χ torsions taking as reference DFT(B97D3/aug-CC-PCTZ) potential energies of a database of nucleosides and dinucleotides, extracted from X-ray structures and supplemented with dihedral scans to sample barriers (>31,000 structures). This approach combines enhanced conformational diversity and torsion correlations. The obtained parameterization, which represents a step forward towards the implementation of automatic approaches for FF-calibration, reduces artifact conformations, favoring A-form-like structures for several tetramers (rGACC, rCCCC, rAAAA and rCAAU), but fails for others, leading for example to bad geometries and stabilities for the rUUCG hairpin loop.

Small details matter. Contrary to DNA, RNA sample a wide range of non-canonical conformations where unexpected details can bias the conformational ensemble breaking the block-transferability principle implicit in FF development. An example was recently reported by Darré *et al.*[6] who combining QM/MM calculations and database analysis described the existence of an unexpected coupling between the 2'OH conformation and sugar puckering. The results highlighted a potential mechanism for protein-induced sugar re-puckering in RNA and demonstrated the need to recalibrate the C2'-O2' torsion and to treat in a different way central and terminal nucleotides.

Empirical potentials. A remarkable shift from canonical RNA FF refinement has recently been proposed by Bussi and coworkers who used experimental data not only for validation, but directly in the fitting of parameters. One example is the elegant combination of enhanced sampling simulations with NMR experimental data, in the framework of the maximum entropy principle with explicit treatment of experimental uncertainties.[61] In their approach, nucleosides (A and C) and dinucleotides (ApA, ApC, CpA, and CpC) J^3-couplings were simultaneously used to generate chemically consistent perturbations to AMBER-ff99-BSC0-χ_{OL3}. The corrected potential was shown to be portable to rAAAA and rCCCC, notably reducing the occurrence of artifacts in previous simulations. Another example from the same group is the use of torsion preferences taken from high resolution (<3 Å) X-ray structures in the Protein

Data Bank (PDB), as reference distributions for RECT-Target-Metadynamics (T-MetaDyn) simulations, from which corrections to AMBER-ff99-BSC0-χ_{OL3} potentials were suggested.[62] Such corrections work well in reproducing NMR observables for rGACC and rCCCC, but fails for example for rAAAA, highlighting potential problems in FF transferability. In any case, irrespective of the success/failure rate, Bussi's work represents a proof of concept of a novel force field refinement approach departing from the pure QM-based parameterization strategy that has dominated the area for a couple of decades.

Reformulating the electrostatics? Until very recently the electrostatic component of the force field was considered a *taboo* and no group wanted to modify the set of charges appearing in the 90's versions of the force field. The reasons are multiple: i) while for other parameters improving the level of the reference QM calculation leads typically to better parameters, this is not the case for charges, where for not well-known reasons the HF/6-31G(d) level provides the best "effective charges", ii) changing the charges should lead to a complete re-parameterization of all the torsional terms in the force field (something most groups try to avoid), and iii) the modification of the charges can modify in an unpredictable way the RNA-solvent interactions. Very recently Shaw's group has presented a new RNA FF[63] which includes a minimum alteration of AMBER nucleobase charges, which along recalibration of LJ and several torsion terms, aims at improving nucleobase stacking, base pairing and key torsional conformers. This upgraded force-field in combination with the TIP4P-D water model,[64] seems to work quite well for several RNAs. However, caveats of this force-field are clear, such as certain over-stabilization of helical regions, which possibly contribute to the deviation from NMR data observed for rUUUU, and to the higher melting temperatures of the rU_{40} hairpin, the RNA duplexes and the tetraloops.[63]

More disruptive approaches explicitly introducing polarization in the force field have been followed by other authors. MacKerell and the CHARMM group were pioneers in this field, suggesting the first nucleic acids polarizable FF,[65] which has been recently revised[66] and which works well, at least for DNA.[67] Very recently, a nucleic-acid version of the AMOEBA FF[68] was published by the groups of Ponder and Ren, that may be also useful for the study of RNAs. Most importantly, both groups have dedicated important efforts in making polarization models computationally efficient, pushing for a wide-spread use of such force fields in the future.[69] We may speculate that when fully refined these force-fields will become the defult for the simulation of medium-size RNA structures.

The risk of overtraining. The description presented in the previous paragraphs highlights the problems of force-field modifications performed to improve a given RNA system, as the fitted parameters can fail to represent many others. This is a well-known risk in bioinformatics named "overtraining", which in classical force-field appears a lack of transferability. The standard approach in FF development to achieve transferability and to reduce then overtraining has been to focus on small model systems, but experimental data on small RNA models is scarce and often of poor quality. Even for medium sized systems, the reference experimental data that FF developers can use is insufficient, for example, we have NMR data in solution of only a small number of 4-mers and the data collected from them is quite limited (mostly sugar J-couplings and a few NOEs). These data might be enough to determine if a FF-based simulation is incorrect, but are unable to guide a full parameterization process. An effort from experimentalists providing data useful to guide FF parameterization would be extremely useful and highly appreciated in the field.

Finally, we should note that the use of QM data as reference (the default in the last decades of FF-refinement) is suffering the problems of sampling solvent-relevant conformations at the QM level. Current QM procedures implement continuum SCRF descriptions of the solvent,[70] which are known to be not accurate enough when solvent-solute boundary is not well defined (for example in the case of systems with potential intra-molecular hydrogen bonds).[70] Most FF-developers have faced this problem by limiting the region of the QM-explored conformational space to the biologically relevant one, an approach that has been extremely successful for DNA FF parametrization.[71] However, for a conformationally-promiscuous molecule such as RNA this is another source of overtraining as canonical structures might be too prevalent and unusal conformations might be poorly described. Clearly, the use of MD simulations based on QM/MM Hamiltonians seems a good choice for future FF refinements,[6] but here the QM level (typically DFT) need to be well calibrated to guarantee that complex interactions such as dispersion are well reproduced and QM/MM simulation should be long enough as to guarantee correct sampling of all degrees of freedom.

Clearly, RNA FFs need to be improved, but, despite their caveats, they have allowed a significant advance in our understanding of RNA. We will discuss next a few recent examples of the successful use of RNA FFs, addressing the reader to recent reviews by Bussi and coworkers 2018,[72] and Šponer and coworkers 201[73] for a more comprehensive revision of RNA FFs applications. One area where FFs have shown certain degree of success is in the study of particular RNA motifs. One example of this type of work was recently published by Yildirim *et al*.[74] who predicted the structure and thermodynamics of the 1x1 internal loop in CUG repeats based on MD and discrete path sampling (DPS) calculations, finding a complex conformational scenario characterized by a dual base pair scheme explaining the binding mode of drugs active against myotonic dystrophy.[75] Similar techniques were used to determine the stability of the bioactive form of anticodon stem-loops of tRNA^{GLY} iso-acceptors when the specific modification G/C$_{34}$↔A$_{34}$ (first position in the anticodon loop) is introduced.[88] MD simulations complemented with experimental measurements demonstrated that the presence of A$_{34}$ kills tRNA functionality explaining the emergence of ADAT (adenosine deaminases that catalyzes the conversion of A$_{34}$ to I$_{34}$). The latter is prevalent in eukaryotes and explains the promiscuity (C,U,A) in the recognition of the last position in the codon.[76] Another remarkable study is the thorough analysis of the binding of Mg^{2+} to RNA, presented by Cunha and Bussi.[77] In this work, accurate binding affinities of Mg^{2+} to all possible sites on an RNA duplex is obtained by means of a trapping-penalized version of the bias-exchange metadynamics approach[78] using AMBER-ff99-BSC0-χ_{OL3}. Worth noting, restraints were applied to both double and single-stranded RNA segments, aimed partially at avoiding force field artifacts. The effects of ion competition and hybridization were well reproduced, and furthermore, RNA conformational entropy was shown to affect cation binding in a site-specific (phosphate or nucleobase) manner. The distribution of Mg^{2+} cations around RNA was also addressed by Lemkul *et al.*[79] using a novel technique that concatenates explicit solvent Grand Canonical Monte Carlo (GCMC) simulations with short MD simulations, allowing the determination and refinement of Mg^{2+} binding sites. Application of this method, in combination with the CHARMM36 force field (using harmonic restraints on the phosphorus atoms), to four challenging RNA structures: a pseudoknot, a ribozyme stem-loop, a 23S rRNA and a Mg^{2+} riboswitch, predicted both inner- and outer-shell Mg^{2+} coordination in agreement with the experimental data. Another study focusing on ion-RNA interactions is that of Havrila *et al.*[80] where the effect of different Na^+ and K^+ parameterizations on the structure of guanine quadruplexes (GQ) and the interchange between ions within the GQ channel and the bulk is evaluated. Also worth to mention is a very recent paper from the group of Cheatham where a reference protocol to address conformational variability in the smallest RNA units *i.e.* dinucleotide monophosphates, is proposed.[81] The authors prove that sampling convergence is achieved at the half microsecond time-scale for such systems when using T-REMD using 18 replicas spanning the 280-396K range. Using this approach in combination with different RNA force fields: AMBER-ff99-BSC0-χ_{OL3}, (with and without vdW_{bb} phosphate corrections), Chen&Garcia, and CHARMM36, and several water models (TIP3P, TIP4PEW, $TIP3P_{CHARMM}$ and OPC), a consensus of five main conformers emerged from the structural sampling of all possible dinucleotides. These conformers and their associated torsional and non-bonded interactions characteristics constitute a reference for future experiments and force field refinements. Also worth mentioning is the protocol for identifying SHAPE reactivity nucleobases in RNA proposed very recently by Mlýnský *et al.*[82] The work not only provided atomistic detail in the dependence between SHAPE reactivity and RNA flexibility, but also holds potential for RNA 3D structure prediction and validation using SHAPE data.

The known shortcomings of current RNA FFs have encouraged several groups to supplement theoretical calculations with experimental restrains. A recent example is the nice work of Vendrusculo and coworkers,[83] which revisited TAR (HIV-1 protein trans-activator of transcription) dynamics, in the TAR-Tat complex (Tat: trans-activation response RNA element). The authors were able to characterize for the first time a low-population intermediate structure by performing Replica-Averaged Metadynamics (RAM) simulations biased by NMR residual dipolar couplings. Another nice example is a recent contribution from the group of Al-Hashimi where a combination of NMR, UV-melting, QM calculations and MD simulations was used to study the occurrence of WC↔HG transitions in A-RNA.[84] Their results clearly demonstrate that the WC↔HG transition is much more difficult in A-RNA than in B-DNA and that while m^1A (a common post-transcriptional modification in adenine), is easily accommodated in B-DNA through Hoogsteen pairing, it leads to the disruption of the duplex in RNA. Finally, rather recently Bottaro *et al.*[85] proposed a method to accurately reconstruct RNA conformational ensembles using the maximum entropy/Bayesian approach to reweight MD simulation ensembles in order to fit NMR experimental data (*i.e.* NOEs and scalar couplings). This elegant approach not only notably reduced force field artifacts, but also improved the interpretation of the NMR data, providing a better picture of RNA tetranucleotides conformational landscape.

What are we losing and what are we gaining on coarse-graining RNA?

Coarse grain (CG) is an ambiguous term used to label a family of models, which allow overstepping the practical limits of the atomistic models by simplifying the representation of the model and/or the complexity of the potential energy functional. Contrary to the situation found for DNA, most CG models for RNA are particle based (pbCG), simplifying the description of the nucleosides by fussing several atoms into a single bead.[73,86-88] The energy functional used to reproduce bead-bead interactions can be defined based on physical or statistical considerations. In the first case, the energy is computed as the addition of terms accounting for pseudo-bonds, pseudo-angles, pseudo-dihedrals (and in some cases pseudo-non-bonded) interactions. IFoldRNAv2,[89-91] TOPRNA,[92] HIRE-RNA[93] and the MARTINI FFs[94] are examples of this type of potentials which in all cases are calibrated to reproduce known experimental observables. In the second case, the statistical (knowledge-based) potentials rely on the statistical analysis of the Protein Data Bank or related databases to derive the frequencies of occurrence of particular interactions from which an inverse Boltzmann transformation yields effective “energies”. NAST[95,96], oxRNA[97], simRNA[98], and RNAkb[99] are popular examples of methods implementing such statistical potentials. Methods such as YUP[100], RACER[101] or NARES-2P[102] combine both physical and statistical potentials. In any case, irrespectively of the method used to define the energy functional, sampling of the conformational space needs to be obtained in order to explore potential conformations of the RNA. For this purpose, most models (NAST, NARES-2P, iFoldRNAv2, TOPRNA, RACER, RNAkb, HIRE-RNA, and MARTINI) use Molecular Dynamics (MD) machinery, others implement Monte Carlo (MC) algorithms (YUP, FARNA[103], and oxRNA), Dokholyan’s group uses discrete Molecular Dynamics (Dmd[89], iFoldRNAv2 and iFoldNMR[104]), and finally the authors of simRNA implement a variety of sampling engines in their codes.

Beside the generality that all these models could share, each of them has been developed for a given purpose and potential users must be aware of their intrinsic limitations before applying them. The following lines will be devoted to summarize the characteristics of some of the most popular CG models (Figure 3 and Table 1 for a summary). To facilitate the discussion, the models are ordered from the most detailed ones, *i.e.* the closest to all-atom representations, to the coarsest ones. The widely used MARTINI[94] CG-FF has now a RNA version created to integrate with CG descriptions of proteins, carbohydrates and lipids. In the RNA MARTINI model, each base is represented by either 6 (pyrimidines) or 7 (purines) beads. The model assumes knowledge of the secondary structure and has demonstrated a good ability to reproduce the structure of a variety of ribosomal and tRNAs. HIRe-RNA[93] is a high resolution model that also uses 6(Pyr)-7(Pur) beads per base, and physical potentials including base pairing, stacking and electrostatics terms. Using their latest version (HIRE-RNA v3), the authors were able to obtain reliable trajectories for many small RNAs. simRNA[98] is a medium resolution model using 5 beads per residue and statistic interaction potentials. The model has been able to reproduce structure and dynamics of medium (up to 190 nt) RNAs, especially when secondary structure is fixed by experimentally-derived restraints[105]. Three other models used 5 beads x base: RNAkb, RACER, and oxRNA. RNAkb[99], is based on a statistical potential and was trained to distinguish between folded RNAs and decoy conformations, showing good results for small RNAs. RACER[101] also uses a hybrid statistical/physical potential which include terms accounting for excluded volumes, electrostatic, and H-bonding of the bases. The method works well for short (<30 nt) RNAs and when experimental data are supplemented, also good results are obtained for medium-sized RNAs (<100 nt). oxRNA[97] also uses 5 beads for each residue, and a dual potential function (different for neighbor and non-neighbor pairs) which was calibrated to reproduce well, not only the structure, but also the thermodynamics of folding of short RNA motifs. Contrary to most of the previous methods that are coupled to MD, oxRNA relies on a MC algorithm for sampling. iFoldRNAv2[89-91] is a coarser model that uses 3 beads representation per residue, and a physical potential including electrostatic, hydrogen bonding and stacking terms. When experimental data is incorporated to the MD sampling methods, the model provides good results for RNA segments up to 200-nt long[90]. iFoldNMR is the latest published model from Dokholyan's group[104]. It uses 3 pseudo beads per residue, with bead interactions described by statistical potentials, and taking advantage of sparse NMR constraints to guide the 3D folding of small to medium RNA fragments in dMD simulations. TOPRNA[92] also considers 3 beads per residue with CHARMM[106] equations parametrized to reproduce known structures. The method assumes previous knowledge of secondary structure and works well for small RNAs when coupled to an MD engine. NARES-2P uses 2 interaction sites and a dipole moment for each nucleotide. [102] In this model the backbone conformation is governed by a statistical potential fitted to reproduce experimental conformational ensembles and heat-capacity curves. The model uses MD simulations to explore the conformational space, and was quite accurate to reproduce properties of small RNAs. Very recently, Bussi and coworkers developed another CG model using 2 beads per residue (choosing an anisotropic particule to represent the nucleoside), called SPQR[107]. The method uses a knowledge-based potential plugged into a MC algorithm, and successfully fold small to medium RNA molecules. Finally, the lower resolution models (NAST, FARNA, YUP-rrRNAv1, and RS3D) use only 1 bead per residue. NAST[95,96] locates the bead at the C3' using statistical potentials supplemented with information on secondary structure, tertiary contacts derived from co-evolution analysis, and eventually with information derived from SAXS (small angle X-ray scattering) or chemical probing experiments. Despite its simplicity, the method coupled to MD simulations works well in the prediction of medium-sized RNA. FARNA[103] uses a fragment library of trinucleotides and a statistical potential coupled to a MC algorithm, showing a good ability to reproduce short RNA motifs. YAMMP/YUP-rrRNAv1[100] locates the bead at the center of mass of each residue using simple harmonic terms for bonds, angles, dihedrals, and non-bonded van der Waals interactions showing a good ability to reproduce tRNAs. Finally, R3SD from Wang and collaborators is applicable to a wide range of RNA folding motifs for which experimental data are available[108], being based on SAXS and solvent accessibility data to build models with a resolution of 1 bead per residue. Sampling is achieved by coupling the potential to a Metropolis Monte Carlo algorithm that must also satisfied secondary structure constraint, in addition to any tertiary structure restraint. R3SD is able to reproduce with good accuracy common RNA folds in small-to-medium RNA molecules.

As noted above, most of the CG methods include the possibility to introduce experimental restraints to avoid the sampling of artefactual conformations. For example, hydroxyl cleavage[88] or chemical probing techniques such as SHAPE (Selective 2′-Hydroxyl Acylation analyzed by Primer Extension) have been used to derive restraints to fix secondary structure elements[109], and gain some information about long range contacts.[110-112] Recently, SHAPE has been coupled with mutational profiling in live cells,[113] or with single-molecule Forster resonance energy transfer in single cells[114] showing very good ability to guide the sampling of CG models towards experimental structures.[114] SAXS is also used to guide CG models[115] as in the case of NAST and R3SD.[95,96,108] A recent work of Lipfert *et al*., used samples labeled with Gold nano-particles together with SAXS to provide a fine-grain 3D structure of an RNA kink-turn motif[116]. Very recently, SAXS and NMR data were combined to supplement RACER (RnA CoarsE-gRained)[101] accomplishing the folding of a long sequence of RNA like the 5S ribosome, and sparse NMR data alone were used in iFoldNMR to derive high quality models of medium-sized RNA.[104] Light Activated Structural Examination of RNA (LASER), is another novel experimental technique which provide solvent accessibility inside cells at the nucleotide level for medium-sized RNA molecules.[117] Also recently, and already in the frontier between bioinformatics and modeling, different groups have used co-evolutionary data to bias CG-models to establish 3D contacts, as in the case of NAST, or the newest 3dRNA from Xiao and collaborators.[118]

Beyond the evident progress, CG RNA models still face many challenges. One of the most important is the development of parameters for partner biomolecules. Large RNA molecules (one of the aims of coarse-graining) like ribosome are usually associated to several proteins, and in other cases, the description of RNA-DNA complexes is needed to understand the replication and transcription machinery of cells. Ions like Mg2+ are also known to be crucial for the stability of certain RNA motives or the folding of tertiary structures.[73,86] This is why, models like oxRNA and HIRE-RNA that can combine DNA and RNA molecules, NARES-2P and MARTINI that can also include proteins, or RACER with his explicit treatment of Mg2+ ions, have the advantage to consolidate themselves in the near future as the reference for RNA coarse-graining.

Bioinformatics: An inescapable complement to RNA structure prediction?

Even the most efficient CG models face severe problems to deal with large RNA structures, which have fueled the development of bioinformatics methods taking profit of the lessons learned from protein structural prediction. These methods (see Table 1 for a summary) have been traditionally classified in two families: i) homology (comparative) modelling techniques[119,120] which are based on the idea that RNA structure is more conserved than sequence, and ii) approaches based on the assumption that RNA is hierarchically assembled from small structural elements[121]. However, the partition between the two predictive paradigms is nowadays somehow artificial as the most recent pipelines of RNA structural prediction are based on hybrid methods combining homology modeling, hierarchical folding and CG simulation engines.

Methods based on the structure conservation principle use techniques based on homology (or comparative) modeling[119,120] and more recently include evolutionary couplings[96] to capture conserved 3D contacts. ModeRNA,[119] MMB[120] and RNA123[122] are examples of currently available homology modelling programs. Homology modelling has been extremely fruitful in the protein field and has the advantage to be easily scaled up with no general size limit, but in the RNA field they suffer caveats derived from the scarcity of non-redundant RNA structural data. Again, involvement of the experimental community seems necessary in the development of more reliable tools. Methods based on the hierarchical folding hypothesis can use as building blocks either secondary structures[121] (experimentally known or predicted[123,124]) or local mini-motifs (<4 nucleotides) whose geometries are arranged in space to minimize some scoring function.[125,126] Such functions are generally statistical potentials based on pairwise interactions either at atomic or coarse grained levels and focus on an ever-evolving play-off between computational efficacy and accurate energetic description.[127-129] This method has been approached in several ways and here we distinguish between the semi-automated (graphics-based) and the fully automated methods, each category comprising of different approaches.

Semiautomatic (graphics-based) modeling consists of interactive software allowing the user to integrate secondary information on RNA into a graphical interface to generate a 3D model. Early examples of this type of software are MANIP[130], and ERNA-3D.[131] More recent programs such as RNA2D3D[132] or Assemble/Assemble2[133] allow the manipulation and assembly of RNA 3D constructs even when they are complexed with other macromolecules. Furthermore, they integrate databases of structural motifs and multiple sequence alignments and facilitate the incorporation of explicit manual annotation of base pairs and stacking interactions. Graphics-based methods can be used to build large 3D structures in a very simple way, but their performance depends heavily on user experience. Automated algorithms such as RNAComposer,[134] RSIM,[135] MC-Sym,[136] and 3dRNA,[137] require a 2D structure input and sample different 3D models taking a fragment assembly approach. RNAComposer splits the 2D structure in stems, hairpins, loops and n-way junctions and then finds matching elements using the RNA FRABASE database.[138] RSIM assembles the 3D models using a fragment database of trinucleotides and a Monte Carlo approach with biased moves preserving secondary structure. MC-Sym uses nucleotide cyclic motifs for the assembly and was optimized to work in a pipeline with MC-Fold, a secondary structure predictor.[133] Reinharz and coworkers[139] expanded further the capabilities of MC-Sym by developing RNA-MoIP

(Table 2), which identifies RNA motifs in the 2D structure and enhances the MC-Sym sampling process, thus allowing the modeling of bigger oligonucleotides. 3dRNA builds its models from smallest secondary elements (SSEs, defined as base pair, hairpins, i-loops, etc.)[118] (see Table 2). Vfold[140] combines a CG- physics- and knowledge-based approach with a hierarchical folding approach, predicting 2D structure from sequence and then constructing the 3D models with a motif assembly method based on entropy and free-energy estimations. Very recently, the same group launched VfoldLA,[141] using loop fragments as building blocks (instead of whole motifs). With a faster scoring function (based on template sequence similarity), VfoldLA allows the prediction of more structures than its predecessor, but with slightly higher RMSD values on average. Recent methods such as F-RAG/RAGTOP,[142] and Ernwin's predictor,[88] rely on graph theory for guiding the assembly of 3D constructs. Graph representation of secondary structure elements is taken one step further in the GARN and GARN2 packages,[143] where the minimization of an energy function is replaced with a regret minimization algorithm using a knowledge-based scoring potential.

Some authors have tried to skip the need for secondary structure annotation by working with smaller (<4 nts) RNA segments. Thus, Rosetta FARFAR[103] assembles experimental tri-residue fragments iteratively (starting from an initially extended chain) in a Monte Carlo simulation directed by a knowledge-based energy function, which can be further refined by atomistic force-field simulations. BARNACLE[125] uses a dynamic Bayesian network (DBN) to model RNA structures, using a maximum-likelihood estimation of di-nucleotide parameters. TreeFolder[126] uses a conditional random fields (CRFs) method trained with tri-nucleotide data, combined with a tree guided conformation-sampling scheme. These methods are applicable in cases where no secondary structure annotation exists, but their performance is still limited when dealing with long RNAs.

Finally, it's worth mentioning the existence of integrated tools for the complete RNA 3D structure prediction starting from the sequence (see Table 2 for a summary). In many cases, they add a layer of analysis with the quantification of non-canonical interactions or evolutionary-conserved 3D-contacts. RMDetect[144] was one of the first of such packages combining different technologies based on the identification of 3D structural modules. The metaRNAmodule pipeline[145] exploits and extends the capabilities of RMDetect, completely automating the extraction of putative modules from the FR3D database to perform tertiary structure prediction. JAR3D[146] is a probabilistic model that can also find 3D motifs from the sequence and has advantadge of a continuous self-training as new structures are deposited in the RNA 3D Motif Atlas.[147] As noted above, the introduction of co-evolutionary data has revolutionized the field, as it allows to introduce non-local restraints in the derivation of RNA models, and several programs such as FARNA/FARFAR[148] and more recently NAST[95,96] take advantadge of such information (see the previous section, and Figure 3).

We are far from being able to predict the structure of complex RNA motifs, but recent theoretical methods, either alone, or coupled with experimental restrains are providing very encouraging results,[123] visible in the latest rounds of the RNA-Puzzles competition (Figure 4).[149] For example Pyle group[150] recently modeled domains D2 and D3 of RepA (lncRNA of 1600 nucleotides) with a variety of experimental restraints using RNAComposer and finding reasonable results, and Bujnicki group (the developers of ModeRNA[119]) extracted structural fragments corresponding to evolutionarily conserved regions and successfully used this information to constrain the geometry of conserved residues when folding large RNA constructs.

CONCLUSIONS

RNA is a complex molecule with a chemical structure resembling DNA, but with a conformational richness similar to that of proteins. RNA molecules are usually large, flexible, and show a large conformational landscape which makes difficult its experimental characterization, making often theoretical methods the only approach to gain structural information. Unfortunately, development of theoretical methods to deal with RNA structure is handicapped by the multiscale nature of the system, which has forced the development of a myriad of modeling and simulation techniques we have revised here. Compared to proteins, RNA structural characterization is still in its infancy, but advances are evident, even for the most agnostic scientist. With a more coordinated effort, and the involvement of experimentalists providing more comprehensive data for method calibration we may see soon a new generation of methods able to make quantitative predictions on the structure and physical properties of RNA.

AWKNOLEDGEMENTS

This work has been supported by the Spanish MINECO (BIO2015-64802-R; BFU2015-61670-EXP), the ERC Council (SimDNA and 2shRNA) and the H2020 projects BioExcel and MuG. IRB is the recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain). P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) researcher, and P.D.D. and L.D. are SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher.

REFERENCES

1. Smith, L.G., Zhao, J., Mathews, D.H., and Turner D.H. (2017). Physics-based all-atom modeling of RNA energetics and structure. WIREs RNA*8*, e1422.
2. Šponer, J.,Krepl, M., Banáš, P., Kührová, P., Zgarbová, M., Jurečka, P., Havrila, M., and Otyepka, M. (2017). How to understand atomistic molecular dynamics simulations of RNA and protein-RNA complexes? WIREs RNA *8*, e1405.
3. Aytenfisu, A.H., Spasic, A., Grossfield, A., Stern, H.A., and Mathews, D.H. (2017). Revised RNA Dihedral Parameters for the Amber Force Field Improve RNA Molecular Dynamics. J. Chem. Theory Comput. *13*, 900-915.

4. Szabla, R., Havrila, M., Kruse, H., and Šponer, J. (2016). Comparative Assessment of Different RNA Tetranucleotides from the DFT-D3 and Force Field Perspective. J. Phys. Chem. B *120*, 10635-10648.

5. Kruse, H., Mladek, A., Gkionis, K., Hansen, A., Grimme, S., andŠponer, J. (2015). Quantum Chemical Benchmark Study on 46 RNA Backbone Families Using a Dinucleotide Unit. J. Chem. Theory Comput. *11*, 4972-4991.

6. Darré, L., Ivani, I., Dans, P.D., Gómez, H., Hospital, A., and Orozco, M. (2016). Small Details Matter: The 2′-Hydroxyl as a Conformational Switch in RNA. J. Am. Chem. Soc. *138*, 16355–16363.

7. Riley, K.E., Pitoňák, M., Jurečka, P., and Hobza, P. (2010). Stabilization andStructure Calculations for Noncovalent Interactions in ExtendedMolecular Systems Based on Wave Function and Density FunctionalTheories. Chem. Rev. *110*, 5023-5063.

8. Řezáč, J., and Hobza, P. (2016). Benchmark Calculations of Interaction Energies in NoncovalentComplexes and Their Applications. Chem. Rev. *116*, 5038-5071.

9. Rypniewski, W., Banaszak, K., Kuliński, T., and Kiliszek, A. (2016). Watson-Crick-like pairs in CCUG repeats: evidence for tautomeric shifts or protonation. RNA *22*, 22.31.

10. Preethi, S.P., Sharma, P., and Mitra, A. (2017). Structural landscape of base pairs containing post-transcriptional modifications in RNA. RNA *23*, 847-859.

11. Preethi, S.P., Sharma, P., and Mitra, A. (2017). Higher order structures involving post transcriptionally modified nucleobases in RNA. RSC Adv. 7, 35694-35703.

12. Wilson, K.A., Holland, D.J., and Wetmore, S.D. (2016). Topology of RNA-protein nucleobase-amino acid π-π interactions and comparison to analogous DNA-protein π-π contacts.RNA*22*, 696-708.

13. Zhao, Y., and Truhlar, D.G., (2008).The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor. Chem. Account *120*, 215-241.

14. Chawla, M., Chermak, E., Zhang, Q., Bujnicki, J.M., Oliva, R. and Cavallo, L. (2017). Occurrence and stability of lone pair-π stacking interactions between ribose and nucleobases in functional RNAs. Nucleic Acids Res. *45*, 11019-11032.

15. Casalino, L., Palermo, G., Abdurakhmonova, N. Rothlisberger, U., and Magistrato, A. (2017). Development of Site-Specific Mg^{2+}-RNA Force Field Parameters: A Dream or Reality? Guidelines from Combined Molecular Dynamics and Quantum Mechanics Simulations. J. Chem. Theory Comput. *13*, 340-352.

16. Bertran, J., and Oliva, A. (2017) Ribozymes. In Simulating Enzyme Reactivity: Computational Methods in Enzyme Catalysis, I. Tuñón and V. Moliner, ed. (The Royal Society of Chemistry), pp. 404-435.

17. Dubecký, M., Walter, N. G., Šponer, J., Otyepka, M. and Banáš, P. (2015). Chemical feasibility of the general acid/base mechanism of *glmS* ribozyme self-cleavage. Biopolymers, *103*, 550-562.

18. Świderek, K., Marti, S., Tuñón, I., Moliner, V., and Bertran, J. (2015). Peptide Bond Formation Mechanism Catalyzed by Ribosome. J. Am. Chem. Soc. *137*, 12024-12034.

19. Zhang, S., Ganguly, A., Goyal, P., Bingaman, J.L., Bevilacqua, P.C., and Hammes-Schiffer, S. (2015). Role of the Active Site Guanine in the *glmS* Ribozyme Self-Cleavage Mechanism: Quantum Mechanical/Molecular Mechanical Free Energy Simulations. J. Am. Chem. Soc. *137*, 784-798.

20. Casalino, L., Palermo, G., Rothlisberger, U., and Magistrato, A. (2016). Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns. J. Am. Chem. Soc. *138*, 10374-10377.

21. Chen, H., Giese, T.J., Golden, B.L. and York, D.M. (2017). Divalent Metal Ion Activation of a Guanine General Base in the Hammerhead Ribozyme: Insights from Molecular Simulations. Biochemistry *56*, 2985-2994.

22. Lee, T.-S., Radak, B.K., Harris, M.E., and York D.M. (2016). A Two-Metal-Ion-Mediated Conformational Switching Pathway for HDV Ribozyme Activation. ACS Catal. *6*, 1853-1869.

23. Mlýnský, V., Walter, N.G., Šponer, J., Otyepka, M., and Banáš, P. (2015). The role of an active site Mg^{2+} in HDV ribozyme self-cleavage: insights from QM/MM calculations. **Phys. Chem. Chem. Phys.*17***, 670-679.

24. Radak, B.K., Lee, T.-S., Harris, M.E., and York, D.M. (2015). Assessment of metal-assisted nucleophile activation in the hepatitis delta virus ribozyme from molecular simulation and 3D-RISM. RNA*21*, 1566-1577.

25. Thaplyal, P., Ganguly, A., Hammes-Schiffer, S., and Bevilacqua, P.C. (2015). Inverse Thio Effects in the Hepatitis Delta Virus Ribozyme Reveal that the Reaction Pathway Is Controlled by Metal Ion Charge Density. *Biochemistry 54*, 2160-2175.

26. Zhang, S., Stevens, D.R., Goyal, P., Bingaman, J.L., Bevilacqua, P.C., and Hammes-Schiffer, S. (2016). Assessing the Potential Effects of Active Site Mg^{2+} Ions in the *glmS* Ribozyme-Cofactor Complex. J. Phys. Chem. Lett. *7*, 3984-3988.

27. Bingaman, J.L., Zhang, S., Stevens, D.R., Yennawar, N.H., Hammes-Schiffer, S. and Bevilacqua, P.C. (2017). The GlcN6P cofactor plays multiple catalytic roles in the *glmS* ribozyme. Nat. Chem. Biol. *13*, 439-445.

28. Chen, H., Piccirilli, J.A., Harris, M.E., and York, D.M. (2015). Effect of Zn^{2+} binding and enzyme active site on the transition state for RNA 2′-O-transphosphorylation interpreted through kinetic isotope effects. Biochim. Biophys. Acta*1854*, 1795-1800.

29. Mlýnský, V., and Bussi, G. (2017). Understanding in-line probing experiments by modeling cleavage of nonreactive RNA nucleotides. RNA *23*, 712-720.

30. Gaus, M., Cui, Q., andElstner, M. (2012). DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). J. Chem. Theory Comput. *7*,931-948.

31. Jin, X., Zhang, J.Z.H., and He, X.(2017). Full QM Calculation of RNA Energy Using Electrostatically Embedded Generalized Molecular Fractionation with Conjugate Caps Method. J. Phys. Chem. A *121*, 2503-2514.

32. Huang, M., Dissanayake, T., Kuechler, E., Radak, B.K., Lee, T.-S., Giese, T.J., and York, D.M. (2017). A Multidimensional B-Spline Correction for Accurate Modeling Sugar Puckering in QM/MM Simulations. J. Chem. Theory Comput. *13*, 3975-3984.

33. Nam, K., Cui, Q., Gao, J., and York, D.M. (2007).Specific reactionparametrization of the AM1/d Hamiltonian for phosphoryl transferreactions: H, O, and P atoms. J. Chem. Theory Comput.*3*, 486-504.

34. Giese, T.J., and York D.M. (2017). Quantum mechanical force fields for condensed phase molecular simulations. J. Phys.: Condens. Matter *29*, 383002.

35. Vangaveti, S., Ranganathan, S.V., and Chen, A.A. (2017). Advances in RNA molecular dynamics: a simulator's guide to RNA force fields: Advances in RNA molecular dynamics. Wiley Interdiscip. Rev. RNA *8*, e1396.

36. Robertson, M.J., Tirado-Rives, J., and Jorgensen, W.L. (2017). Improved treatment of nucleosides and nucleotides in the OPLS-AA force field. Chem. Phys. Lett. *683*, 276–280.

37. 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.

38. 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.

39. Yildirim, I., Stern, H.A., Kennedy, S.D., Tubbs, J.D., and Turner, D.H. (2010). Reparameterization of RNA χ Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine. J. Chem. Theory Comput. *6*, 1520–1531.

40. Yildirim, I., Kennedy, S.D., Stern, H.A., Hart, J.M., Kierzek, R., and Turner, D.H. (2012). Revision of AMBER Torsional Parameters for RNA Improves Free Energy Predictions for Tetramer Duplexes with GC and iGiC Base Pairs. J. Chem. Theory Comput. *8*, 172–181.

41. Chen, A.A., and Garcia, A.E. (2013). High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. Proc. Natl. Acad. Sci. *110*, 16820–16825.

42. Denning, E.J., Priyakumar, U.D., Nilsson, L., and Mackerell, A.D. (2011). Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. J. Comput. Chem. *32*, 1929–1943.

43. MacKerell, A.D., Wiorkiewicz-Kuczera, J., and Karplus, M. (1995). An all-atom empirical energy function for the simulation of nucleic acids. J. Am. Chem. Soc. *117*, 11946–11975.

44. MacKerell, A.D., Banavali, N., and Foloppe, N. (2000). Development and current status of the CHARMM force field for nucleic acids. Biopolymers *56*, 257–265.

45. Aduri, R., Psciuk, B.T., Saro, P., Taniga, H., Schlegel, H.B., and SantaLucia, J. (2007). AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA. J. Chem. Theory Comput. *3*, 1464–1475.

46. Xu, Y., Vanommeslaeghe, K., Aleksandrov, A., MacKerell, A.D., and Nilsson, L. (2016). Additive CHARMM force field for naturally occurring modified ribonucleotides: CHARMM Potential Energy Function. J. Comput. Chem. *37*, 896–912.

47. Bergonzo, C., Henriksen, N.M., Roe, D.R., and Cheatham, T.E. (2015). Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. RNA *21*, 1578–1590.

48. Bottaro, S., Banáš, P., Šponer, J., and Bussi, G. (2016). Free Energy Landscape of GAGA and UUCG RNA Tetraloops. J. Phys. Chem. Lett. *7*, 4032–4038.

49. Schrodt, M.V., Andrews, C.T., and Elcock, A.H. (2015). Large-Scale Analysis of 48 DNA and 48 RNA Tetranucleotides Studied by 1 μs Explicit-Solvent Molecular Dynamics Simulations. J. Chem. Theory Comput. *11*, 5906–5917.

50. Bergonzo, C., and Cheatham, T.E. (2015). Improved Force Field Parameters Lead to a Better Description of RNA Structure. J. Chem. Theory Comput. *11*, 3969–3972.

51. 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.

52. 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.

53. Horn, H.W., Swope, W.C., Pitera, J.W., Madura, J.D., Dick, T.J., Hura, G.L., and Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J. Chem. Phys. *120*, 9665–9678.

54. Izadi, S., Anandakrishnan, R., and Onufriev, A.V. (2014). Building Water Models: A Different Approach. J. Phys. Chem. Lett. *5*, 3863–3871.

55. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B.L., Grubmüller, H., and MacKerell, A.D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat. Methods *14*, 71–73.

56. Steinbrecher, T., Latzer, J., and Case, D.A. (2012). Revised AMBER Parameters for Bioorganic Phosphates. J. Chem. Theory Comput. *8*, 4405–4412.

57. Yang, C., Lim, M., Kim, E., and Pak, Y. (2017). Predicting RNA Structures via a Simple van der Waals Correction to an All-Atom Force Field. J. Chem. Theory Comput. *13*, 395–399.

58. Yang, C., Kulkarni, M., Lim, M., and Pak, Y. (2017). In silico direct folding of thrombin-binding aptamer G-quadruplex at all-atom level. Nucleic Acids Res. *45*, 12648–12656.

59. Havrila, M., Zgarbová, M., Jurečka, P., Banáš, P., Krepl, M., Otyepka, M., and Šponer, J. (2015). Microsecond-Scale MD Simulations of HIV-1 DIS Kissing-Loop Complexes Predict Bulged-In Conformation of the Bulged Bases and Reveal Interesting Differences between Available Variants of the AMBER RNA Force Fields. J. Phys. Chem. B *119*, 15176–15190.

60. Wales, D.J., and Yildirim, I. (2017). Improving Computational Predictions of Single-Stranded RNA Tetramers with Revised α/γ Torsional Parameters for the Amber Force Field. J. Phys. Chem. B *121*, 2989–2999.

61. Cesari, A., Gil-Ley, A., and Bussi, G. (2016). Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. J. Chem. Theory Comput. *12*, 6192–6200.

62. Gil-Ley, A., Bottaro, S., and Bussi, G. (2016). Empirical Corrections to the Amber RNA Force Field with Target Metadynamics. J. Chem. Theory Comput. *12*, 2790–2798.

63. Tan, D., Piana, S., Dirks, R.M., and Shaw, D.E. (2018). RNA force field with accuracy comparable to state-of-the-art protein force fields. Proc. Natl. Acad. Sci. *115*, E1346–E1355.

64. Piana, S., Donchev, A.G., Robustelli, P., and Shaw, D.E. (2015). Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. J. Phys. Chem. B *119*, 5113–5123.

65. 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.

66. Lemkul, J.A., and MacKerell, A.D. (2017). Polarizable Force Field for DNA Based on the Classical Drude Oscillator: I. Refinement Using Quantum Mechanical Base Stacking and Conformational Energetics. J. Chem. Theory Comput. *13*, 2053–2071.

67. Lemkul, J.A., and MacKerell, A.D. (2017). Polarizable Force Field for DNA Based on the Classical Drude Oscillator: II. Microsecond Molecular Dynamics Simulations of Duplex DNA. J. Chem. Theory Comput. *13*, 2072–2085.

68. Zhang, C., Lu, C., Jing, Z., Wu, C., Piquemal, J.-P., Ponder, J.W., and Ren, P. (2018). AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. J. Chem. Theory Comput. *14*, 2084–2108.

69. Dans, P.D., Walther, J., Gómez, H., and Orozco, M. (2016). Multiscale simulation of DNA. Curr. Opin. Struct. Biol. *37*, 29–45.

70. Orozco, M., and Luque, F.J. (2000). Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. Chem. Rev. *100*, 4187–4226.

71. Dans, P. D., Ivani, I., Hospital, A., Portella, C., González, C., Orozco, M. How accurate are accurate force fields for B-DNA. Nucl. Acids Res., 45, 4217-4230.

72. Mlýnský, V., and Bussi, G. (2018). Exploring RNA structure and dynamics through enhanced sampling simulations. Curr. Opin. Struct. Biol. *49*, 63–71.

73. Šponer, J., Bussi, G., Krepl, M., Banáš, P., Bottaro, S., Cunha, R.A., Gil-Ley, A., Pinamonti, G., Poblete, S., Jurečka, P., *et al.* (2018). RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. Chem. Rev. *118*, 4177–4338.

74. Yildirim, I., Chakraborty, D., Disney, M.D., Wales, D.J., and Schatz, G.C. (2015). Computational Investigation of RNA C U G Repeats Responsible for Myotonic Dystrophy 1. J. Chem. Theory Comput. *11*, 4943–4958.

75. Childs-Disney, J.L., Stepniak-Konieczna, E., Tran, T., Yildirim, I., Park, H., Chen, C.Z., Hoskins, J., Southall, N., Marugan, J.J., Patnaik, S., *et al.* (2013). Induction and reversal of myotonic dystrophy type 1 pre-mRNA splicing defects by small molecules. Nat. Commun. *4*, 2044–2055.

76. Saint-Leger, A., Bello, C., Dans, P.D., Torres, A.G., Novoa, E.M., Camacho, N., Orozco, M., Kondrashov, F.A., and Ribas de Pouplana, L. (2016). Saturation of recognition elements blocks evolution of new tRNA identities. Sci. Adv. *2*, e1501860–e1501860.

77. Cunha, R.A., and Bussi, G. (2017). Unraveling Mg $^{2+}$ –RNA binding with atomistic molecular dynamics. RNA *23*, 628–638.

78. Piana, S., and Laio, A. (2007). A Bias-Exchange Approach to Protein Folding. J. Phys. Chem. B *111*, 4553–4559.

79. Lemkul, J.A., Lakkaraju, S.K., and MacKerell, A.D. (2016). Characterization of Mg $^{2+}$ Distributions around RNA in Solution. ACS Omega *1*, 680–688.

80. Havrila, M., Stadlbauer, P., Islam, B., Otyepka, M., and Šponer, J. (2017). Effect of Monovalent Ion Parameters on Molecular Dynamics Simulations of G-Quadruplexes. J. Chem. Theory Comput. *13*, 3911–3926.

81. Hayatshahi, H.S., Henriksen, N.M., and Cheatham, T.E. (2018). Consensus Conformations of Dinucleoside Monophosphates Described with Well-Converged Molecular Dynamics Simulations. J. Chem. Theory Comput. *14*, 1456–1470.

82. Mlýnský, V., and Bussi, G. (2018). Molecular Dynamics Simulations Reveal an Interplay between SHAPE Reagent Binding and RNA Flexibility. J. Phys. Chem. Lett. *9*, 313–318.

83. Borkar, A.N., Bardaro, M.F., Camilloni, C., Aprile, F.A., Varani, G., and Vendruscolo, M. (2016). Structure of a low-population binding intermediate in protein-RNA recognition. Proc. Natl. Acad. Sci. *113*, 7171–7176.

84. Zhou, H., Kimsey, I.J., Nikolova, E.N., Sathyamoorthy, B., Grazioli, G., McSally, J., Bai, T., Wunderlich, C.H., Kreutz, C., Andricioaei, I., *et al.* (2016). m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. Nat. Struct. Mol. Biol. *23*, 803–810.

85. Bottaro, S., Bussi, G., Kennedy, S.D., Turner, D.H., and Lindorff-Larsen, K. (2018). Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. Sci. Adv. *4*, eaar8521.

86. 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.

87. Dawson,W.K., Maciejczyk,M., Jankowska,E.J. and Bujnicki,J.M. (2016) Coarse-grained modeling of RNA 3D structure. Methods, 103, 138–156.

88. Kerpedjiev,P., Höner zu Siederdissen,C. and Hofacker,I.L. (2015) Predicting RNA 3D structure using a coarse-grain helix-centered model. RNA, 21, 1110–1121.

89. Ding,F., Lavender,C.A., Weeks,K.M. and Dokholyan,N. V (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. Nat. Methods, 9, 603-08.

90. Krokhotin,A., Houlihan,K. and Dokholyan,N. V (2015) iFoldRNA v2: folding RNA with constraints. Bioinformatics, 31, 2891–93.

91. Ding,F., Sharma,S., Chalasani,P., Demidov,V. V, Broude,N.E. and Dokholyan,N. V (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA, 14, 1164–73.

92. Mustoe,A.M., Al-Hashimi,H.M. and Charles L. Brooks,I. (2014) Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges. J. Phys. Chem. B, 118, 2615–27.

93. Cragnolini,T., Laurin,Y., Derreumaux,P. and Pasquali,S. (2015) Coarse-Grained HIRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. J. Chem. Theory Comput., 11, 3510–3522.

94. Uusitalo,J.J., Ingólfsson,H.I., Marrink,S.J. and Faustino,I. (2017) Martini Coarse-Grained Force Field: Extension to RNA. Biophys. J., 113, 246–256.

95. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA, 15, 189–99.

96. Weinreb,C., Riesselman,A.J., Ingraham,J.B., Gross,T., Sander,C., Marks,D.S., Adachi,A., Gendelman,H.E., Koenig,S., Folks,T., et al. (2016) 3D RNA and Functional Interactions from Evolutionary Couplings. Cell, 165, 963–75.

97. Šulc,P., Romano,F., Ouldridge,T.E., Doye,J.P.K. and Louis,A.A. (2014) A nucleotide-level coarse-grained model of RNA. 140, 235102.

98. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M. and Bujnicki,J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res., 44, e63.

99. Bernauer,J., Huang,X., Sim,A.Y.L. and Levitt,M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. RNA, 17, 1066–75.

100. Tan,R.K.Z., Petrov,A.S. and Harvey,S.C. (2006) YUP: A Molecular Simulation Program for Coarse-Grained and Multi-Scaled Models. J. Chem. Theory Comput., 2, 529–40.

101. Bell,D.R., Cheng,S.Y., Salazar,H. and Ren,P. (2017) Capturing RNA Folding Free Energy with Coarse-Grained Molecular Dynamics Simulations. Sci. Rep., 7, 45812.

102. He,Y., Liwo,A. and Scheraga,H.A. (2015) Optimization of a Nucleic Acids united-RESidue 2-Point model (NARES-2P) with a maximum-likelihood approach. J. Chem. Phys., 143, 243111.

103. Das,R., Karanicolas,J. and Baker,D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat. Methods, 7, 291–294.

104. Williams,B., Zhao,B., Tandon,A., Ding,F., Weeks,K.M., Zhang,Q. and Dokholyan,N. V. (2017) Structure modeling of RNA using sparse NMR constraints. Nucleic Acids Res., 45, 12638–12647.

105. Magnus,M., Boniecki,M.J., Dawson,W. and Bujnicki,J.M. (2016) SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. Nucleic Acids Res., 44, W315–W319.

106. 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.

107. Poblete,S., Bottaro,S. and Bussi,G. (2018) A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. Nucleic Acids Res., 46, 1674–1683.

108. Bhandari,Y.R., Fan,L., Fang,X., Zaki,G.F., Stahlberg,E.A., Jiang,W., Schwieters,C.D., Stagno,J.R. and Wang,Y.-X. (2017) Topological Structure Determination of RNA Using Small-Angle X-Ray Scattering. J. Mol. Biol., 429, 3635–3649.

109. Tan, Z., Sharma, G. and Mathews, D.H. (2017) Modeling RNA Secondary Structure with Sequence Comparison and Experimental Mapping Data. Biophys. J., 113, 330–338.

110. Low, J.T. and Weeks, K.M. (2010) SHAPE-directed RNA secondary structure prediction. Methods, 52, 150–58.

111. Lorenz, R., Luntzer, D., Hofacker, I.L., Stadler, P.F. and Wolfinger, M.T. (2015) SHAPE directed RNA folding. Bioinformatics, 32, btv523.

112. Hajdin, C.E., Bellaousov, S., Huggins, W., Leonard, C.W., Mathews, D.H. and Weeks, K.M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc. Natl. Acad. Sci. U. S. A., 110, 5498–503.

113. Smola, M.J. and Weeks, K.M. (2018) In-cell RNA structure probing with SHAPE-MaP. Nat. Protoc., 13, 1181–1195.

114. Vieweger, M. and Nesbitt, D.J. (2018) Synergistic SHAPE/Single-Molecule Deconvolution of RNA Conformation under Physiological Conditions. Biophys. J., 114, 1762–1775.

115. Cantara, W.A., Olson, E.D. and Musier-Forsyth, K. (2017) Analysis of RNA structure using small-angle X-ray scattering. Methods, 113, 46–55.

116. Zettl, T., Mathew, R.S., Shi, X., Doniach, S., Herschlag, D., Harbury, P.A.B. and Lipfert, J. (2018) Gold nanocrystal labels provide a sequence–to–3D structure map in SAXS reconstructions. Sci. Adv., 4, 1–10.

117. Feng,C., Chan,D., Joseph,J., Muuronen,M., Coldren,W.H., Dai,N., Corrêa,I.R., Furche,F., Hadad,C.M. and Spitale,R.C. (2018) Light-activated chemical probing of nucleobase solvent accessibility inside cells. Nat. Chem. Biol., 14, 276–283.

118. Wang,J., Mao,K., Zhao,Y., Zeng,C., Xiang,J., Zhang,Y. and Xiao,Y. (2017) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. Nucleic Acids Res., 45, 6299–6309.

119. Rother, M., Rother, K., Puton, T. and Bujnicki, J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acid Research, 39, 4007-22.

120. Flores, S.C., Wan, Y., Rusell, R. and Altman, R.B. (2010) Predicting RNA structure by multiple template homology modeling. Pacific Symposium Biocomputing, 216-27.

121. Miao, Z. and Westhof, E. RNA (2017) Structure: Advances and Assessment of 3D Structure Prediction. Annual Reviews of Biophysics, 46, 483-503.

122. Sijenyi, F., Saro, P., Ouyang, Z., Damm-Ganamet, K., Wood, M., Jiang, J. and SantaLucia Jr, J. (2012) The RNA folding problems: different levels of sRNA structure prediction. In Leontis, N. and Westhof, E. (eds) RNA 3D Structure Analysis and Prediction. Nucleic Acids and Molecular Biology, ed (Springer, Berlin, Heidelberg), 27, 91-117.

123. Magnus, M., Matelska, D., Łach, G., Chojnowski, G., Boniecki, M.J., Purta, E., Dawson, W., Dunin-Horkawicz, S. and Bujnicki, J.M. (2014) Computational modeling of RNA 3D structures, with the aid of experimental restraints. RNA Biology, 11, 522-36.

124. Mathews, D.H., Moss, W.N. and Turner. D.H. (2010) Folding and finding RNA secondary structure. Cold Spring Harbor Perspectives in Biology, 2, a003665.

125. Frellsen, J., Moltke, I., Thiim, M., Mardia, K.V., Ferkinghoff-Borg, J. and Hamelryck, T. A (2009) Probabilistic Model of RNA Conformational Space. PLOS Computational Biology, 5, e1000406.

126. Wang, Z. and Xu, J. (2011) A conditional random fields method for RNA sequence-structure relationship modeling and conformation sampling .Bioinformatics, 27, i102-i110.

127. Capriotti, E., Norambuena, T., Marti-Renom, M. A. & Melo, F. (2011) All-atom knowledge-based potential for RNA structure prediction and assessment. Bioinformatics 27, 1086–1093.

128. Masso, M. (2017) All-Atom Four-Body Knowledge-Based Statistical Potentials to Distinguish Native Protein Structures from Nonnative Folds. Biomed Res. Int. 2017, 1–17.

129. Yang, Y., Gu, Q., Zhang, B.-G., Shi, Y.-Z. & Shao, Z.-G. (2018) A novel knowledge-based potential for RNA 3D structure evaluation. Chinese Phys. B, 27, 38701.

130. Massire, C. and Westhof, E. (1998) MANIP: An interactive tool for modelling RNA. Journal of Molecular Graphics and Modelling, 16, 197-205.

131. Zwieb, C. and Müller, F. (1997) Three-dimensional comparative modeling of RNA. Nucleic Acids Symposium Series, 69-71.

132. Martinez, H.M., Maizel Jr, J.V. and Shapiro, B.A. (2008) RNA2D3D: A program for generating, viewing and comparing 3-Dimensional Models of RNA. Journal of Biomolecular Structure and Dynamics, 25, 669-683.

133. Jossinet, F. (2015) Assemble2: an interactive graphical environment dedicated to the study and construction of RNA architectures. Virtual and Augmented Reality for Molecular Science, 37-38.

134. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K.J., Lukasiak, P., Bartol, N., Blazewicz, J. and Adamiak, R.W. (2012) Automated 3D structure composition for large RNAs. Nucleic Acids Research, 40, e112.

135. Bida, J.P. and Maher 3rd, L.J. (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. RNA, 18, 385-393.

136. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature, 452, 51-55.

137. Zhao, Y., Huang, Y., Gong, Z., Wang, Y., Man, J. and Xiao, Y. (2012) Automated and fast building of three-dimensional RNA structures. Scientific Reports, 2, 734.

138. Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E.K., Blazewicz, J., and Adamiak, R.W. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics, 11, 231.

139. Reinharz, V., Major, F. and Waldispühl, J. (2012) Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. Bioinformatics, 28, i207-14.

140. Cao, S. and Chen, S.J. (2011) Physics-based de novo prediction of RNA 3D structures. Journal of Physical Chemistry, 115, 4216-4226.

141. Xu, X. & Chen, S.-J. (2018) Hierarchical Assembly of RNA Three-Dimensional Structures Based on Loop Templates. J. Phys. Chem. B, 122, 5327–5335.

142. Jain, S. and Schlick, T. (2017) F-RAG: Generating atomic coordinates from RNA graphs by fragment assembly. Journal of Molecular Biology, 429, 3587-3605.

143. Boudard, M., Barth, D., Bernauer, J., Denise, A. and Cohen, J. (2017) GARN2: coarse-grained prediction of 3D structure of large RNA molecules by regret minimization. Bioinformatics, 33, 2479-2486.

144. Cruz, J.A. and Westhof, E. (2011) Sequence-based identification of 3D structural modules in RNA with RMDetect. Nature Methods, 8, 513-21.

145. Theis, C., HönerZuSiederdissen, C., Hofacker I.L., and Gorodkin, J. (2013) Automated identification of RNA 3D modules with discriminative power in RNA structural alignments. Nucleic Acids Research, 41, 9999-10009.

146. Zirbel, C.L., Roll, J., Sweeney, B.A., Petrov, A.I., Pirrung, M, and Leontis, N.B. (2015) Identifying novel sequence variants of RNA 3D motifs. Nucleic Acids Research, 43, 7504-7520.

147. Petrov, A.I., Zirbel, C.L., Leontis, N.B. (2013) Automated classification of RNA 3D motif and the RNA 3D Motif Atlas. RNA, 19, 1327-1340.

148. De Leonardis, E., Lutz, B., Ratz, S., Cocco, S., Monasson, R., Schug, A. and Weigt, M. (2015) Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic Acids Research, 43, 10444-10455.

149. Miao, Z., Adamiak, R.W. Antczak, M., Batery, R.T., Becka, A.J., Biesiada, M., Boniecki, M.J., Bujnicki, J.M., Chen, S.J., Cheng, C.Y., Chou, F.C., Ferré-D'Amaré, A.R., Das, R., Dawson, W.K., Ding, F., Dokholyan, N.V., Dunin-Horkawicz, S., Geniesse, C., Kappel, K., Kladwang, W., Krokhotin, A., Łach, G.E., Major, F., Mann, T.H., Magnus, M., Pachulska-Wieczorek, K., Patel, D.J., Piccirilli, J.A., Popenda, M., Purzycka, K.J., Ren, A., Rice, G.M., Santalucia, J Jr., Sarzynska, J., Szachniuk, M., Tandon, A., Trausch, J.J., Tian, S., Wang, J., Weeks, K.M., Williams, B 2nd., Xiao, Y., Xu, X., Zhang, D., Zok, T. and Westhof, E. (2017) RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA, 23, 655-6.

150. Liu, F., Somarowthu, S. and Pyle, A.M. (2017) Visualizing the secondary and tertiary architectural domains of lncRNARepA. Nature Chemical Biology, 13, 282-289.

TABLES

Table 1. Summary of available coarse-grained methods and stand-alone bioinformatics tools to model RNA 3D structure.

Particle-based coarse grain models					
CG force field name	Sampling method	Potential Energy Function	Compatible molecules / Experimental complementation	Sequence length range	Number of beads
MARTINI	Molecular Dynamics (MD)	Physicle principles	Water, ions, lipids, carbohydrates, polymers, proteins, DNA / Knowledge of secondary and tertiary structure is required	Tested on ~5000 nt	7 (Pur), 6 (Pyr) per nt
HIRE-RNA v3	Molecular Dynamics	Hybrid (statistical+physical)	DNA, proteins (combines with OPEP FF) / No	76 nt tested	7 (Pur), 6 (Pyr) per nt
3dRNA	Simulated Annealing Monte Carlo (SAMC) algorithm	Physicle principles	--- / Could be combined with secondary and tertiary contacts from co-evolution data	Up to 377 nt	6 per nt
simRNA	Variety of samping engines (MD and MC)	Statistical, knowledge based	--- / Good results with secondary structure fixed with restraints	Up to 190 nt	5 per nt
RNAkb	Molecular Dynamics	Statistical, knowledge based	--- / ---	76 nt tested	5 per nt
RACER	Molecular Dynamics	Hybrid (statistical+physical)	Mg^{2+} / Could be complemented with SAXS and NMR data		5 per nt
oxRNA	Monte Carlo (MC)	Statistical, knowledge based	DNA / ---	10^3 nt	5 per nt
iFoldRNA v2	Discrite Molecular Dynamics (DMD)	Physicle principles	--- / Could be complemented with base-	Up to 200 nt	3 per nt

			probing		
iFoldNMR	Discrite Molecular Dynamics	Physicle principles	--- / Sparce NMR data	Up to 56 nt	3 per nt
TOPRNA	Molecular Dynamics	Physicle principles	--- / Knowledge of secondary structure is required	12 nt tested	3 per nt
NARES-2P	Molecular Dynamics	Hybrid (statistical+physical)	DNA, proteins with UNRES / ---	59 nt tested	2 per nt + dipole moment
SPQR	Monte Carlo	Statistical, knowledge based	--- / ---	12 nt tested	2 per nt, one anisotropic
NAST	Molecular Dynamics	Statistical, knowledge based	--- / Could be combined with secondary and tertiary contacts from co-evolution data and SAXX	Up to 160 nt	1 per nt
FARNA	Monte Carlo	Statistical, knowledge based	--- / Could be complemented with secondary structure data	Up to 23 nt	1 per nt
YUP-rrRNAv1	Monte Carlo	Hybrid (statistical+physical)	DNA / ---	76 nt tested	1 per nt
RS3D	Monte Carlo	Statistical, knowledge based	--- / SAXS data	Up to 387 nt	1 per nt
Bioinformatics tools based on structure conservation and comparative analyses					
Software name	Sampling method	Scoring Function	Source of knowledge	Sequence length range	Written in
ModeRNA	Template global structure and sequence mutation according to alignment. For insertion/deletion regions fragment assembly is applied	Based on input sequence alignment	Based on homologous 3D template and fragment database	More than >70 nt	Python
Macro	Torsions and interatomic distances are taken	Force field based	Based on homologous 3D	More than >200	C++

Molecule Builder (MMB), formerly RNABuilder	from template and further refinement is done with a multi-resolution modeling approach	on user-defined constraints and secondary structure	template	nt	
Galvanek & Coworkers (unnamed)	Transference of template global structure and Rosetta de novo prediction (see below) for non-conserved/in-del regions	---	Based on homologous 3D template	From 50 to >500 nt	---
RNA123	Template global structure and sequence mutation according to alignment and BUILDER algorithm based on fragment assembly	RNA123 force field	Based on homologous 3D template and Motifs database	Up to ~ 1500 nt	C++
Bioinformatics tools based on hierarchical folding, interactive modeling					
Software name	Sampling method	Scoring Function	Source of knowledge	Sequence length range	Written in
ERNA-3D	User driven	---	User experience	---	---
MANIP	Fragment Assembly and manual manipulation	None declared (user decides)	3D Modules database and user experience	---	C
RNA2D3D	First model automatic, then user driven modifications	Molecular Mechanics Minimizations	3D representations of base pairs & single-stranded nucleotides and user experience	No limits (user defined)	C
S2S/ Assemble/ Assemble2	Fragment Assembly and manual manipulation	Convergence to predicted secondary structure	3D Modules database and user experience	No limits (user defined)	Java and Python
Bioinformatics tools based on hierarchical folding, automated methods					
Software name	Sampling method	Scoring Function	Source of knowledge	Sequence length range	Written in
MC-Sym	Fragment Assembly with Las Vegas algorithm	Knowledge-based energy function	3D NCM (Nucleotide Cyclic Motifs) database and energy function	Up to ~ 50 nt	---

RNAComposer	Fragment Assembly from RNA FRABASE structural elements	Satisfaction of input sequence and 2D structure	Provided secondary structure and RNA FRABASE database	Tested between 31 and 161 nt; server limited to 500 nt	---
RSIM	Monte Carlo Fragment Assembly with Closed Moves	Satisfaction of input 2D structure	Provided secondary structure and 3D fragment database	Tested between 17 and 46 nt	C++
3dRNA	Fragment assembly of SSE (Smallest Secondary Elements)	3dRNAscore	Provided secondary structure and 3D SSE database	Tested between 12 and 101 nt	---
Ernwin	Markov Chain Monte Carlo simulation	Knowledge-based scoring function	Provided secondary structure and 3D fragment database	Tested between 66 and 417 nt	Python
Vfold	Template match or fragment assembly of multiple motif templates	Free energy estimation	3D structural template database	Server restricted to 200 nt	---
VfoldLA	Fragment assembly of multiple loop templates	Sequence-based	3D structural template database	Tested up to 84	---
GARN/ GARN2	Game algorithm	Knowledge-based scoring function	Provided secondary structure and set of RNA molecules with available 3D and 2D structures	Tested between 49 and 172 nt	Java

Bioinformatics tools based on local structural variability

Software name	Sampling method	Scoring Function	Source of knowledge	Sequence length range	Written in
Rosetta (FARNA/ FARFAR)	Monte Carlo Fragment Assembly	Full atom high-resolution refinement guided by a force field for atom interactions	Trinucleotide 3D database and scoring function	Up to ~ 40 nt	Python
BARNACLE	Dynamic Bayesian Network combined with directional statistics for continuous	Satisfaction of input 2D structure	Probabilistic model based on available non-redundant	Tested between 12 and 46 nt	Python

	conformational sampling		set of 3D RNA structures		
TreeFolder	Tree-guided conformation sampling algorithm, based on conditional random fields (CRFs)	Satisfaction of input 2D structure	Conditional Random Fields (CRF) model trained on a representative set of 3D RNA structures	Tested between 12 and 46 nt	---

Table 2. Integrated tools and methods for RNA 3D structure prediction starting sequence which are implemented in pipelines.

Software name	Sampling	Scoring Function	Se len
Improvements using evolutionary information			
Evolutionary Couplings (EC)	Guided NAST sampling	K-means clustering with lowest energy-per-contact, using EC	Te be an
Direct Coupling Analysis	Guided Rosetta sampling	Rosetta function with additional terms based on predicted tertiary contacts	Te be an
3dRNA	Fragment assembly of SSE (Smallest Secondary Elements) guided by coevolutionary signals	3dRNAscore	Te be an
Hierarchical folding pipelines			
MC-Fold | MC-Sym	Fragment Assembly with Las Vegas algorithm	Knowledge-based energy function	Up
RNA-MoIP (RNAsubopt | RNA-MoIP | MC-Sym)	Guided MC-Sym sampling	Objective function based on minimum entropy	Te be an
F-RAG | RAGTOP | RAG-3D	Monte Carlo Simulated Annealing	Knowledge-based scoring function	Te be an

FIGURES

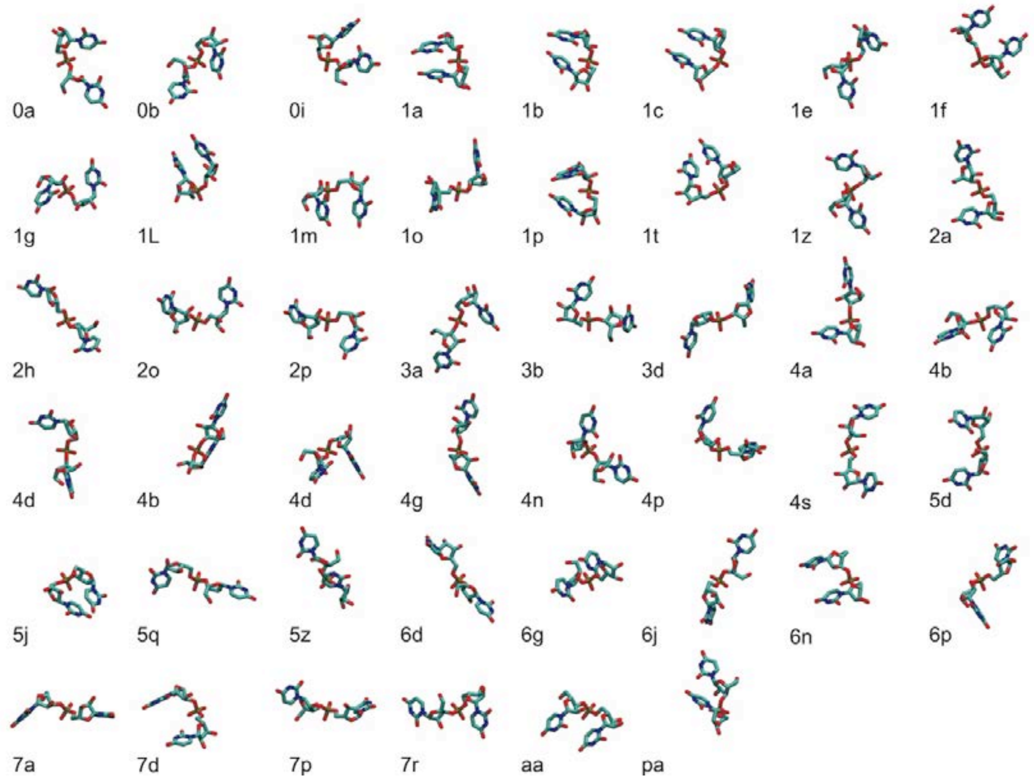
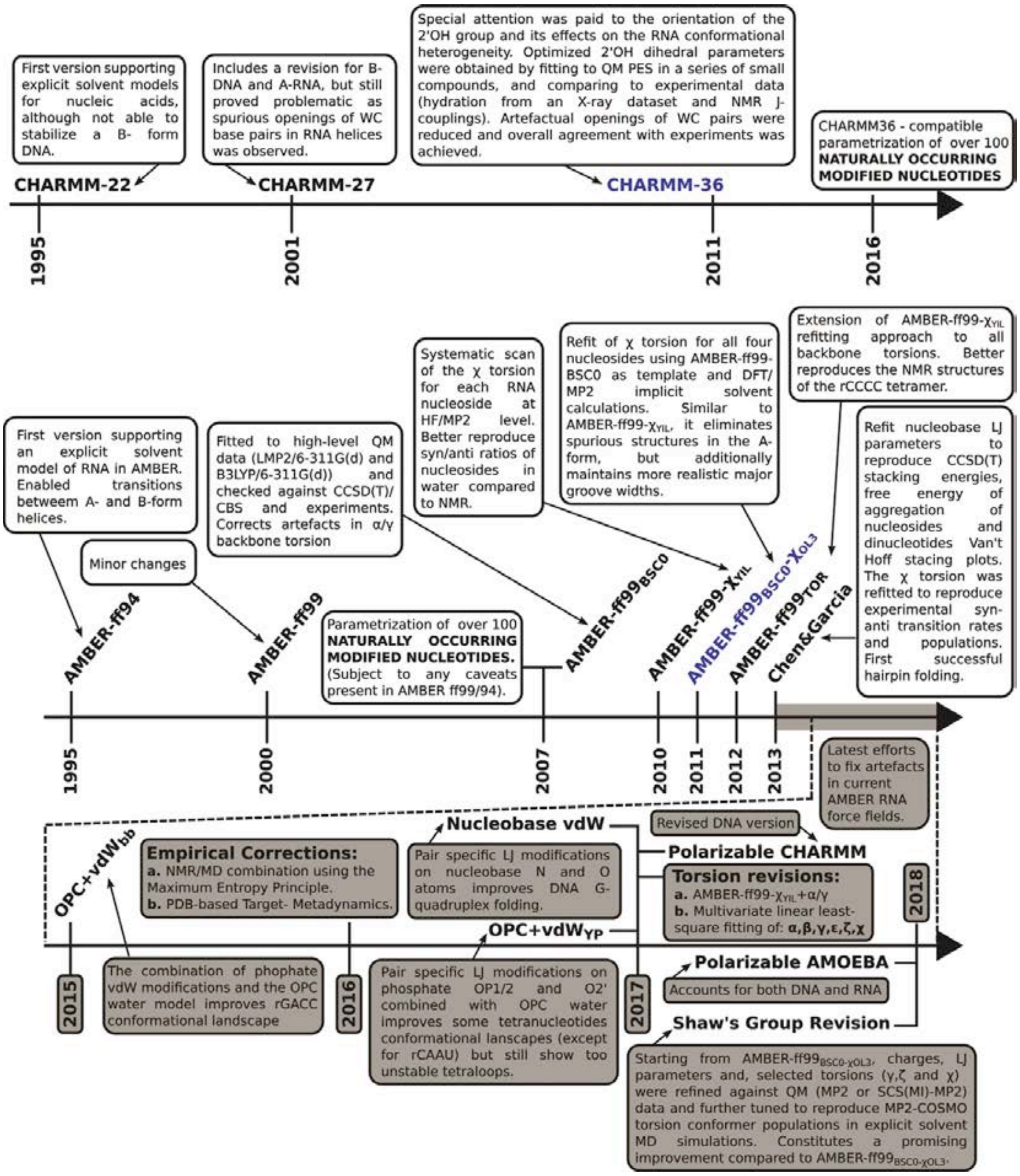


Figure 1. Representation of the 46 conformers present in the UpU46 benchmark set. The set consists of the 46 uracil dinucleotides (UpU), representing all known 46 RNA backbone conformational families described by the RNA Ontology Consortium. The conformers are given two-character names that reflect the seven-angle (δεζαβγδ) combinations empirically found favorable for the sugar-to-sugar ''suite'' unit within which the angle correlations are strongest. Note that all conformers were aligned to a common reference coordinate.



First version supporting explicit solvent models for nucleic acids, although not able to stabilize a B- form DNA.

Includes a revision for B-DNA and A-RNA, but still proved problematic as spurious openings of WC base pairs in RNA helices was observed.

Special attention was paid to the orientation of the 2'OH group and its effects on the RNA conformational heterogeneity. Optimized 2'OH dihedral parameters were obtained by fitting to QM PES in a series of small compounds, and comparing to experimental data (hydration from an X-ray dataset and NMR J-couplings). Artefactual openings of WC pairs were reduced and overall agreement with experiments was achieved.

CHARMM36 - compatible parametrization of over 100 **NATURALLY OCCURRING MODIFIED NUCLEOTIDES**

First version supporting an explicit solvent model of RNA in AMBER. Enabled transitions betweem A- and B-form helices.

Minor changes

Fitted to high-level QM data (LMP2/6-311G(d) and B3LYP/6-311G(d)) and checked against CCSD(T)/CBS and experiments. Corrects artefacts in α/γ backbone torsion

Systematic scan of the χ torsion for each RNA nucleoside at HF/MP2 level. Better reproduce syn/anti ratios of nucleosides in water compared to NMR.

Refit of χ torsion for all four nucleosides using AMBER-ff99-BSC0 as template and DFT/MP2 implicit solvent calculations. Similar to AMBER-ff99-χ_{YIL}, it eliminates spurious structures in the A-form, but additionally maintains more realistic major groove widths.

Extension of AMBER-ff99-χ_{YIL} refitting approach to all backbone torsions. Better reproduces the NMR structures of the rCCCC tetramer.

Refit nucleobase LJ parameters to reproduce CCSD(T) stacking energies, free energy of aggregation of nucleosides and dinucleotides Van't Hoff stacing plots. The χ torsion was refitted to reproduce experimental syn-anti transition rates and populations. First successful hairpin folding.

Parametrization of over 100 **NATURALLY OCCURRING MODIFIED NUCLEOTIDES.** (Subject to any caveats present in AMBER ff99/94).

Latest efforts to fix artefacts in current AMBER RNA force fields.

Empirical Corrections:
a. NMR/MD combination using the Maximum Entropy Principle.
b. PDB-based Target- Metadynamics.

Pair specific LJ modifications on nucleobase N and O atoms improves DNA G-quadruplex folding.

Revised DNA version

Torsion revisions:
a. AMBER-ff99-χ_{YIL}+α/γ
b. Multivariate linear least-square fitting of: **α,β,γ,ε,ζ,χ**

The combination of phophate vdW modifications and the OPC water model improves rGACC conformational landscape

Pair specific LJ modifications on phosphate OP1/2 and O2' combined with OPC water improves some tetranucleotides conformational lanscapes (except for rCAAU) but still show too unstable tetraloops.

Accounts for both DNA and RNA

Starting from AMBER-ff99$_{BSC0-\chi OL3}$, charges, LJ parameters and, selected torsions (γ,ζ and χ) were refined against QM (MP2 or SCS(MI)-MP2) data and further tuned to reproduce MP2-COSMO torsion conformer populations in explicit solvent MD simulations. Constitutes a promising improvement compared to AMBER-ff99$_{BSC0-\chi OL3}$.

Figure 2. Time line of the evolution of the CHARMM (top) and AMBER (bottom) force fields for RNA. Corrections to each original version are indicated and their main characteristics are commented. The region highlighted in gray corresponds to the latest period where several efforts to eliminate artifacts in the current available version have been done. Such period is particularly addressed in the present work. The current standard versions are indicated in blue.

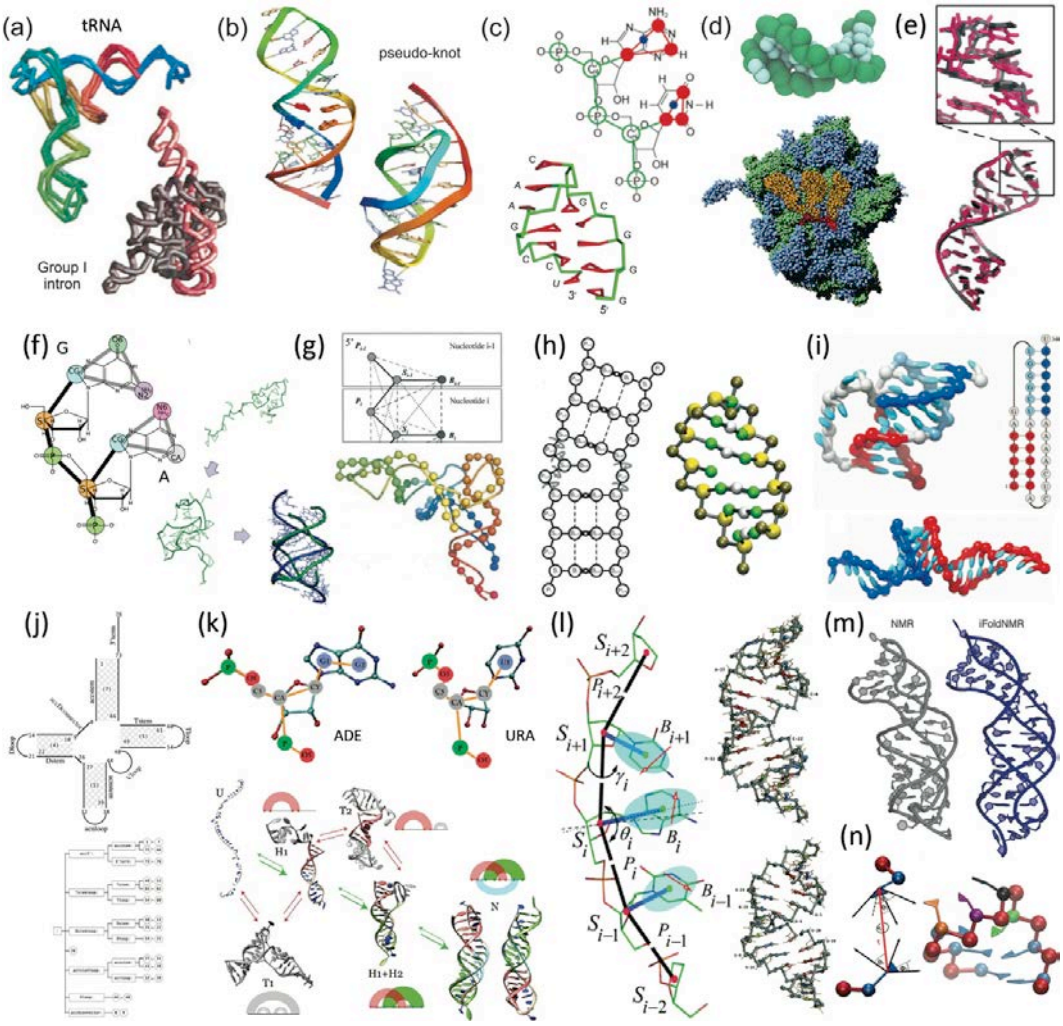


Figure 3. Mapping strategies used for coarse graining RNA in different models, as illustrated by the corresponding authors. (a) The model of Jonikas*et al*. with 1 bead per residue, called NAST. (b) The FARNA model developed by Das & Baker. 1 bead per residue is used to reduce the complexity. (c) simRNA from Boniecki*et al*., with 5 beads per residue. The mapping scheme and a hairpin are shown. (d) The model developed by Faustino *et al*. for the MARTINI coarse-grain force field. Representation of a hairpin, and a complete ribosome containing mRNA and tRNA. (e) The RNAkb model of Levitt *et al*. use 5 beads per residue. (f) The 5 beads per residue model of Ren *et al*. (RACER) was used to fold short RNA oligomers. (g) iFoldRNAv2 from Dokholyan *et al*. uses 3 beads per residue to fold a tRNA molecule. (h) TOPRNA, developed by Brooks *et al*., uses 3 pseudoatoms to represent the phosphate, sugar, and base. (i) The oxRNA model from Ouldridge*et al*. uses a single rigid body with 5 interaction sites to represent a nucleotide. Representation of a pseudoknot. (j) The YUP-rrRNAv1 developed of Harvey *et al*., with 1 bead per residue. Representation of the tRNA^{PHE} hierarchy. (k) The HIRe-RNA model of Derremaux *et al*. with 6/7 beads per residue. Folding pathway of a triple helix. (l) The NARES-2P for RNA developed by Scheraga *et al*. with 2 beads per residue. (m) iFoldNMR from Dokholyan *et al*., with 3 beads per residue coupled to

sparse NMR restraints data. (n) The SPlit-and-conQueR (SPQR) model of bussi *et al.*, with 2 beads per residue (one anisotropic). Representation of a CG tetraloop.

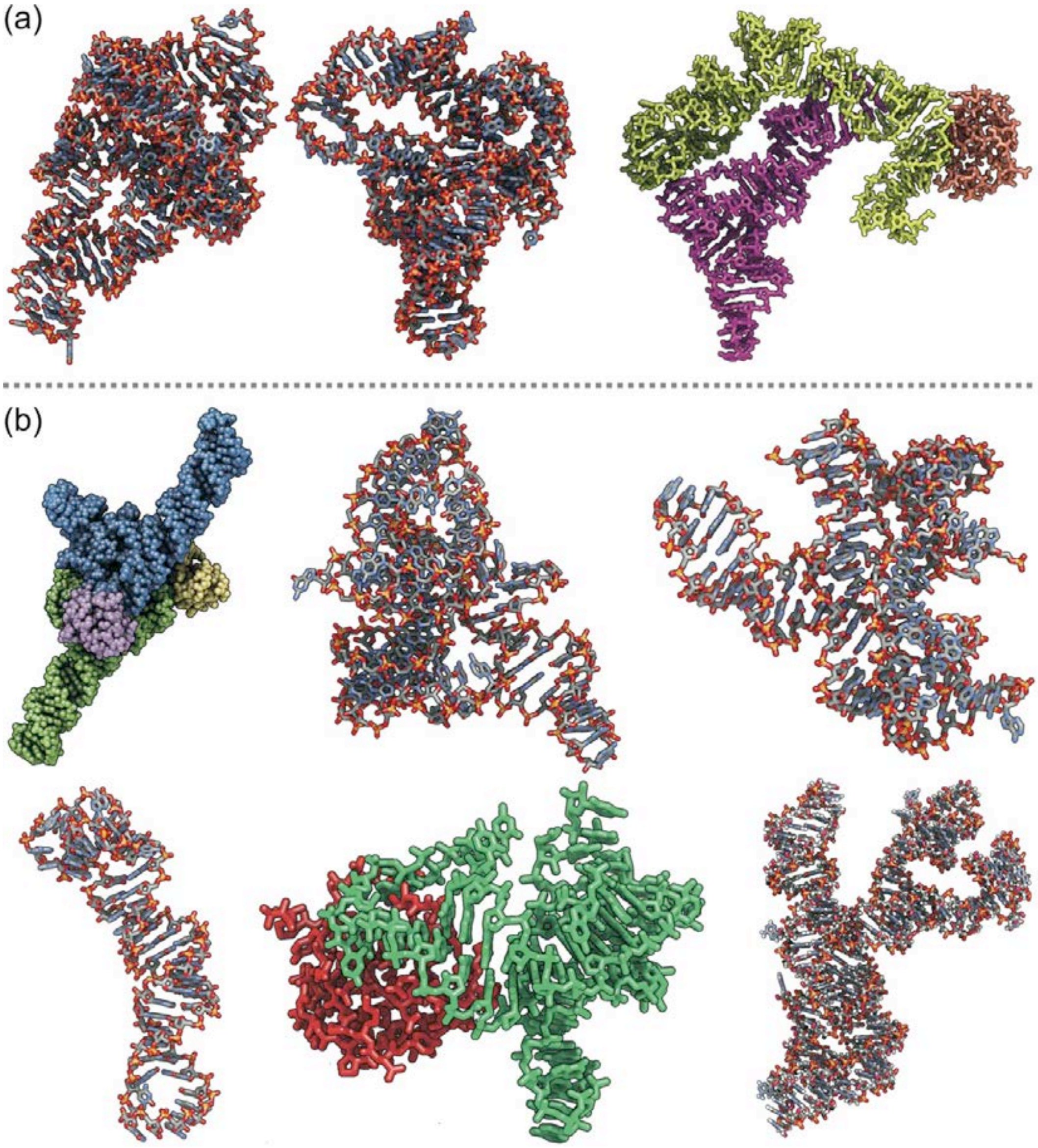


Figure 4. RNA-Puzzles determined in the collective and blind experiments in 3D RNA structure prediction. (a) Structures assessed in RNA-Puzzles round II (from left to right): Lariat capping ribozyme (PDB id 4P95); Adenosylcobalaminriboswitch (PDB id 4GXY); and, the T-box-tRNA complex structure (PDB id 4LCK). (b) The six 3D RNA structures predicted in the RNA-Puzzles round III (From left to right, and from top to bottom): SAM-I riboswitchaptamer (PDB id 3V7E); SAM-I/IV riboswitch (PDB id 4L81); c-di-AMP bound to ydaOriboswitch (PDB id 4QLM); The ZTP riboswitch (PDB id 4XW7); Apo form of L-glutamine riboswitch (PDB id 5DDO); The Varkud satellite (VS) ribozyme (PDB id 4R4P).

Resumen en español

1 Introducción

Estructura del ADN

En las células eucariotas, el ADN es un polímero muy largo dividido en unas pocas unidades independientes denominadas cromosomas, los cuales pueden llegar a medir dm (si se extienden) en organismos como el humano, pero están muy compactados dentro del núcleo celular. La fibra de ADN consiste en dos cadenas complementarias de biopolímeros formadas a partir de unidades repetitivas llamadas nucleótidos, enrolladas entre sí para formar una doble hélice. Cada nucleótido se compone de una de cuatro bases nitrogenadas lipófilas (citosina [C], guanina [G], adenina [A] o timina [T]), un azúcar desoxirribosa y un grupo fosfato completamente ionizado al pH fisiológico. Los nucleótidos se mantienen unidos en una cadena mediante enlaces fosfodiéster, lo que da como resultado una cadena principal de azúcar-fosfato alternante. Los grupos fosfato están unidos al carbono 5' de un nucleótido y al carbono 3' del otro, de modo que la unidad repetitiva completa en un ácido nucleico es un 5',3'-nucleótido. Las bases nitrogenadas son (principalmente) anillos aromáticos de estructura plana que se conectan al anillo de la (desoxi)ribosa por un enlace glicosídico entre el nitrógeno base endocíclico y el átomo C1 del azúcar. Las bases en dos cadenas opuestas de polinucleótidos forman enlaces de hidrógeno para formar ADN bicatenario, de acuerdo con las reglas de apareamiento de bases (A con T y C con G), y se apilan una encima de otra para estabilizar la doble hélice (ver la Figura 1).

La refinada forma del ADN permite la descripción de sus elementos estructurales constitutivos en términos de un conjunto reducido de coordenadas internas helicoidales. Una serie de parámetros rotacionales y traslacionales han sido delineados para describir las relaciones geométricas entre bases y pares de bases (definidos en la reunión EMBO en Cambridge en 1988, también llamado "Cambridge Accord", y estandarizados en el Taller Tsukuba sobre Estructura e Interacciones de Ácidos Nucleicos [1] de modo que se utilice como marco de referencia único para calcular los parámetros de morfología base y generar valores uniformes en todos los estudios). Definido con respecto a un sistema de coordenadas unido a cada par de bases o a pasos de pares de bases (*bps*), un conjunto de 18 parámetros helicoidales caracterizaría por completo a una unidad de dinucleótidos (ver Figura 2).

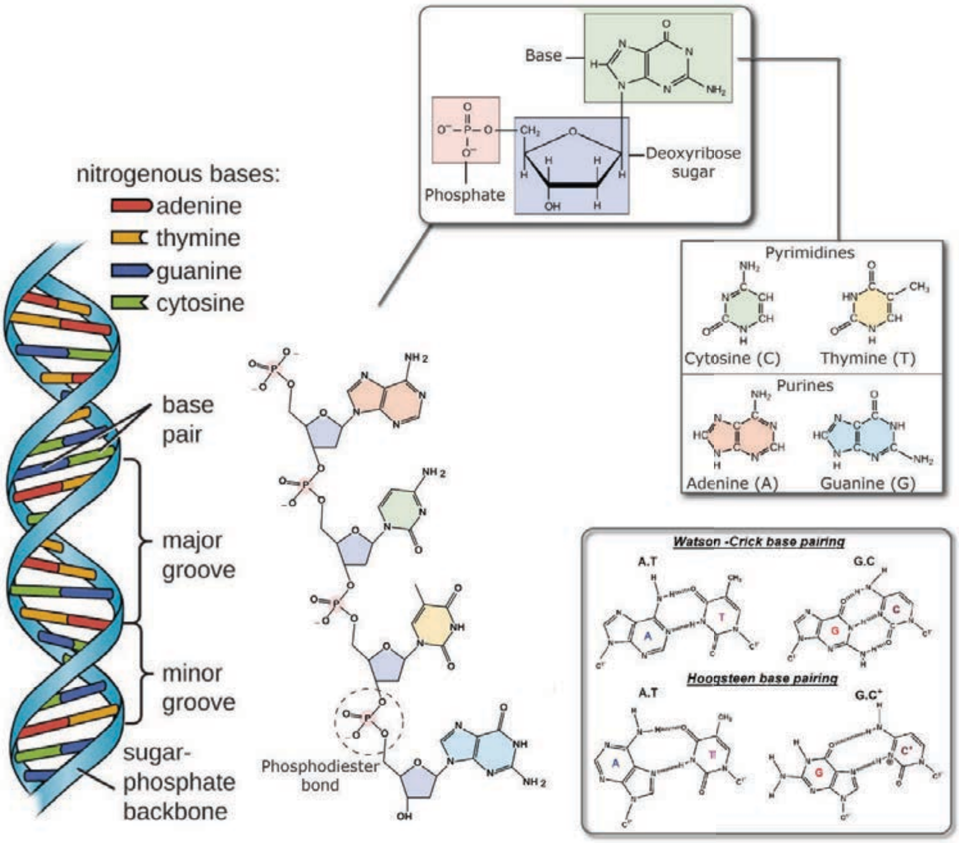


Figure 1 Estructura de la doble hélice de ADN, composición de bases y pares de bases

La cadena principal (*backbone*) del ADN impone un espacio conformacional que se puede entender mejor al considerar sus grados de libertad de torsión. Así, la geometría de la estructura está controlada por torsiones alrededor de seis cadenas principales, cinco enlaces de azúcar y uno glicosídico (Figura 3).

En general, la estructura local del ADN es el resultado de la interacción entre los parámetros helicoidales óptimos del par de bases, la conformación de los azúcares y los diedros de la cadena principal predilectos. Cuando se describe la hélice del ADN como un todo, una característica dominante es que, a lo largo de la secuencia de nucleótidos, los grupos de azúcar se unen al mismo lado de un par de bases y definen dos tipos de hendiduras: un surco mayor, delineado por el N7 de la purina y el C6 de la pirimidina, y un surco menor con el N3 de la purina y el O2 de la pirimidina. La nomenclatura se hace en referencia a la forma más común, ADN-B, donde el volumen vacío formado por el surco mayor es más grande que el del menor, pero por consistencia se mantiene para otras formas del ADN.

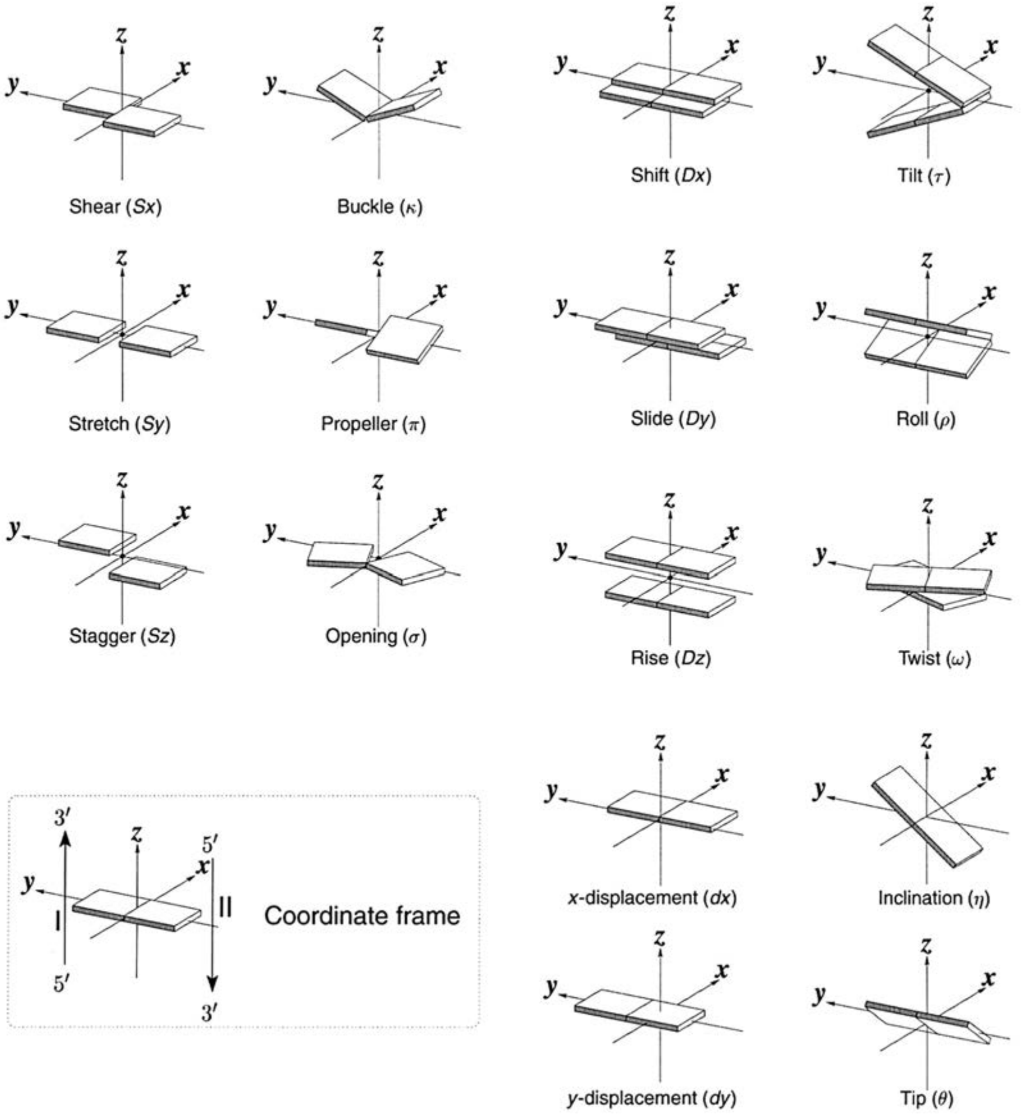


Figure 2 Modelo de base rígida y par de bases: definición de parámetros helicoidales para el par de bases y el paso del par de bases [2].

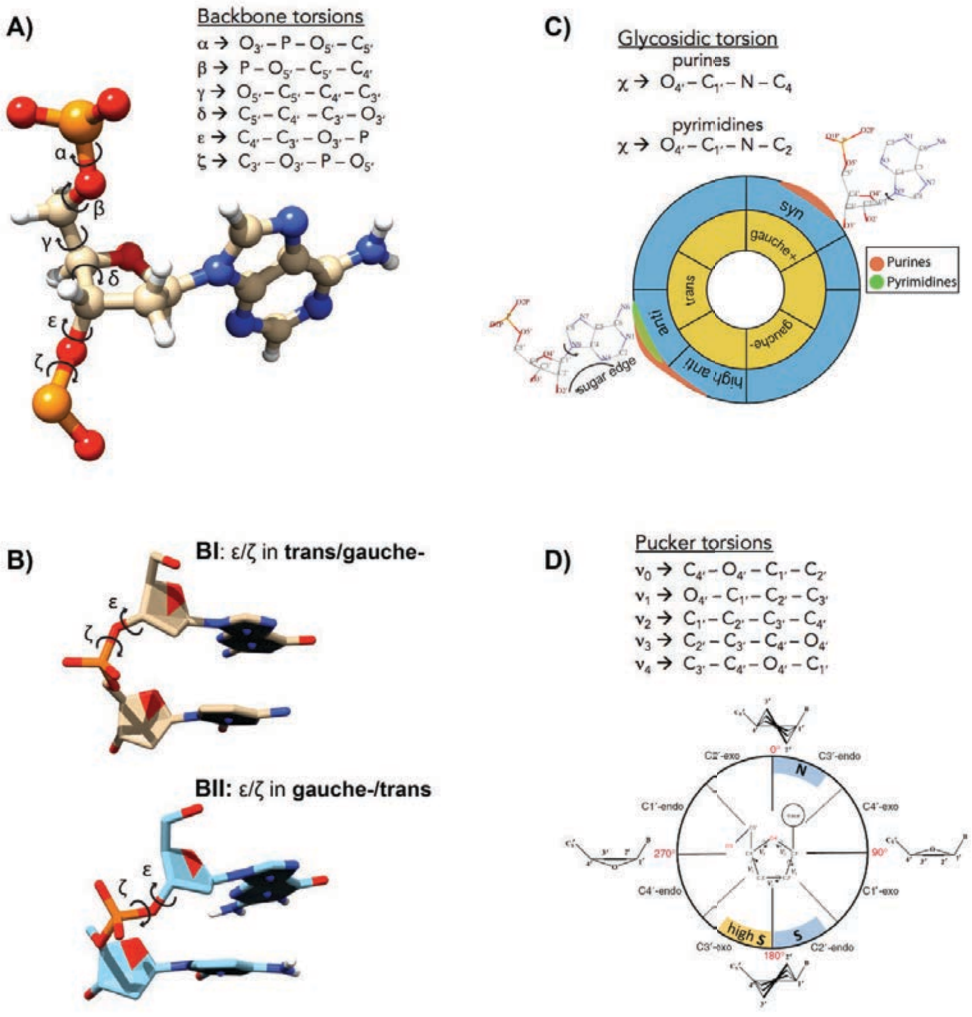

A)
Backbone torsions
α → $O_{3'} - P - O_{5'} - C_{5'}$
β → $P - O_{5'} - C_{5'} - C_{4'}$
γ → $O_{5'} - C_{5'} - C_{4'} - C_{3'}$
δ → $C_{5'} - C_{4'} - C_{3'} - O_{3'}$
ε → $C_{4'} - C_{3'} - O_{3'} - P$
ζ → $C_{3'} - O_{3'} - P - O_{5'}$
α
β
γ
δ
ε
ζ
C)
Glycosidic torsion
purines
χ → $O_{4'} - C_{1'} - N - C_4$
pyrimidines
χ → $O_{4'} - C_{1'} - N - C_2$
syn
gauche+
trans
anti
gauche-
high anti
sugar edge
Purines
Pyrimidines
B)
BI: ε/ζ in trans/gauche-
ε
ζ
BII: ε/ζ in gauche-/trans
ζ
ε
D)
Pucker torsions
ν_0 → $C_{4'} - O_{4'} - C_{1'} - C_{2'}$
ν_1 → $O_{4'} - C_{1'} - C_{2'} - C_{3'}$
ν_2 → $C_{1'} - C_{2'} - C_{3'} - C_{4'}$
ν_3 → $C_{2'} - C_{3'} - C_{4'} - O_{4'}$
ν_4 → $C_{3'} - C_{4'} - O_{4'} - C_{1'}$
C2'-exo
0°
C3'-endo
N
C1'-endo
C4'-exo
270°
90°
C4'-endo
C1'-exo
high S
S
C3'-exo
180°
C2'-endo

Figure 3 Definición de torsiones de la cadena principal del ADN. A) torsiones de la cadena principal; B) transiciones BI/BII en la cadena principal; C) Torsión glicosídica con rangos para pirimidinas (verde) y purinas (naranja); D) Tipos de fruncimiento.

Hay tres familias principales de hélices de ADN: ADN-B, que es el más común y específico para las secuencias mixtas (aunque la conformación exacta varía entre las diferentes combinaciones de secuencias); ADN-A, que puede formarse en ciertas secuencias de purinas (por ejemplo, GAGGGA) en condiciones no fisiológicas o en híbridos de ADN-ARN; y ADN-Z, que se encuentra favorecida en secuencias alternantes de pirimidina-purina (por ejemplo, CGCGCG) y está presente en una cantidad muy pequeña en el núcleo de la célula. Las familias de ADN A y B son hélices dextrógiras, mientras que la familia de ADN-Z tiene una orientación levógira de la hélice (véase la Figura 4 y la Tabla 1).

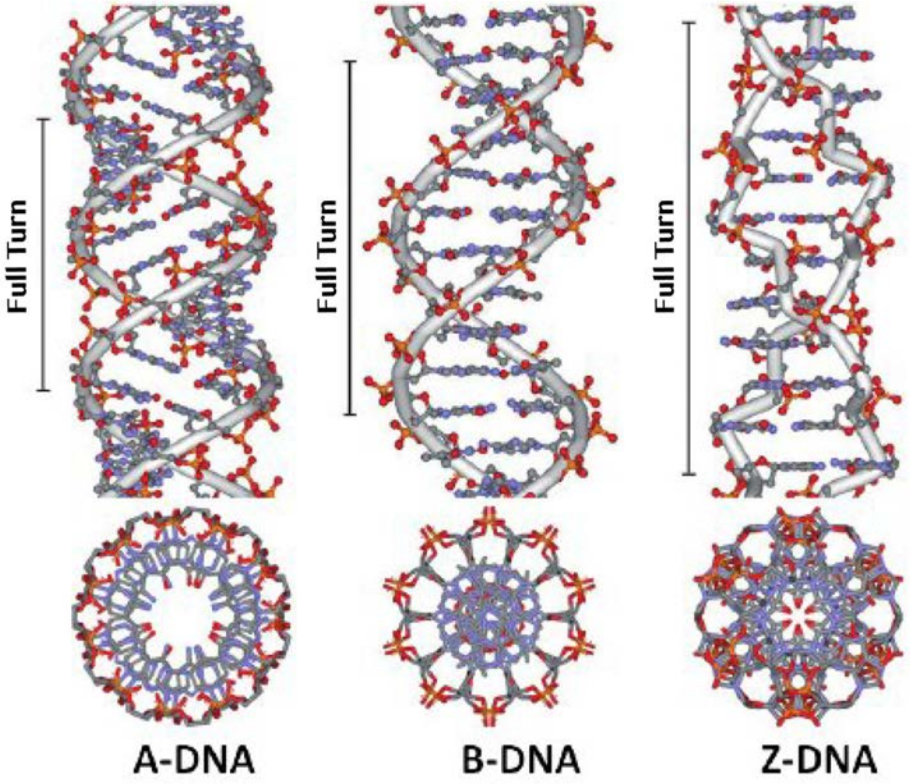


Figure 4 Las tres formas principales de ADN doble hélice.

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 Características geométricas de las 3 principales familias de hélice de ADN (Neidle 2008).

Estructura del ARN

Similar al ADN, el ARN es un oligopolímero con una cadena principal de fosfatos. Sus unidades básicas (llamadas ribonucleótidos) también están compuestas por un azúcar y una base nitrogenada, pero difieren de los nucleótidos del ADN en dos aspectos químicos clave: (i) los grupos de azúcares ribosa del ARN tienen un grupo hidroxilo unido en la posición 2' y (ii) las bases de timina del ADN son reemplazadas por uracilos (ver Figura 5) que emplean el mismo modo de apareamiento de bases, pero les falta el grupo metilo en la posición 5'.

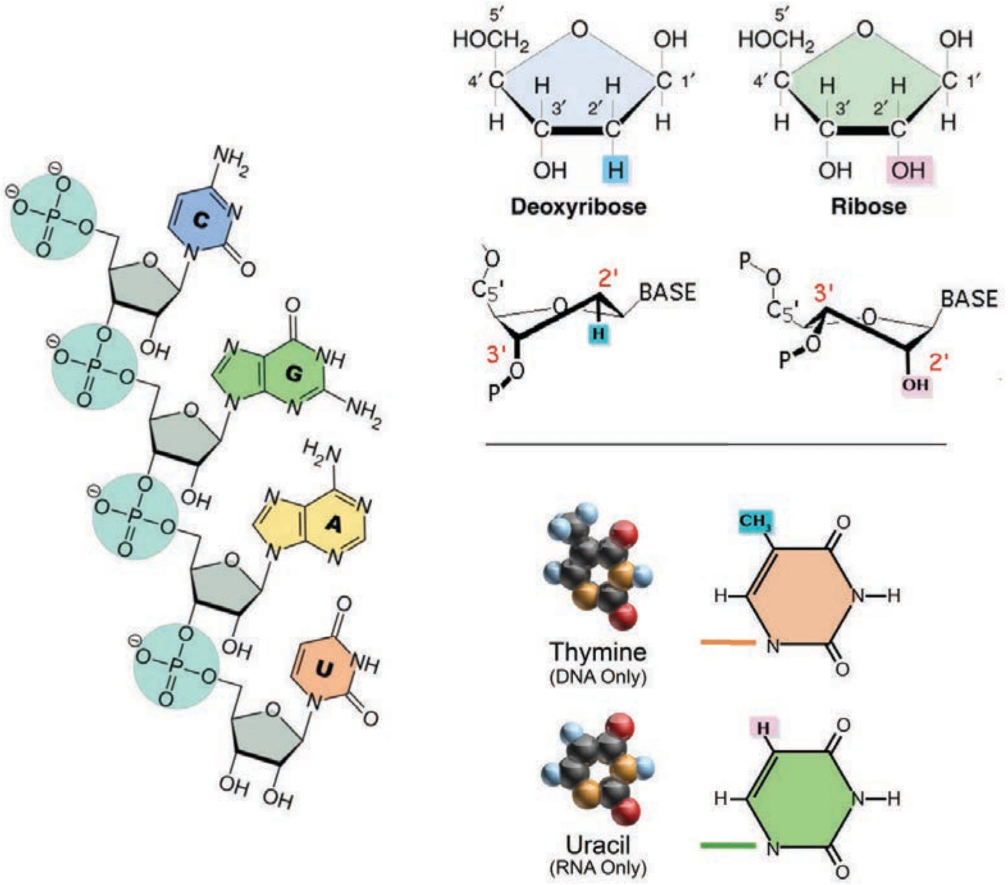


Figure 5 Estructura y composición de bases de una cadena de ARN; Derecha: Principales diferencias entre la composición molecular de ADN y ARN

Los pares de bases en el ARN, análogos a los apareamientos canónicos de Watson-Crick en el ADN, son A·U y G·C; pero hay un modo de apareamiento más, que preserva las dimensiones generales de una hélice de ARN, que es el apareamiento wobble G·U. El par G·U tiene aproximadamente la misma estabilidad que un A·U [3,4] y se observa comúnmente en moléculas de ARN.

Estos tres principales modos de apareamiento de bases en el ARN también mantienen el grupo ribosa de ambas bases en el mismo lado del par, lo que permite la definición de un surco mayor y menor. El ARN es típicamente un polímero monocatenario y la formación de hélices y otras interacciones son posibles porque la molécula se repliega sobre sí misma, de forma similar a una proteína.

En un nivel intermedio de análisis, denominado estructura secundaria, el elemento estructural fundamental de la secuencia de ARN es la doble hélice. Una vez que se especifican las hélices, las regiones no apareadas entre ellas se pueden clasificar en varios tipos de elementos estructurales, denominados colectivamente bucles. Los bucles pueden ser internos, entre dos tallos de hélice (un bucle interno de una sola cara se llama una protuberancia), bucles en horquilla, que consisten en varias bases separadas que están limitadas a cada lado por la misma hélice o bucles de múltiples bifurcaciones en la intersección de tres o más tallos de hélice (Figura 6). La estructura terciaria del ARN se compone de unos pocos tipos de entidades estructurales recurrentes, denominados colectivamente *motivos de ARN*, que se usan en diferentes combinaciones (como bloques de construcción) para generar una rica variedad de formas moleculares. En el artículo de revisión "*Modeling, Simulations, and Bioinformatics in the service of RNA Structure*", que forma parte de esta tesis, se ha recopilado un compendio de métodos teóricos para ayudar a la caracterización del complejo espacio conformacional multi-escala del ARN.

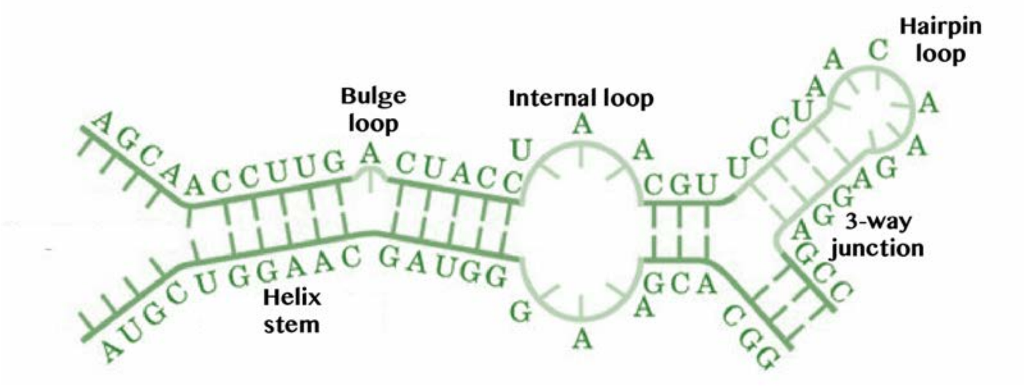


Figure 6 Representación de los elementos de la estructura secundaria más comunes en ARN.

Dinámica dependiente de la secuencia del ADN.

El conocimiento actual de las características estructurales del ADN revela irregularidades dependientes de la secuencia en el espacio conformacional de diferentes hélices dobles. Algunos de los principios que relacionan la secuencia con la estructura han sido derivados a partir del análisis de los datos de cristalografía disponibles para el ADN y complejos proteína-ADN, que proporcionan información no solo sobre la geometría de equilibrio, sino también sobre la flexibilidad esperada de los *bps* [95,96,98]. Una fuente alternativa de parámetros para describir la estructura y flexibilidad del ADN son las simulaciones de dinámica molecular (DM) atomística. Los métodos de DM pueden cubrir mucho mejor el espacio de las secuencias que el análisis de estructuras experimentales y sus resultados están libres de artefactos de red y son consistentes con la presencia de ligandos y el solvente del entorno. Sin embargo, las descripciones de propiedades del ADN derivadas de DM son solo tan precisas como la calidad de los parámetros del campo de fuerza utilizado para describir las interacciones del ADN (ver la sección sobre DM a continuación) y la capacidad de muestrear suficientemente el espacio conformacional.

Notablemente, los resultados basados en DM recopilados por el Ascona B-DNA Consortium [8,9] revelaron dos hallazgos principales que desafiaron los modelos actuales de flexibilidad derivada de la secuencia del ADN. En primer lugar, el modelo del vecino más cercano es insuficiente para describir la flexibilidad del ADN ya que los parámetros de la hélice asociados a un dinucleótido pueden diferir en mayor cantidad según las bases que lo rodean que si se lo compara con otros dinucleótidos. En segundo lugar, una gran cantidad de distribuciones de equilibrio de los parámetros helicoidales tienen desviaciones no despreciables de la normalidad, está claro que la aproximación armónica (definida por los valores de equilibrio y la rigidez asociada) implícita en los modelos elásticos es inexacta. Una cuidadosa caracterización de estos efectos polimórficos y dependientes de la secuencia se ha realizado extensamente en la presente tesis y estos resultados se recogen en los trabajos "*The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA*", "*The Physical Properties of B-DNA beyond Calladine's rules*" y "*Long-Range Effects Modulate Helical Properties of some DNA Dinucleotide Pairs*".

Interacciones proteína-ADN

La interacción de las proteínas reguladoras con el ADN es esencial para la realización adecuada de un gran número de procesos biológicos que van desde la regulación de la expresión génica hasta la replicación, reparación y empaquetamiento del ADN. Como primera clasificación de las interacciones, el reconocimiento de una molécula de ADN por un pequeña molécula o proteína puede ser altamente específico, reconociendo únicamente una secuencia determinada en un gen o incluso un genoma, o inespecífica, sin una unión preferencial significativa a una secuencia de nucleótidos particular.

La unión **no específica** proteína-ADN ocurre a través de dos mecanismos clave relacionados [10], por los cuales se reconoce la atracción electrostática general entre las proteínas y el ADN, o la geometría global del ADN (13). Las proteínas se unen **específicamente** a las secuencias de ADN mediante dos estrategias comúnmente denominadas lectura "directa" e "indirecta". En una lectura directa, la secuencia de ADN es leída a través de contactos específicos entre las cadenas laterales de aminoácidos y los grupos funcionales básicos expuestos en la interfaz proteína-ADN. En una lectura indirecta, las proteínas reconocen las secuencias de ADN a través de variaciones dependientes de la secuencia en la flexibilidad y los parámetros estructurales tales como el ancho del surco, el *twist* entre los pares de bases o la conformación de la cadena principal.

Los aspectos dinámicos de las interacciones proteína-ADN pueden separarse principalmente en dos tipos: la dinámica del proceso de unión en sí y la dinámica del complejo después de su formación. La visión consensuada de un evento de reconocimiento típico es que la proteína primero se une no específicamente al ADN, y luego se difunde a lo largo de la doble hélice mediante la formación de puentes de sal transitorios entre los grupos funcionales cargados de aminoácidos y los átomos de la cadena principal en el ADN [11]. Una vez que alcanza su sitio de unión, la proteína reconoce las propiedades de deformabilidad específica de su motivo de unión al ADN, que le permiten deformar fácilmente el ADN para alcanzar una cierta conformación [12,13]. Los complejos proteína-ADN, después de su formación, también se someten a procesos dinámicos interesantes e importantes y existe una cantidad sustancial de evidencia reciente de que el conjunto configuracional no se colapsa al unirse. En esta tesis estudiamos un caso muy interesante de interacción proteína-ADN, acerca de la cooperatividad de pares de proteínas en la unión al ADN, sin haber interacción directa entre las dos proteínas. El estudio se llama "*Allosterism and signal transfer in DNA*".

Simulaciones de dinámica molecular de ácidos nucleicos.

Hay tres aspectos principales para un cálculo de DM: 1) el modelo que describe las interacciones moleculares; 2) el cálculo de energías y fuerzas del modelo, que debe hacerse con precisión y eficiencia; 3) el algoritmo utilizado para integrar las ecuaciones de movimiento. En la forma más simple de DM, las trayectorias de átomos y moléculas se determinan resolviendo numéricamente las ecuaciones de movimiento de Newton para un sistema de partículas interactuantes (Alder 1959, Rahman 1964, Lifson 1968). Por defecto, el equilibrio de la dinámica corresponde al conjunto microcanónico de la mecánica estadística (energía total constante), pero se pueden introducir términos adicionales si se espera que la simulación mantenga constantes algunas propiedades macroscópicas, como la presión o la temperatura.

La simulación de dinámica molecular requiere la definición de una función de energía potencial, que junto con un conjunto de parámetros empíricos se denomina campo de fuerza (*forcé-field*). Los campos de fuerza más comunes consisten en una suma de fuerzas de enlace (asociadas con longitudes de enlace químico, ángulos de enlace y diedros de enlace) y fuerzas entre moléculas sin enlace (interacciones electrostáticas asociadas e interacciones de Van der Waals). Los valores de los parámetros se obtienen ajustando a cálculos electrónicos detallados (simulaciones de mecánica cuántica, QM) o una variedad de propiedades físicas experimentales.

Los primeros intentos de describir el ADN a partir de trayectorias de varios nanosegundos condujeron a la destabilización de las estructuras. La mejora más exitosa para esos campos de fuerza fue el estándar por excelencia durante más de una década, ya que produjo descripciones razonables de las propiedades del ADN en el régimen de multi-nanosegundos. Muy recientemente, a medida que se hicieron disponibles simulaciones de multi-microsegundos, fueron detectados varios errores en la parametrización parmbsc0 [14–20] y varios grupos desarrollaron correcciones de variable confiabilidad [19,21]. Nuestra propia contribución a este esfuerzo es presentada y discutida en la presente tesis en el trabajo "*Parmbsc1: a refined force field for DNA simulations*".

2 Objetivos.

El objetivo principal de esta tesis es ofrecer una visión integral y consensuada de las propiedades estructurales y dinámicas del ADN en condiciones fisiológicas. Los trabajos presentados aquí se han ido construyendo gradualmente unos sobre otros, y los resultados se han ido acumulando sucesivamente para formar un compendio de mecanismos interdependientes. El orden en que se presentan los resultados no siempre es cronológico. En el proceso de una tesis doctoral, hay muchas situaciones en las que uno está estancado o en las que "los árboles no están dejando ver el bosque", pero en retrospectiva, al analizar el resultado final, es fácil construir una sucesión de logros graduales. A continuación, presento la acumulación lógica de objetivos, como se ve después de una tesis doctoral.

- **Benchmarking** del campo de fuerza del ADN de vanguardia parmbsc1 probándolo en una gran variedad de sistemas de ADN bajo diversas condiciones. Esto es claramente un prerrequisito para usar con confianza simulaciones de DM en el estudio de ADN-B. Se debe probar que las trayectorias muestrean el espacio conformacional del ADN-B a fondo y de manera exhaustiva.
- **Explicar los polimorfismos del ADN-B** es probablemente la clave para elucidar el rompecabezas de sus intrincadas propiedades mecánicas dependientes de la secuencia que, en última instancia, rigen la mayoría de las funciones biológicamente relevantes de la doble hélice.
- Desarrollar un conjunto exhaustivo de reglas que rigen los **efectos de la secuencia en un ADN-B a nivel de tetranucleótidos**. Combinamos el nuevo campo de fuerza parmBSC1 y los últimos conocimientos en el área de polimorfismos en el espacio helicoidal, para brindar una descripción y explicación completas a nivel de tetranucleótidos para los polimorfismos de las diferentes bases, pares de bases y *bps*, y sus interconexiones.
- Descifrar los **efectos más allá de los tetrámeros** en el espacio conformacional del ADN-B, para conocer su contribución a la dinámica del ADN. Nuestro objetivo es determinar la fuerza, la relevancia y, en última instancia, los mecanismos de modulación conformacional de largo alcance mediante patrones de secuencia específicos.
- Aplicar el conocimiento de las propiedades intrínsecas del ADN al estudio del reconocimiento entre proteínas y ADN, así como la **unión cooperativa de proteínas al ADN**. Finalmente nos propusimos descubrir cómo la comunicación de largo alcance a través del ADN, como se demuestra a partir de los efectos de la secuencia, tiene un impacto en su papel en las interacciones proteína-ADN.
- Elaborar un compendio de **enfoques computacionales para el modelado de ARN**, que obliga a los investigadores a mirar más allá de los esquemas comunes de simulación clásica o cuántica. Nuestro objetivo es

resumir el alcance y el desafío de los enfoques más recientes creados para caracterizar la gran variedad conformacional del ARN, lo que debería ayudar a guiar el desarrollo de una nueva generación de métodos capaces de hacer predicciones cuantitativas sobre la estructura y las propiedades físicas del ARN.

3 Resumen de resultados.

3.1 Parmbsc1: a refined force field for DNA simulations

(Parmbsc1: un campo de fuerza refinado para simulaciones de ADN)

Sinopsis. Presentamos parmbsc1, un nuevo campo de fuerza para la simulación atomística del ADN, que se ha parametrizado a partir de datos mecánicos cuánticos de alto nivel y ha sido probado en casi 100 sistemas (~ 140 μs) que cubren la mayor parte del espacio estructural del ADN. Parmbsc1 proporciona resultados de alta calidad en diversos sistemas, resolviendo problemas de campos de fuerza previos. Parmbsc1 pretende ser un campo de fuerza de referencia para el estudio del ADN en la próxima década. Los parámetros y las trayectorias están disponibles en http://mmb.irbbarcelona.org/ParmBSC1/.

Corregimos el perfil acoplado ε/ζ, la torsión glicosídica χ, así como los parámetros de *puckering* usando cálculos de QM de alto nivel tanto en fase gaseosa como en solución. Los nuevos parámetros ε/ζ mejoraron la representación de equilibrio BI/BII (que por definición está determinado por estos dos diedros, con trans/gauche- que representa el estado canónico de BI, mientras que BII está dado por gauche-/trans de ε/ζ) y rectificado las distribuciones de *twist* y *roll*, que están estrechamente correlacionadas con el estado de la cadena principal. La corrección de χ abordaba el equilibrio anti/syn de la orientación de la base y permite simulaciones precisas de estructuras de ADN no canónicas, y también reduce el *fraying* de la base del extremo. El *puckering* se actualizó debido a su acoplamiento a las torsiones modificadas y mejoró el sesgo de parmbsc0 hacia las conformaciones *East*.

También validamos minuciosamente no solo la reproducción correcta de estructuras experimentales, sino también contra observables experimentales, tales como NOE y RDC de experimentos de RMN, longitudes de persistencia (*persistence length*) y tasas de transición en diferentes tipos de solventes. Las propiedades dinámicas se evaluaron mediante experimentos de plegamiento y desplegamiento y transiciones conformacionales (tales como A- a B-ADN). Probamos su aplicabilidad a varias configuraciones de ADN exóticas y no canónicas, al ADN en complejos con proteínas y ligandos, y en una serie de solventes. Esto significó un tiempo de simulación acumulativo de ~ 140 μs, y más de 100 sistemas simulados diferentes. La Figura 7 ilustra la excelente correspondencia entre las estructuras experimentales con los promedios del conjunto de las simulaciones con parmbsc1 en comparación con parmsbsc0.

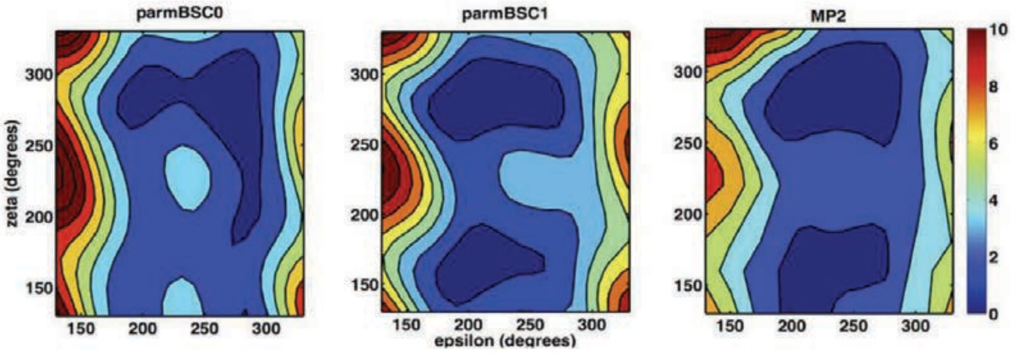


Figure 7 Perfiles de contorno de las distribuciones épsilon/zeta de los cálculos de QM MP2 (derecha) y perfiles de PMF usando los campos de fuerza parmbsc0 (izquierda) y parmbsc1 (centro).

3.2 The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA.

(La función de los enlaces de hidrógeno no convencionales en la determinación de las propensiones BII en B-DNA)

Sinopsis. Es probable que una comprensión precisa de las transiciones de la cadena principal del ADN sea la clave para dilucidar el rompecabezas de las intrincadas propiedades mecánicas dependientes de la secuencia que rigen la mayoría de las funciones biológicamente relevantes de la doble hélice. Un factor que se cree que es importante en el reconocimiento indirecto dentro de complejos proteína-ADN es el efecto combinado de dos torsiones de la cadena principal del ADN (ε y ζ) que dan lugar al bien conocido equilibrio conformacional BI/BII. En este trabajo explicamos la propensión BII dependiente de la secuencia observada en los pasos RpY (R = purina; Y = pirimidina) a nivel de tetranucleótidos con la ayuda de un contacto C-H ···O, no detectado previamente entre átomos, que pertenecen a bases adyacentes. Nuestros resultados están respaldados por extensas simulaciones de dinámica molecular multi-microsegundo del Ascona B-DNA Consortium, de cálculos mecánicos cuánticos de alto nivel y minería de datos de las estructuras experimentales depositadas en Protein Data Bank.

Completamos el rompecabezas de poblaciones de estado BI/BII en los diferentes tetranucleótidos de ADN-B y relacionamos las transiciones de la cadena principal con una compleja interacción de cambios coordinados en la geometría de las bases [15,20], donde las interacciones de puentes de hidrógeno inusuales y los cambios sutiles en el solvente juegan un papel clave. Observamos en las extensas trayectorias de DM de µABC la formación de un contacto C6H6-O3′ en pasos RpY, una interacción análoga al contacto C8H8-O3′ de los pasos RpR. Nuestro análisis claramente relacionó la presencia de esta interacción en la cadena principal de la unión entre dos bases con la transición de la cadena principal a BII en la misma unión (R^2> 0.9) y se encontró que el contacto estabiliza este estado.

Proporcionamos una imagen exhaustiva del mecanismo que induce las transiciones BI/BII en la cadena principal dependientes de la secuencia, señalando cómo esto se logra en los diferentes tipos de *bps*. Aunque la formación de enlaces está en todos los casos extremadamente correlacionada con el estado de la cadena principal, la coreografía más compleja de los cambios en los parámetros helicoidales en los mismos *bps* y las bases vecinas es bastante diferente dependiendo de la secuencia. Además, el análisis de estructuras experimentales de alta resolución de ADN aislado apoya nuestras conclusiones de un acoplamiento dependiente de la secuencia entre los sub-estados de la cadena principal y la formación de enlaces de hidrógeno. Los cálculos Ab initio nos permitieron cuantificar la fuerza relativa de estas interacciones y especular sobre las implicaciones en la estabilidad de la cadena principal a nivel de los tetrámero.

3.3 The Physical Properties of B-DNA beyond Calladine's rules

(Las propiedades físicas del B-DNA más allá de las reglas de Calladine)

Sinopsis. Presentamos un esfuerzo de múltiples laboratorios para describir las propiedades físicas del dúplex ADN-B en condiciones fisiológicas. Al procesar una gran cantidad de datos de simulaciones de dinámica molecular atomística, determinamos las propiedades estructurales dependientes de la secuencia del ADN expresadas en la distribución de equilibrio de su dinámica estocástica. Nuestro análisis incluye un estudio de los momentos de primer y segundo orden (o media y covarianza) de la distribución de equilibrio, que pueden ser capturados con precisión por un modelo Gaussiano o armónico, pero con dependencia de secuencia no local. Posteriormente, caracterizamos la coreografía dependiente de la secuencia de la cadena principal y los movimientos de la base que modulan los efectos no Gaussianos o anarmónicos manifestados en los momentos superiores de la dinámica del dúplex, al muestrear la distribución de equilibrio. Contrariamente a las suposiciones anteriores, tales deformaciones anarmónicas no son raras en el ADN y pueden jugar un papel importante en la determinación de la conformación del ADN dentro de los complejos. Los polimorfismos en las geometrías helicoidales son particularmente frecuentes para ciertos contextos de secuencias de tetranucleótidos, y siempre están acoplados a una red compleja de cambios coordinados en la cadena principal, siendo los equilibrios BI/BII un determinante principal. El análisis de nuestras simulaciones, que contienen ejemplos de las 136 secuencias distintas de tetranucleótidos, nos permite reformular las reglas de Calladine, utilizadas durante décadas para interpretar la geometría promedio del ADN de acuerdo con la supuesta dependencia de secuencia local y las fluctuaciones armónicas, de una manera más precisa, lo que lleva a un conjunto extendido de reglas con poder predictivo cuantitativo que abarca dependencia de secuencia no local y fluctuaciones anarmónicas.

Nuestros resultados determinan que los parámetros helicoidales son transferibles (con pocas excepciones) a nivel de tetranucleótidos y nos alientan a hacer observaciones cualitativas de su variabilidad e interdependencia que resultarían confiables (Figura 8). El ADN-B muestrea sus coordenadas internas de forma concertada, generando una coreografía compleja de transiciones conformacionales que modula los polimorfismos de ADN. Por lo tanto, muchos parámetros helicoidales y torsiones de la cadena principal muestran patrones de correlación coherentes específicos de secuencia entre los 3 *bps* de un tetrámero. Los cationes representan un jugador adicional en esta negociación, que tiene la capacidad de modificar sutilmente el paisaje polimórfico del ADN, particularmente a nivel de *bps*.

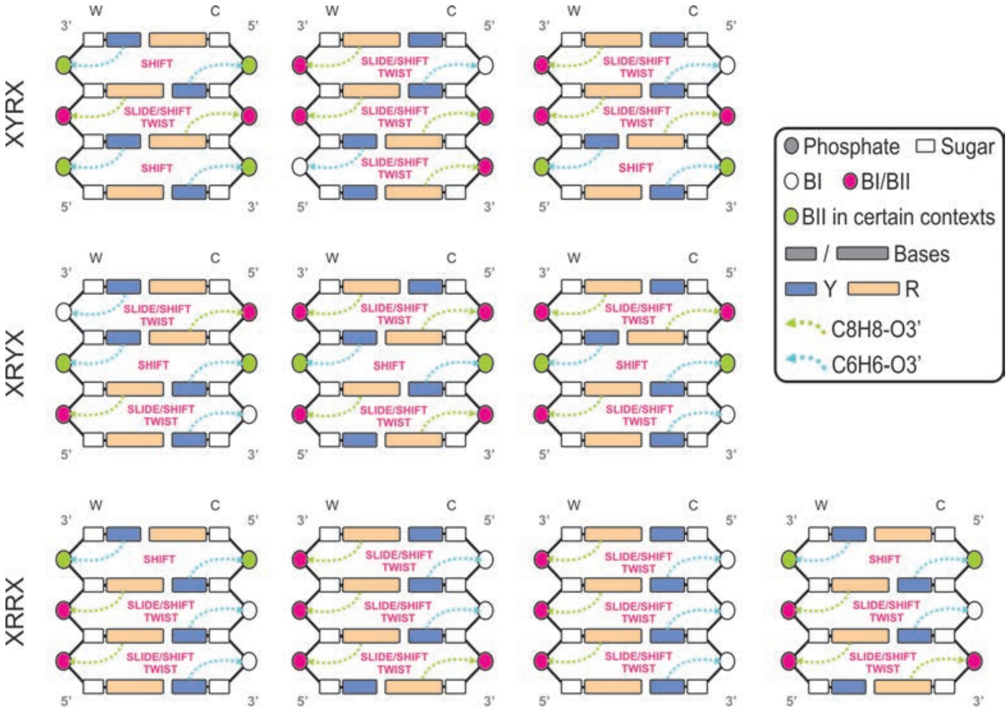


Figure 8 Esquema del paisaje polimórfico del B-ADN a nivel de tetranucleótidos. Los 136 tetranucleótidos únicos se agruparon de acuerdo con purinas (R) y pirimidinas (Y), para las cuales solo existen 10 combinaciones únicas.

3.4 Long-Range Effects Modulate Helical Properties of some DNA Dinucleotide Pairs.

(Efectos de largo alcance modulan las propiedades helicoidales de algunos pares de dinucleótidos de ADN)

Sinopsis. Usamos extensas simulaciones de dinámica molecular para estudiar las propiedades estructurales y dinámicas del paso central d(TpA) en el tetrámero altamente polimórfico d(CpTpApG). Contrariamente a la suposición de los vecinos cercanos (modelo dímero) y los próximos vecinos (modelo tetrámero), las propiedades del paso central d(TpA) cambian de manera bastante significativa en función del correspondiente hexámero. Aún más sorprendente, encontramos que en algunos casos las propiedades del d(TpA) central parecen depender de vecinos remotos (más allá del nivel de hexámeros), lo que destaca la existencia de mecanismos para la transmisión a largo plazo de información estructural en el ADN.

Presentamos aquí un análisis detallado de CTAG en diferentes contextos de secuencia. Los resultados demuestran que los efectos de largo alcance modulan las propiedades geométricas del paso central d(TpA). Dichos efectos de largo alcance son muy visibles a nivel de hexámeros, pero sorprendentemente se extienden más allá de este nivel, lo que indica la existencia de un complejo mecanismo de transferencia de información a través del ADN mediante movimientos coordinados de la cadena principal.

3.5 Allosterism and signal transfer in DNA.

(Allosterismo y transferencia de señal en el ADN)

Sinopsis. Analizamos los mecanismos básicos de transmisión de señal en el ADN y los orígenes de la alostería exhibidos por sistemas tales como el complejo ternario BAMHI-DNA-GRDBD. Encontramos que la información de perturbación generada por un evento de unión a proteína primaria viaja como una onda a regiones distantes de ADN, siguiendo un mecanismo de salto. Sin embargo, tal perturbación estructural es transitoria y no conduce a cambios permanentes en la geometría del ADN y las propiedades de interacción en el sitio de unión secundario. El mecanismo alostérico BAMHI-DNA-GRDBD no ocurre a través de ningún modelo tradicional: lectura directa (proteína-proteína), indirecta (reorganización del sitio secundario) o liberación de solvente. Por el contrario, es generado por un mecanismo sutil y menos común mediado por entropía, que podría tener un papel importante para explicar otros efectos cooperativos mediados por el ADN.

Primero discutimos la respuesta estructural, observando las correlaciones y la causalidad en los descriptores geométricos que explicarían la transferencia de información de sitio a sitio en el ADN. Encontramos que la presencia de BAMHI enriquece el acoplamiento entre los grados de libertad de los dos sitios de unión.

Desde una perspectiva termodinámica, eliminamos la posibilidad de una explicación predominantemente entálpica y encontramos que el mecanismo está mediado por la entropía. En términos de cambios de energía libres, la formación del complejo ternario es cooperativa porque además de pagar una penalización por entropía para unirse a su propio sitio, la primera proteína también hace algo del "trabajo" termodinámico desfavorable requerido para endurecer el sitio de unión secundario, puesto que los dos sitios de enlace están acoplados dinámicamente.

Además, adaptamos una metodología de cálculo y descomposición de la entropía de transferencia de la teoría de la información (previamente utilizada en proteínas) para el estudio de nuestro sistema e identificamos fuentes y sumideros de flujo de información a través del ADN.

3.6 Modeling, Simulations, and Bioinformatics in the service of RNA Structure.

(Modelizaciones, simulaciones y bioinformática al servicio de la estructura del ARN)

Sinopsis. Aunque químicamente se acercan al ADN, los ARN pueden adoptar una amplia gama de estructuras, desde hélices regulares hasta conformaciones globulares que muestran una complejidad similar a la de las proteínas. La determinación de la estructura de las moléculas de ARN, crucial para la comprensión de la función, se ve gravemente obstaculizada por su tamaño y flexibilidad, lo que dificulta el uso sistemático de enfoques experimentales. Las técnicas de simulación también están sufriendo problemas muy graves, relacionados con la precisión de los métodos y su capacidad para muestrear un espacio conformacional grande y complejo. Los enfoques recientes creados para reducir las limitaciones de la generación actual de métodos de simulación serán revisados aquí, siguiendo una descripción sistemática de modelos altamente precisos capaces de tratar con sistemas pequeños, con enfoques *coarse grained* (CG), menos precisos, pero aplicables para tratar con modelos grandes.

Abordamos enfoques computacionales recientes y señalamos sus fortalezas y debilidades, así como las lecciones del pasado que impulsaron su desarrollo. Seguimos una descripción sistemática de modelos QM altamente precisos específicamente aplicables a sistemas pequeños, a representaciones atomísticas clásicas de DM, modelos CG, menos precisos, pero capaces de tratar con modelos grandes y finalmente los enfoques bioinformáticos en auge actualmente.

4 Discusión y conclusiones

Problema de exactitud del campo de fuerza

La utilidad y aplicabilidad de las simulaciones de DM para modelar sistemas biomoleculares depende de su capacidad para muestrear suficientemente el espacio conformacional y la descripción correcta del potencial en términos de la forma funcional del campo de fuerza y el conjunto de parámetros. Claramente, el campo de fuerza define la forma del espacio conformacional para un conjunto dado de posiciones atómicas y también el acceso a los mínimos energéticos. Al simular sistemas en equilibrio, especialmente en sistemas bastante estables como el ADN, los campos de fuerza se esfuerzan por generar conjuntos que reproducen sistemas reales y no tiene por qué ser una gran desventaja con el poder de muestreo. En los últimos años, se ha convertido en tarea de los ingenieros informáticos y los desarrolladores de software abordar el problema de lograr escalas de tiempo largas y biológicamente relevantes.

La convergencia y reproducibilidad de simulaciones de ADN atomístico con campos de fuerza de última generación, como nuestro parmbsc1, se ha demostrado de forma convincente [22,23]. También parece que hasta llegar a una revolución significativa, donde los milisegundos de simulación se vuelven rutinarios, los rangos de muestreo actuales cubren por completo las estructuras internas y la dinámica de los ADN-B en esta escala de tiempo [24].

La creciente confianza ha permitido a muchos investigadores utilizar DM para estudios muy detallados sobre la naturaleza dependiente de la secuencia de oligómeros de ADN y sobre el complejo arsenal de mecanismos que rigen su comportamiento. En cualquiera de estos estudios es necesaria una validación cuidadosa de los resultados ya que aún no está del todo claro qué tan bien y en qué grado se reproducen los efectos de secuencia en DM. El hecho de que la última generación de campos de fuerza coincida muy bien entre sí y que se ajusten a los escasos datos experimentales es seguramente muy alentador, pero pasará algún tiempo hasta que se puedan validar pequeñas diferencias en las geometrías de las secuencias.

Nuestra propia validación extensiva del campo de fuerza parmbsc1, así como una gran cantidad de otros trabajos que, desde su publicación, se han establecido específicamente para evaluar su rendimiento [22,23], o simplemente lo han aplicado con éxito, hablan de una parametrización muy estable capaz de tratar con una amplia gama de ADN. Vale la pena mencionar que en condiciones especiales podrían ser necesarias pequeñas mejoras, lo que podría lograrse con la inclusión de términos de polarización. Sin embargo, hasta la fecha, ningún campo de fuerza ha sido capaz de modelar la polarización sin desestabilizar finalmente el sistema y esto a un costo enorme (un factor de 10) a la velocidad de cálculo.

En resumen, con base en el notable desempeño de parmbsc1, nosotros y otros grupos podemos emplearlo con confianza en el estudio detallado de la dinámica del ADN y esperamos que el número de resultados de soporte solo aumente.

Dependencia de la secuencia y polimorfismos del ADN-B.

Entonces, ¿qué es lo que realmente aprendemos al analizar la variabilidad de conformación del ADN sobre su espacio de secuencia a nivel de los tetrámeros? Está bien establecido que diferentes *bps* tienen diferentes preferencias con respecto a sus geometrías internas, y hasta cierto punto, el conjunto de reglas heurísticas de Calladine es capaz de dar sentido a estas diferencias.

A nivel de *bps*, algunas secuencias son extremadamente estables, como ApT, y algunas secuencias, como CpG, tienen un equilibrio biestable y convierten entre diferentes disposiciones de sus geometrías internas. Hay casos en que esta frustración puede explicarse por la distribución de cargas, el volumen o la fuerza de sus interacciones de apilamiento y los puentes de hidrógeno, pero en muchos casos requiere una visión más integral, teniendo en cuenta los efectos de secuencia de más alto nivel.

En simulaciones de DM de multi-microsegundos, los parámetros de pares intra-base son siempre unimodales ya que los estados alternativos a los que se puede acceder a través de la apertura de la base no se muestrean en esta escala de tiempo. Sin embargo, sus promedios de conjunto muestran diferencias considerables de acuerdo con el cambio en la secuencia. Los parámetros de pares de bases pueden ser bimodales, pero solo en ciertas combinaciones de tetranulceótidos que constituyen aproximadamente el 5% de los casos. Esto puede explicarse teniendo en cuenta que el *bps* central de una combinación particular de cuatro nucleótidos tiene una preferencia estructural que está en conflicto con la de sus pasos vecinos. Con el fin de minimizar el costo de energía y satisfacer de la mejor manera posible todos los requisitos conformacionales, un *bps* más flexible poblará varios estados, generalmente un máximo de dos.

La optimización de las geometrías entre varios *bps* generalmente implica reorganizaciones de la red troncal, con el azúcar fosfato actuando como una bisagra que permite la coordinación consecutiva de *bps* en una coreografía compleja que a menudo involucra otros factores, tales como cambios sutiles en el entorno del solvente. En los ADN-B, la transición principal más importante es BI/BII, que se puede relacionar con la química a través de la fuerza relativa dependiente de la secuencia de puentes de hidrógeno no convencionales que estabilizan las conformaciones BII. En un modelo de tetrámero de ADN-B, las transiciones de la cadena principal de diferentes tetrámeros se traducen en movimientos a lo largo de diferentes grados internos de libertad, dependiendo de la secuencia.

Por lo tanto, ahora podemos construir una imagen del espacio conformacional interconectado del ADN como una superposición de secuencias de tetranucleótidos con descriptores estructurales transferibles. Todavía es una cuestión de especulación cómo estas propiedades podrían ser explotadas por proteínas y otras moléculas que se unen al ADN para diferentes funciones biológicas.

Transferencia de información a través del ADN.

Sin embargo, hay algunos casos especiales en los que el modelo de tetrámero no parece ser suficiente. El CTAG es uno de esos casos que demuestra que, para un tetrámero altamente flexible y polimórfico, la composición de la secuencia de largo alcance puede tener un efecto notable sobre las propiedades estructurales del *bps* central. Analizar el mecanismo detrás de esta comunicación de largo alcance a través del ADN ha significado más que nada una oportunidad para comprender los raros eventos de modulación de secuencia que podrían ser mucho más generales en casos de distorsiones mayores e inducidas en la hélice. En CTAG pudimos observar la influencia de la secuencia no solo desde el nivel del hexámero, sino incluso más allá, y los datos apuntan a un complejo mecanismo de transferencia de información a través del ADN mediante movimientos coordinados de la cadena principal.

En la realización de la función biológica, el ADN a menudo se considera erróneamente como un retículo inerte sobre el cual las proteínas se ensamblan para replicar o transcribir genes. Sin embargo, los experimentos demuestran que la transferencia de información en el ADN puede ocurrir incluso a largas distancias y puede producir efectos alostéricos sobre la unión al ligando (17, 18).

Sin lugar a duda, la unión de proteínas o moléculas pequeñas al ADN puede producir cambios conformacionales acoplados que pueden afectar a un sitio de unión vecino y aumentar su afinidad por la proteína de unión secundaria. Tales cambios no necesitan alterar los promedios del conjunto y solo potencian modificaciones en la forma del pozo de energía en el sitio de unión secundario. Como se ve a partir de la información dinámica proporcionada por una trayectoria de DM, tal vez en más de un caso de parejas de proteínas, el ADN actúa como un cable que transmite pulsos de información originados en el sitio primario de unión que viajan a regiones distantes.

Mostramos que los métodos de DM pueden proporcionar explicaciones razonables para los fenómenos de unión cooperativa en el ADN y abren por primera vez la posibilidad de la "alostería sin cambio conformacional" en el reclutamiento de proteínas al ADN. Desde un punto de vista termodinámico, este tipo de enlace cooperativo parece estar impulsado por la entropía. Por lo tanto, el primer evento vinculante congela algunos de los grados de libertad alrededor de su propia región de unión, pero también reduce el costo de entropía asociado al segundo enlace.

Bibliografía

1. Olson WK, Bansal M, Burley SK *et al.* 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* 2001;**313**:229–37.

2. Lu X-J, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* 2003;**31**:5108–21.

3. Xu D, Landon T, Greenbaum NL *et al.* The electrostatic characteristics of G.U wobble base pairs. *Nucleic Acids Res* 2007;**35**:3836–47.

4. Varani G, McClain WH. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep* 2000;**1**:18–23.

5. Suzuki M, Amano N, Kakinuma J *et al.* Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. *J Mol Biol* 1997;**274**:421–35.

6. Subirana JA, Faria T. Influence of sequence on the conformation of the B-DNA helix. *Biophys J* 1997;**73**:333–8.

7. Olson WK, Gorin AA, Lu XJ *et al.* DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 1998;**95**:11163–8.

8. Pasi M, Maddocks JH, Beveridge D *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* 2014;**42**:12272–83.

9. Lavery R, Zakrzewska K, Beveridge D *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* 2010;**38**:299–313.

10. von Hippel PH, Berg OG. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A* 1986;**83**:1608–12.

11. Mirny L, Slutsky M, Wunderlich Z *et al.* How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J Phys A Math Theor* 2009;**42**:434013.

12. Kalodimos CG, Biris N, Bonvin AMJJ *et al.* Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science (80-)* 2004;**305**:386–9.

13. Ansari A, Kuznetsov S V. Dynamics and Mechanism of DNA-Bending Proteins in Binding Site Recognition. Springer, New York, NY, 2010, 107–42.

14. Lane AN, Chaires JB, Gray RD *et al.* Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res* 2008;**36**:5482–515.

15. Dršata T, Pérez A, Orozco M *et al.* Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J Chem Theory Comput* 2013;**9**:707–21.

16. Heddi B, Foloppe N, Oguey C *et al.* Importance of Accurate DNA Structures in Solution: The Jun–Fos Model. *J Mol Biol* 2008;**382**:956–70.

17. Pérez A, Lankas F, Luque FJ *et al.* Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res* 2008;**36**:2379–94.

18. Fadrná E, Sˇpacˇková N, Sarzyñska J *et al.* Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J Chem Theory Comput* 2009;**5**:2514–30.

19. Krepl M, Zgarbová M, Stadlbauer P *et al.* 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* 2012;**8**:2506–20.

20. Dans PD, Pérez A, Faustino I *et al.* Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res* 2012;**40**:10668–78.

21. Zgarbová M, Šponer J, Otyepka M *et al.* Refinement of the Sugar–

Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J Chem Theory Comput* 2015;**11**:5723–36.

22. Galindo-Murillo R, Robertson JC, Zgarbová M *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* 2016;**12**:4114–27.

23. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:gkw1355.

24. Galindo-Murillo R, Roe DR, Cheatham TE. On the absence of intrahelical DNA dynamics on the μs to ms timescale. *Nat Commun* 2014;**5**:5152.