UNIVERSITAT DE BARCELONA

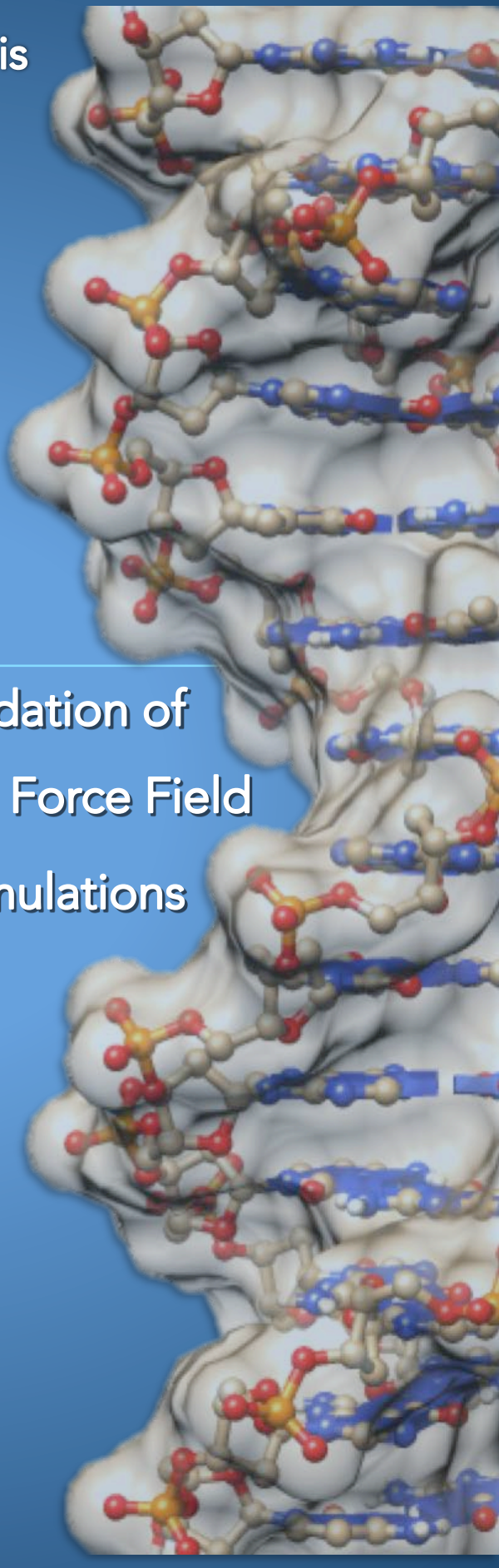
# Parmbsc1: Parameterization and Validation of a new State-of-the-art Force Field for DNA Simulations

Iván Ivani

Doctoral thesis

# Parmbsc1

## Parameterization and Validation of a new State-of-the-art Force Field for DNA Simulations



Ivan Ivani

UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DOCTORAT EN BIOMEDICINA

# Parmbsc1: Parameterization and validation of a new state-of-the-art force field for DNA simulations

IVAN IVANI

SEPTEMBER 2016

UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DOCTORAT EN BIOMEDICINA



# Parmbsc1: Parameterization and validation of a new state-of-the-art force field for DNA simulations

Memòria presentada per Ivan Ivani per optar al grau de doctor/a per la Universitat de Barcelona

TUTOR I DIRECTOR:

MODESTO OROZCO LOPEZ

DOCTORAND:

IVAN IVANI

# ACKNOWLEDGMENTS

Even though this acknowledgment is right at the beginning of this thesis, I write it last, as it is difficult not to forget anyone who in any way helped me in making this work and influenced me during my PhD years. Firstly, this work would not have been possible without the help and mentorship of my supervisor, Prof. Modesto Orozco, from whom I have learned so much during my PhD study. The entire lab, which changed so much during my stay, making me feel old and slowly progressing. I have learned so much from talking and sharing ideas with almost everyone who has been in our lab. A mention of honor would go Michel•la, with who I have spend so much time discussion great ideas that never came to light, but always had fun doing whatever we did from traveling and drinking; Rima, who magically appeared and in the same fashion disappeared to a much better future, leaving us miserable envying her life, but I still love her anyways; Pedro, who doesn't share to much, but is a very good listener (and a great person); others great people I met: Francesco, Hansel, Adam, Pablo, Oscar, Guillem, Rosana, Athi, Nadine, Nacho, Florian, as well as all the newbies, Alexandra, Jurgen, Diana…Neću zaboraviti ni Sanju, perspektivnu naučnicu i buduću zvezdu srpske naučne scene (ako u mežuvremenu ne postane domaćica i majka novog Paca), Antoniju za vrhunske sladolede (batali znanstvenost i otvori neko slastičarnu na Rivi), Stipeta za to što je ljudina.

No puedo olvidar a Marga, una muy buena persona que me ayudo muchisimo. Friends I made at IRB, Mariano, Milos, Vanya, Pep, Giulio… many who have successfully finished and left IRB much sooner than me. Last but not least (in a fashion of scientific papers), Federica, because I know you are reading this thesis just for this, whose young spirit and motivationally laughter always filled the room with joy and melody. I have prayed so many times to God to give her a husband, but it goes to proof that God is dead.

Nikako ne smem da zaboravim svoju majku, koja je najponosnija od svih i koja će jedina čuvati i pokazivati ovaj rad drugima. Moja sestra koja je ogromna podrška u mom životu i koja me razume i voli. Moj otac, koji iako to ne pokazuje, ipak se raduje što sam konačno postao prvi doktor nauka u porodici. Moje tetke koje me vole kao svog rođenog sina i kojima pripada veliki deo moje slave zbog uticaja na moju ličnost i karakter. Čitavo moje društvo u Novom Sadu, Šakotu Milicu, sjajnog zubara i divnu osobu, za njenu redovnu oralnu podršku, Vukču za motivacioni hejt na nivou republičkog zavoda za , Ćureta za iskreno i verno prijateljstvo kao i za sve spojlere koje je ispričao, Krnju, zavodnika svih generacija i pojavu nad pojavama; Borisa i Sikimu, buduće milione i disrupter-e svetskog bankovnog poredka…Na kraju moram da posvetim većinu svog rada svojoj boljoj polovini, Petrović Marini (Pale, 1984), koja me održavala i čistila i prala, ali pre svega volela i bila tu uz mene. Kok

# CONTENTS

# LIST OF FIGURES



# LIST OF TABLES

# OVERVIEW

Classical force fields are the core of classical simulations, particularly of molecular dynamics (MD), a technique that is changing our view on the structure, flexibility and function of biological macromolecules. Originated from the pioneering work of Lifson's group in the sixties, force fields have been in continuous evolution, improving in each generation the accuracy in the representation of proteins and nucleic acid.

Force field development is tightly connected to the refinement of simulation procedures and to the extension of simulation time scales. Thus, as simulation time passed the microsecond barrier, MD simulations have revealed the existence of some errors in the default force field for DNA simulations, parmbsc0 (developed in the group). The goal of this thesis is to address these problems by a reparameterization of AMBER force field that aims to represent a wide range of DNA structures under physiological and non-physiological conditions. Keeping $\alpha/\gamma$ parmbsc0 corrections and parm99 non-bonded parameters, we systematically reparameterized sugar puckering, $\varepsilon$, $\zeta$ and $\chi$ torsions using high level QM calculations both in gas phase and solution. The refined force field has been tested for more than 3 years to an unprecedented level of detail, considering a large variety of DNAs, and analyzing structural, mechanical and dynamical properties of the DNAs resulting from the corresponding MD simulations. The refined force field parameters have been also subjected for more than 1 year of $\beta$-testing by different groups, finding to our knowledge no major drawbacks.

In the world of RNA simulations, despite the recent efforts to improve the description of RNA in MD simulations, RNA force fields are still far in accuracy from those of DNA. A probable cause could be the incomplete understanding of the mechanism of 2'-OH orientation, which in big extent determines the RNA conformation and most probably serves as the molecular switch.

## THESIS ORGANIZATION

This thesis is compiled of five publications (or in the process of publication) works; first three consider DNA force field development and following validation and benchmark while the last two are focused on RNA efforts. For better understanding of this work **Chapter 1** introduces the central concepts related to nucleic acids, their structures and ways to study them. **Chapter 2** goes into more details of the methodology employed here, briefly explaining basic QM formalism and MD

simulations with an emphasis on force fields. **Chapter 3** is a small handbook of methods employed in the analysis in this work. All together first three chapters should provide a solid ground to better understand the details and the relevance of the five publications in the following two chapters. **Chapter 4** is based on the development of new force field, called parmbsc1, its further testing on the Drew-Dickerson sequence and benchmarking. **Chapter 5** focuses on efforts to understand the mechanism of complexity of RNA structures studying 2'-OH rotation, and computational design of a new RNA dumbbell structure. A summary of the major results and a general discussion that reflects on the five projects and future work are presented in **Chapter 6**, with the main conclusions at the end of this work.

*"All great ideas and all great thoughts have a ridiculous beginning."*
*Albert Camus*

# 1| INTRODUCTION

This chapter briefly introduces the reader to topics that will be discussed in this thesis, such as principles of nucleic acids' structures, definition of DNA in helical space and theoretical methods of studying nucleic acids. For more extensive description, the reader is referred to (Saenger 1984; Neidle 2008).

## MY FIRST ENCOUNTER WITH DNA

In 1995, a short scientific report was aired on Serbian National Television. It was about Human Genome Project. Several leading scientist were talking about how we are going to better understand complex processes that occur within our body, from a single sequence consisting of just 4 letters. These 4 letters, or 4 nucleic bases, is what makes the DNA, our "personal code". It governs how we look, how we age and possibly how we behave. I was 9 y.o. at the time and could hardly keep attention on anything for longer than 5 minutes, but the representation of biomolecules in our cells stayed in my mind ever since. The idea that we can visualize and see the small world of building units that makes us, and the processes that are going on inside of us got me truly inspired and eager to understand more. My mother, on the other hand, wanted to understand why my behavior was hyperactive and how could she change it.

## 1.1. One molecule to rule them all

In 1944, years after revolution in physics and birth of quantum mechanics, the famous physicist, Erwin Schrödinger, wrote a book called "What is life?", which focused on one important question: *How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?*. By that time, Charles Darwin's work had already explained life after it got started, but a big question still bothered scientist, *What is the essence of life?* Schrödinger's answer was that the essence was the information that had to be present in a form of a molecule, an "aperiodic crystal" as he called it, and in configuration of covalent chemical bonds. In that time most of the scientific community believed that proteins carried the hereditary material, but two young scientists from Cambridge, Francis Crick and James Watson, changed everybody's mind by discovering the structure of DNA. Inspired by Schrödinger's idea, Crick and Watson, managed to construct the correct model of DNA double-helix based on X-ray diffraction images collected by Rosalind Franklin and Maurice Wilkins (see Figure 1.1). Francis Crick poetically called this molecule the *secret of life*. Their work was published in Nature in 1953 (Watson & Crick 1953), for which they jointly received a Nobel Prize in 1962.[1]

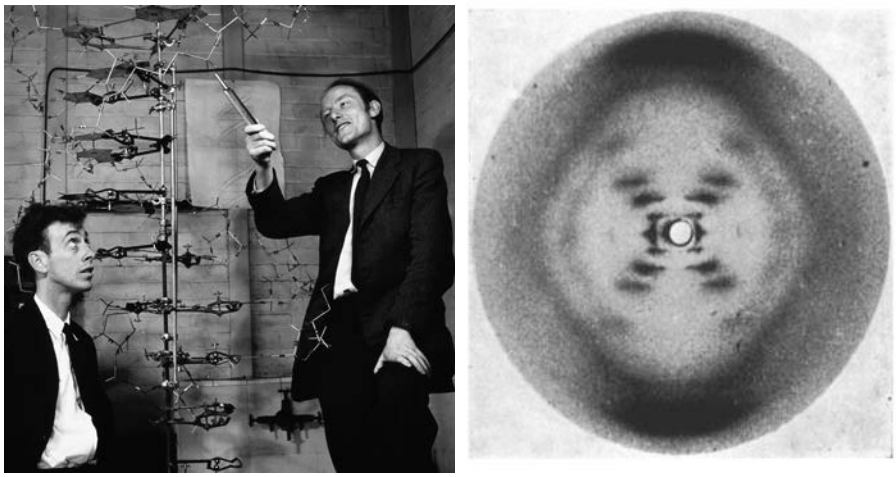


**Figure 1.1. Watson and Crick's discovery.** Watson and Crick next to their model of DNA (left), and the X-ray diffraction image of DNA that they used for constructing the model (right).

Previous experimental findings from Erwin Chargaff, who found that the molar ratios of adenine:thymine and cytosine:guanine in DNA are both unity inspired and coincide with the Watson-Crick's proposed model. Chargaff also discovered that the

[1] The term 'double-helix' got popular with the publication of James Watson's book 'The Double-helix: A Personal Account of the Discovery of the Structure of DNA' in 1968.

proportion of the bases in the DNA molecule differed widely from species to species, which all together matched with Schrödinger's hypothesis of non-repetitive molecule. Combined with Watson and Crick structural model Chargaff's experiments explained basic principles of base-pairing, how genetic code is written and how it can be replicated and transcribed. Since that, genetics has advanced significantly. Scientists have discovered many aspects of biological function of DNA (Crick 1970); how it is transcribed into messenger RNA, which are later translated into protein sequences, the exact mechanism of DNA replication, the process of code recognition, the mechanisms of DNA compaction into chromatin structure, and the basic principles of DNA regulation, amongst many other relevant processes. For more detailed view, the reader is referred to specific books, such as (Watson et al. 2003; Boyle 2008).

## 1.2 The structure of DNA

DNA is a long polymer of non-static structure made of repeating units called nucleotides. In living organisms it does not exist as a single molecule, but usually as a pair of molecules held tightly together in the shape of a double-helix. A nucleotide unit consists of a phosphate-deoxyribose segment, which is the part of the backbone holding the chain together, and a nucleobase, which interacts with the other DNA strand in the helix. Two main forces stabilize the DNA in a double-helix: *base-stacking interaction* among aromatic nucleobases and *hydrogen bonding* between them.

### Nucleic bases

There are only 4 coding nucleo bases found in DNA classified in two types: *the purines*, adenosine (A) and guanine (G), and *the pyrimidines*, cytosine (C) and thymine (T). The bases are planar aromatic heterocyclic molecules and are divided into two groups – the pyrimidine bases, C and T, and the purine bases, A and G.[2] In natural base-pairing scheme, also called *Watson-Crick pairing*, adenine pairs with thymine and guanine pairs with cytosine (see Figure 1.2). They are generally abbreviated as A-T and G-C base-pairs. Modified forms of these 4 DNA bases occur naturally and some of them have biological significance (like methylated-Cytosine). Recent findings have showed that levels of modified bases naturally occurring in DNA, mostly modification on cytosine, play a major role in gene silencing and DNA compaction (Keshet et al. 1986; Kriaucionis & Heintz 2009; Hashimoto et al. 2012).

[2] For those cracking their head on how just 4 letter can bring such a huge verity between species, consider that one DNA molecule has on average 220 million bps, giving an approximate $2*10^{32}$ combinations.

## Base pairing

In their DNA model, Watson and Crick proposed that purines and pyrimidine bases are held together by specific hydrogen bonds, to form planar base pairs, what is now knows as *Watson-Crick base pairing* (W-C pairing). In this arrangement, adenine and thymine (A•T) base pair has two hydrogen bonds compared to the three in guanine and cytosine (G•C) pair. Having an additional hydrogen bond makes G•C base pair more stable (around 1.5 kcal/mol, coming from theoretical studies (Stofer et al. 1999)). In native, double-helical DNA the two bases in a base pair necessarily arise from two separate strands of DNA and so hold the DNA double helix together (see Figure 1.2).

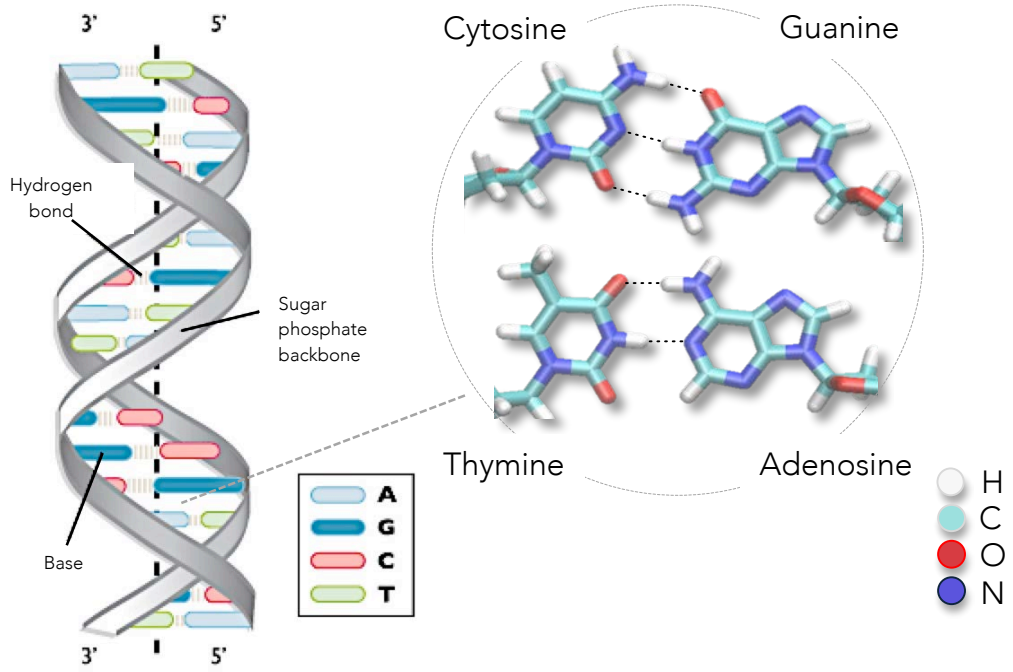



**Figure 1.2. Structure of DNA double-helix.** DNA double-helix with a highlighted nucleotide unit (left), and the 4 DNA bases in paired conformation (called Watson-Crick pairing) (right).

## Helical parameters

The individual bases in a nucleic acid are mostly flat, but base pairs can show considerable flexibility, mainly related to their base-stacking environments. Thus descriptions of base morphology are important in describing and understanding many sequence-dependent features and deformations of nucleic acids. 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”). Most common approach to calculate these parameters consists of defining them with respect to a global helical axis, which need not be linear. Accordingly, the geometry of a base pair can be characterized by means of 6 translational and rotational coordinates:

(a) **Propeller twist** between bases is the dihedral angle between normals to the bases, or their longer axis, **Buckle** is the dihedral angle between bases, along their short axis, while **Opening** is the dihedral angle between bases along the helix axis.
(b) **Stretch**, **Shear** and **Stagger** are displacements of a base pair from each other regarding their long axis, their short axis and the helix axis, respectively.

Base-pair step is characterized with 10 coordinates, 4 of which are referred to the previous step:

(a) **Inclination** is the angle between the long axis of a base pair and a plane perpendicular to the helix axis, while **Tip** is the angle between the short axis of the base pair and a plane perpendicular to the helix axis.
(b) **X-displacement** and **Y-displacement** define translations of a base pair within its mean plane in terms of the distance of the midpoint of the base pair long axis from the helix axis. For example, positive values of X displacement mean displacement towards the major-groove (see later).

And 6 coordinates referred to the helix axis, of which 3 rotational and 3 translational:

(a) **Helical twist** is the angle between successive base pairs about the helix 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 its neighbor, about the long 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 short axis of the base pair.
(b) **Slide** is the relative displacement of one base pair compared to another in the direction of long axis of the base pair, measured between the midpoints of each C6-C8 base pair long axis. Similarly, **Shift** is defined as the relative displacement of a base pair from another in the direction of the base pair short axis, and **Rise** being the displacement of one base pair from its neighbor in the direction of helix axis.

All these parameters are called **helical parameters** (see Figure 1.3). Together, they fully characterize the helical structure of the molecule.

Twist and rise determine the *handedness* (in the direction of the helix axis) and *pitch* (distance between successive nucleotide units per complete turn of the helix), respectively, of the helix, while others can be zero. Some of the helical parameters exhibit coupled behavior, for example, twist, roll and slide change simultaneously with overall bending. Similarly, some helical parameters are connected with particular backbone dihedrals (like twist is connected with epsilon/zeta).

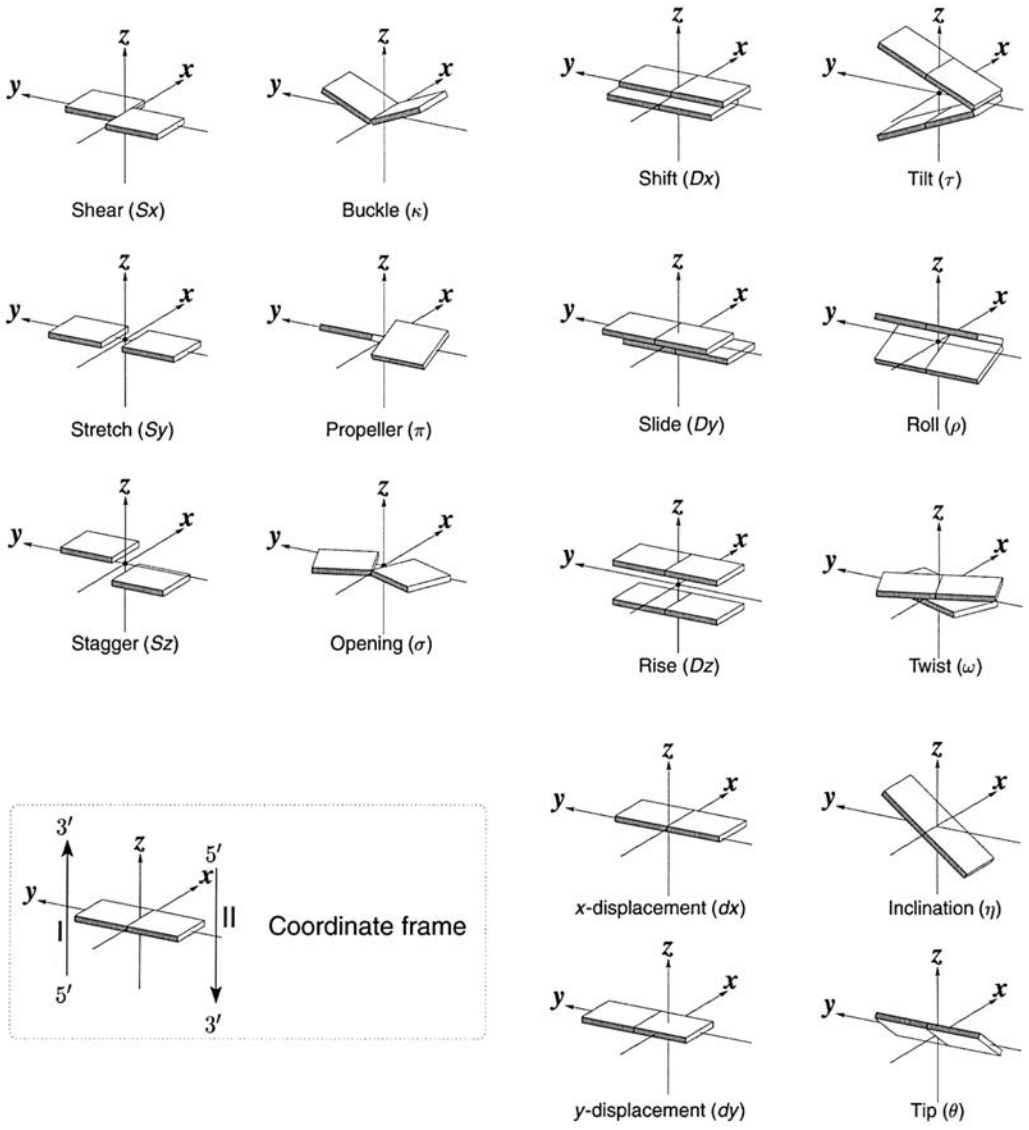



**Figure 1.3. Base-pair geometry.** Helical parameter for base-pair (A) and base-pair step (B) (taken from (Lu & Olson 2003)).

# DNA degrees of freedom

## Backbone dihedrals

The detailed conformation of the DNA is defined by torsion angles. DNA torsion angles in a nucleotide consist of *6 backbone*, *5 sugar* and *one glycosidic torsion angle*. Following the generally accepted atomic numbering scheme of DNA (see Figure 1.4), the torsion can be defined as described in Figure 1.4 (where $\alpha = O_{3'} - P - O_{5'} - C_{5'}$ is the rotation about $P\text{-}O_{5'}$ bond). Backbone torsions have 3 major ranges, *gauche-*, *gauche+* and *trans*, where some is more dominant while others could be rarely occupied. For the definition of torsion angle ranges see Figure 1.4.

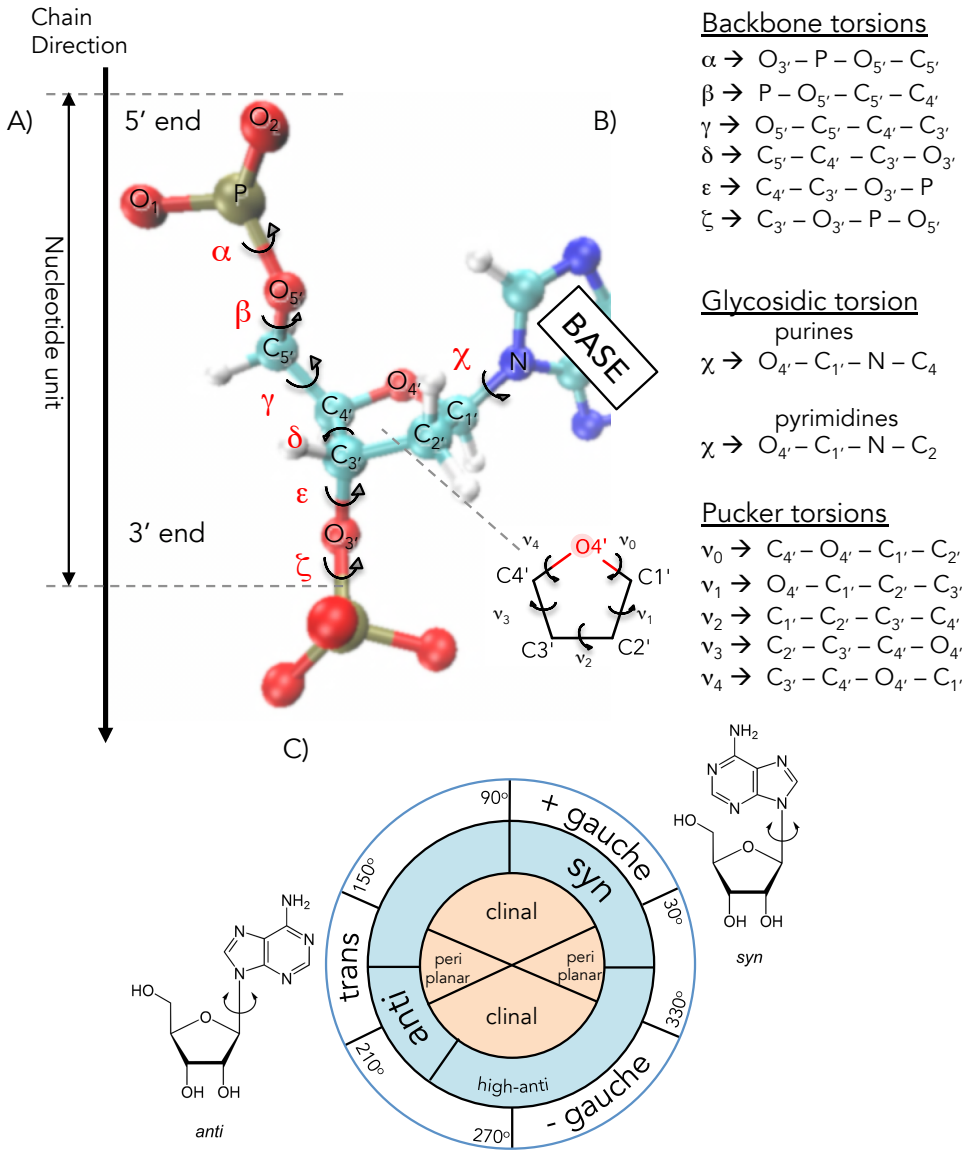



**Figure 1.4. Definition of DNA torsions.** (A) DNA atomic numbering (B) the definition of torsion angles and (C) definition of torsion angle ranges commonly used to describe them (in blue circle are defined sterically allowed regions of glycosidic

torsion).

Similarly to sugar moiety, the base can adopt two main orientations about the glycol $C_{1'}$ - N bond, **syn** and **anti,** and a minor one, *high-anti*. They are defined by torsion angle $\chi$ ($O_{4'}$ - $C_{1'}$ - $N_9$ - $C_4$ for purine and $O_{4'}$ - $C_{1'}$ - $N_9$ - $C_2$ for pyrimidine nucleosides). The sterically preferred ranges for the three domains of $\chi$ are *anti* = (180°, 240°) and *syn* = (0°, 90°), with *high-anti* region being about 270°. In *anti* conformation the Watson-Crick hydrogen bonding groups are directed away from the sugar ring, while in *syn* conformations these groups are oriented towards the sugar and especially its $O_{5'}$ atom (see Figure 1.5). For purine bases, the *syn* conformation is slightly less preferred than the *anti*, except for the guanosine nucleotides, where the *syn* is stabilized due to favorable electrostatic interactions between the $N_2$ amino group of guanine and the 5' phosphate atom. For pyrimidine nucleotides, the *anti* conformation is largely favored respect to the *syn* conformation, because of unfavorable contacts between the $O_2$ oxygen atom of the base and the 5'-phosphate group. The energy difference between the two angle conformers vary depending on the nucleotide unit and the arrangement, but are in general range of 0.9 to 3.6 kcal/mol (Neidle 2008). In most of the DNA (and RNA) conformations bases stay in *anti* orientation, but in some exotic conformations of DNA (as well as in RNA), such as quadruplexes or H-DNAs, *syn* conformation plays a significant role (see later).

Even though, having 6 backbone torsions means (in principle) 6 degrees of freedom, motions of some torsional angles are correlated, generating torsional couples. First couple, ***α/γ***, controls the orientation of the phosphate group to the furanose in the nucleotide unit. As $\beta$ torsion stays mainly in the trans region (180º), $\alpha/\gamma$ couple controls the orientation of the phosphate group and contribute to the overall structure of the molecule. For example, the canonical form of $\alpha/\gamma$, *gauche-/gauche+*, forms a B-DNA double-helix, while *gauche+/trans* would yield a ladder-like structure (a common pam99 problem (Várnai & Zakrzewska 2004; Pérez et al. 2007); see Chapter 2.4).

Another torsional couple is ***ε/ζ***, torsions around oxygen $O_{3'}$. Their local changes are coupled with the motion of $O_{3'}$ atom and the phosphate group from the following nucleotide unit (on the 3' side). Two major regions in $\varepsilon/\zeta$ landscape are associated with *high-twist* and *low-twist* states, also known as $B_I$ and $B_{II}$ (see Figure 1.5). In the more popular high-twist mode, $\mathbf{B_I}$, the helix is in the canonical form, while in low-twist mode, $\mathbf{B_{II}}$, the phosphate group is pushed towards the minor groove, narrowing it. Moreover the $B_I$ and $B_{II}$ states are sequence dependent, making $B_I/B_{II}$ dynamics biologically relevant (Heddi et al. 2006). Other studies have shown the coupling of $B_I/B_{II}$ states with sugar puckering (Wu et al. 2003).

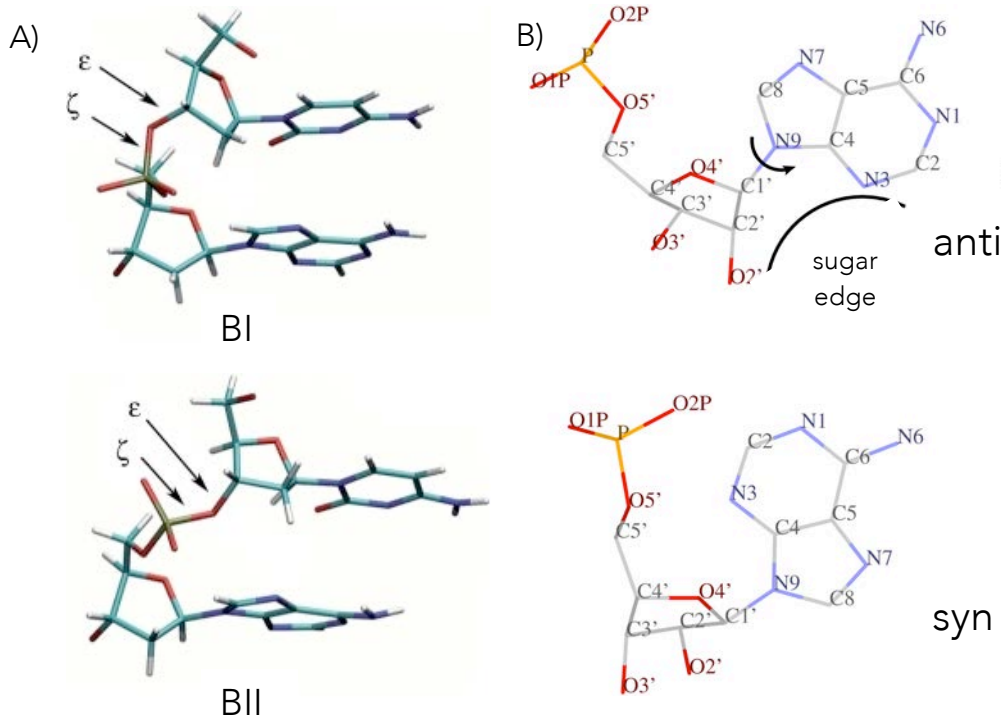



**Figure 1.5. Important torsional conformations.** (A) $B_I$ / $B_{II}$ and (B) *syn* and *anti* orientations of the base.

## Sugar pucker

DNA five-membered furanose ring consists of 5 torsional angles, $\nu_0$, $\nu_1$, $\nu_2$, $\nu_3$, $\nu_4$, and it is generally non-planar. Atoms displaced from the three- or four-atom plane on the same side as $C_{5'}$ are called *endo*, while those on the opposite side are called *exo*, giving the puckering modes its definition (see Figure 1.6.). Sugar pucker states are called after cardinal directions, where $C_{3'}$-*endo*-$C_{2'}$-*exo* is called **North**, $O_{4'}$-*endo* is called **East**, $C_{3'}$-*exo*-$C_{2'}$-*endo* is called **South** and $O_{4'}$-*exo* is called *West*. A more elegant representation of the five-membered ring can be done by the concept of **pseudorotation**, described in terms of the degree of pucker, $\nu_{max}$, and the pseudorotation phase angle, *P*. The standard formalism of pseudorational phase angle *P* was described by (Altona & Sundaralingam 1972) and it is calculated from the endocyclic sugar torsion angles according to

$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2 \cdot \nu_2 \cdot (sin36^\circ + sin72^\circ)} \qquad \text{Eq. 1.1}$$

where $P = 0^o$ is defined in such way that torsion angle $\nu_2$ is maximally positive. The maximum torsion angle $\nu_{max}$ describes the maximum out-of-plane pucker, given by:

$$\nu_{max} = \frac{\nu_2}{cos\ P} \qquad \text{Eq. 1.2}$$

Another fact used in determination of *P* is that the *sum of all sugar pucker torsions is equal to zero*. The theoretical determination of the change of the ring torsion angles as the angle of pseudorotation, and the energy profile associated, were first done by Levitt and Warshel (Levitt & Warshel 1978) (see Figure 1.6B).

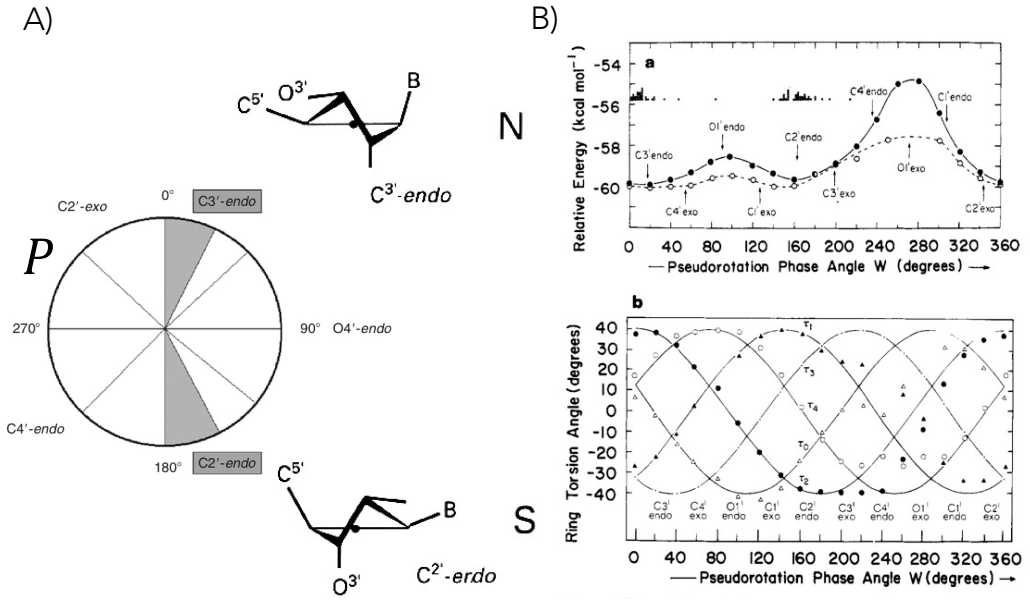



**Figure 1.6. Sugar puckering.** (A) sugar ring and puckers, (B) Pseudorotation phase angle (taken from (Levitt & Warshel 1978)).

In practice, nucleic acids display to main conformations of the sugars: *North* (predominant in RNAs) and *South* (predominant in DNAs), with *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 (Levitt & Warshel 1978).

## Helices

Double helical form of DNA yields several structural characteristics. Very obvious are the **grooves**, which are spaces between two strands. The asymmetry in the base pairs results in two parallel types of grooves, called *major* (MG) and *minor* (mG) *grooves*, whose dimensions are related to the distances of base pairs from the axis of the helix and their orientation with respect to the axis. Grooves are characterized by two parameters, *groove width*, defined as the perpendicular distance between phosphate groups on opposite strands in 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. These voids 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. A common example is binding of dye Hoechst 33258 to the minor groove (Harris et al. 2001) (see Figure 1.7).

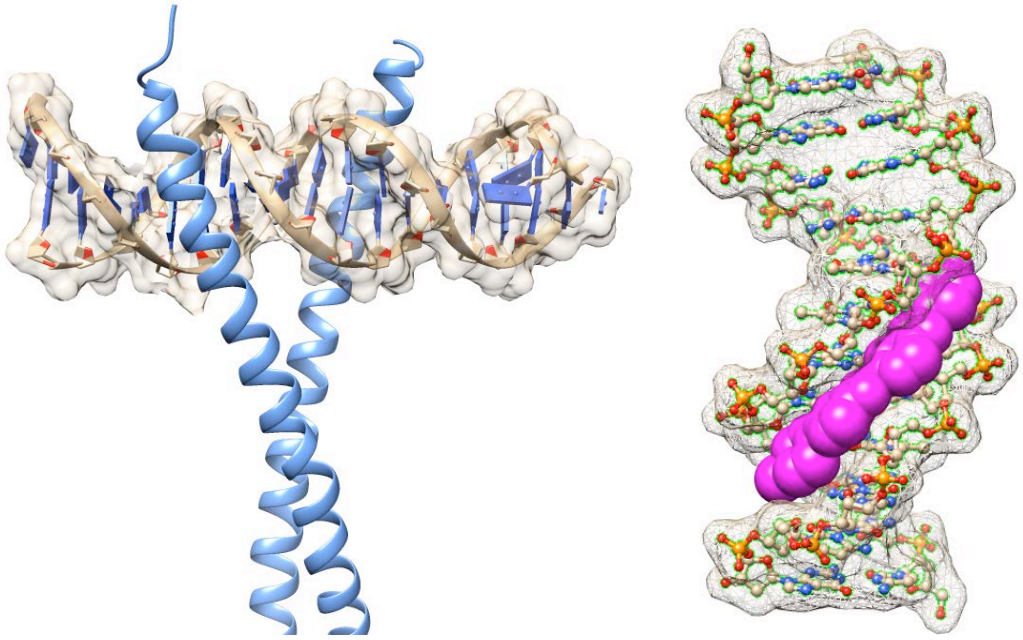


**Figure 1.7. DNA complexes.** DNA-protein complex with Leucine zipper bounded to the DNA major groove (left) [PDB:1YSA], and DNA-ligand complex with the Hoeschst stain dye 33258 (in magenta) bound to DNA minor groove (right) [PDB:264D].

Already original fiber diffraction experiments show that the DNA is a polymorphic structure. The most relevant one is what is called the B-form of DNA (or **B-DNA**), characterized by *right-handed spiral* and clear major and minor grooves (Arnott & Hukins 1972). The double helix of B-DNA has 10 base pairs per complete turn, with helical repeat of a single nucleotide unit with sugar pucker in $C_{2'}$-*endo* conformer. The two polynucleotide chains wound antiparallel to each other and are linked by Watson-Crick base pairing. The wide major groove of B-DNA is richer in base substituents ($O_6$, $N_6$ of purines and $O_4$, $N_4$ of pyrimidines) compared to the narrower minor groove, which has hydrophobic hydrogen atoms of the sugar groups forming its walls. B-DNA is the most common form of DNA molecules occurring at high relative humidity (~92%), but it is not the only form of DNA double-helix. The **A-DNA** is a wider right-handed spiral with a shallow, wide minor groove and a narrower, deeper major groove (Dickerson et al. 1982) (see Figure 1.8). Similarly to B-DNA form, A-DNA has the helical repeat of a single nucleotide unit, but in different sugar conformer. The $C_{3'}$-*endo* sugar pucker brings consecutive phosphate groups closer together, twisting and tilting the base pairs with respect to the helix axis, displacing them nearly 5 Å from it. As consequence, A-DNA helix is in striking contrast to the B-DNA one. The A form occurs under non-physiological conditions,

such as in a dehydrated environment, in complexes of DNA with certain proteins and in hybrids of DNA and RNA. Furthermore, the A-form is the most common form of RNA duplexes (see Chapter 1.3). DNA can also adopt the Z- form, (**Z-DNA**), in the alternating sequence poly(dC-dG)•poly(dC-dG) (Drew et al. 1980). The Z-form is characterized by a *left-handed* form with guanosines in the *syn* conformation, resulting in a "zig-zag" arrangement of phosphate groups (thus the term Z-DNA). This unusual structure requires unrealistically large ionic strength, but can be recognized by binding proteins and maybe involved in regulation of transcription (Oh et al. 2002). Other types include H-DNA (with Hoogsteen base-pairing instead of the common W-C one; (Abrescia et al. 2002); see below), parallel WC-DNA (Watson-Crick base-pairing with both strands propagating in the same direction; (Otto et al. 1991)) or C-DNA (occurring under relatively low-humidity conditions; (Brahms et al. 1973)) exist as minor forms.

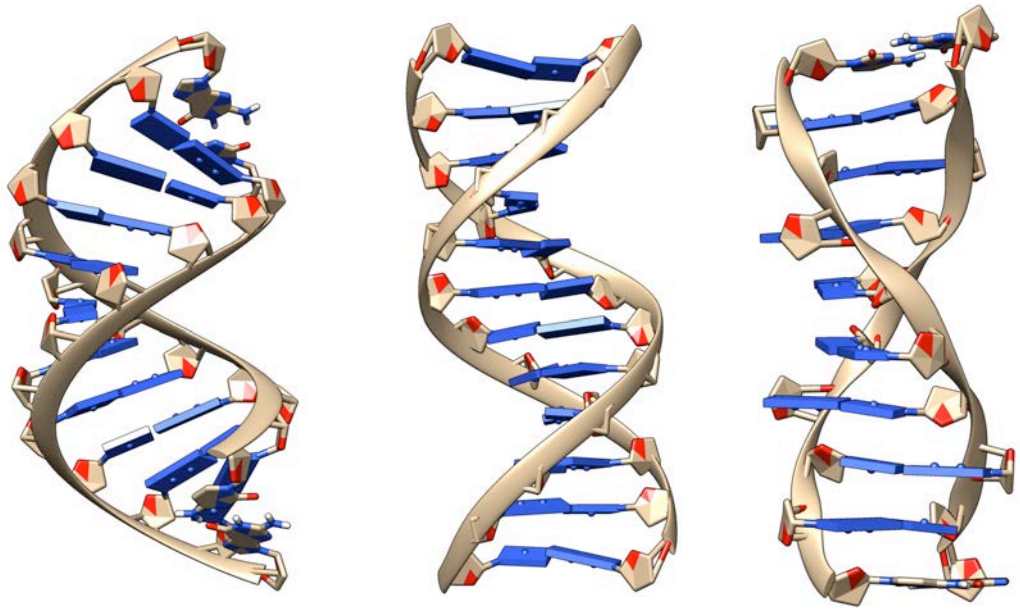


**Figure 1.8. Three major forms of DNA double-helix.** A-DNA (left), B-DNA (middle) and Z-DNA (right) form.

Putting it all together we can characterize the major forms of DNA by their structural features (see Table 2.1). For an overview of DNA conformations and their sequence preferences see (Neidle 2008; Svozil et al. 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: -49.3°<br>G: -10.3° |
| Roll | 0° | 0° | C: 5.6°<br>G: -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: anti<br>G: syn |
| Sugar pucker | C3′-endo | C2′-endo | C: C2′-endo<br>G: 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 1.1. Characteristics of the major DNA double helices.** Geometry attributes of the 3 conformations of DNA double-helix (Neidle 2008).

## Nonstandard and higher-order DNA structures

A part from double helical form, other tertiary structures of DNA of biological significance exist under physiological conditions. One of them is the **triple-stranded DNA**, often named a triplex, a structure in which one strand binds to the major groove of a B-form DNA double-helix and pairs through *Hoogsteen* or *reversed-Hoogsteen hydrogen bonding* to form a triple helix (see Figure 1.9). **Hoogsteen base pairing** is a variation on nucleic acids' base-pairing in which pairing applies the N7 position of the purine base and C6 amino group (see Figure 1.9). This orientation gives Hoogsteen pairing quite different properties and slightly lower stability than regular Watson-Crick pairing (Cubero et al. 2006). Hoogsteen base-pairing have been observed in protein-DNA complexes (Aishima et al. 2002), and might coexist as a minor species in regular duplexes in fast equilibrium with Watson-Crick pairings

(Nikolova et al. 2011; Cubero et al. 2006).[3] Hydrogen-bonding of the triplex-forming strand to the purine strand can occur in *parallel*, when both strands are parallel, or *anti-parallel*, when both strands are antiparallel to each other, manner. The convention for the triplet is X•YZ, where YZ represent the W-C hydrogen-bonded base-pair, with base X being in the third strand and hydrogen-bonded to base Y. Study on the parallel triplexes, poly(U•AU) and poly(T•AT), have shown some that they have some features of classical A-DNA (significant displacement of bases from the helix axis, ~ 3.5 Å, an average helical twist of ~ 30º, and almost identical minor-groove with of 10.7 Å), but with wider major-groove width (9.8 Å), which is necessary to accommodate the third strand (Chandrasekaran et al. 2000), but many others, for example puckering characteristics of a B-DNA (Shields et al. 1997). Antiparallel triplexes consist mostly of purine motives (Pu•PuPy), most readily formed with G•GC base triplets, together with pyridine-purine motive, namely T•AT base triplet (Radhakrishnan & Patel 1993). Antiparallel triplexes show higher stability than parallel ones (Soyfer 1996). Triplex-forming oligonucleotide (TFO) can target sequences on duplex DNA to form intermolecular triplex that (Campos et al. 2005). Triplex DNA can be used as gene-drugs in anti-gene strategy (Hélène 1991). Very intriguing, polypurine strands able to form triplexes are very abundant in the promoter region (Goñi et al. 2004), suggesting that they might have a role in gene regulation.

Another significant structure of the DNA *in vitro* and *in vivo* is **quadruplex DNA**. It is formed by *guanine-rich sequences* that form four-base units sets, stabilized with a monovalent ion in the middle, and stacked on top of each other. Quadruplex formation involves a total of eight hydrogen bonds between the W-C face of one guanine and the Hoogsteen major groove face of another. In case of four guanine bases forming the core, the structure is called G-tetrad, or G-quadruplex (see Figure 1.9) (Arnott et al. 1974; Wang & Patel 1993). The four guanosine nucleosides in an individual tetrad can in exist in either *anti* or *syn* conformation, and thus there are 16 possible combinations (Burge et al. 2006). The mutual orientation of individual strands in a quadruplex depends on the glycosidic angles. In an all-parallel orientation of all four strands, all guanosines have to adopt an *anti* conformation, while in antiparallel setup glycosidic angles be in both *syn* and *anti* conformation, as antiparallel quadruplexes can be formed from separate strands or from a single folded strand (like human telomere G-quadruplex in Figure 1.9). These structures play a significant role stabilizing the chromosome ends, called *telomeres*, and preventing them from being treated as damage to be corrected (Sun et al. 1997; Nugent & Lundblad 1998), allowing the cell to replicate chromosome ends using the enzyme called telomerase.

[3] Before Watson and Crick published their model of DNA double-helix, Karst Hoogsteeen reported a crystal structure of a complex in which analogues of A and T formed a base-pair with a different geometry, thus giving its name (Hoogsteen 1963).

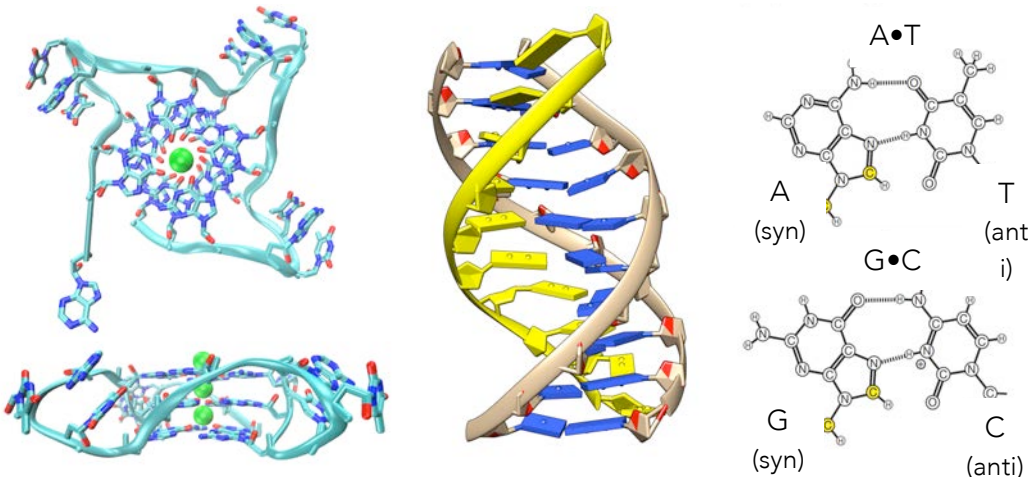



**Figure 1.9. Alternative DNA structures.** Structure of parallel Human talomeric G-quadruplex [PDB: 4G0F] top and side view (left), the structure of an anti-parallel triplex DNA with the Hoogsteen strand colored in yellow (middle), and the representation of Hoogsteen base-pairing (right).

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 interaction with the small proteins, called *histones*, leads to formation of *nucleosome*, which is compacted further more to give the known form of DNA called *chromosome* (Kornberg & Lorch 1999) (see Figure 1.10.).

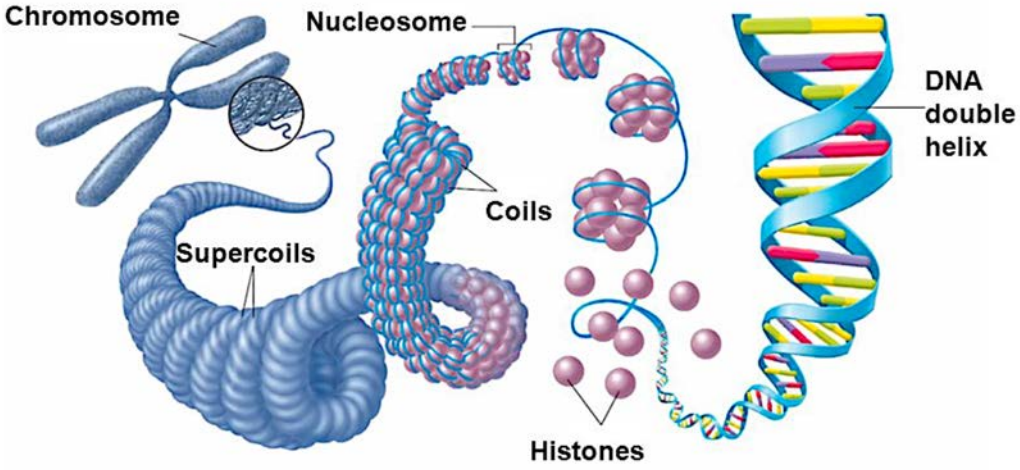



**Figure 1.10. Eukaryotic DNA compaction.** DNA quartenary structures, from double stranded linker (far right) to nucleosome and other compact structures, to the single chromosome (far left) [taken from IU South Bend Biochemistry webpage].

## 1.3. DNA's little brother – RNA

### The central dogma of molecular biology

Besides storing all genomic data, DNA has no functional ability. Yet, working indirectly, genomic sequences define instructions for metabolic processes; a guideline explained by *the central dogma of molecular biology* (Crick 1970) (Figure 1.11), which describes the normal flow of biological information where DNA can be copied into another DNA (called DNA replication) or copied into messenger-RNA (transcription), which can be used to synthesize proteins with help from transfer-RNA (translation) and the ribosomes (a supra-molecule rich in ribosomal RNAs). A third type of RNA, called ribozyme, has enzymatic activity for catalytic cleavage of RNA. Finally many others small and long non-coding RNAs with potential regulatory activities have been characterized in nuclei and cytoplasm. As seen from Figure 1.11, RNA has many biological roles in decoding, regulation and expression of genes. For that reason, many scientists believed that RNA is the central molecule in the molecular biology dogma.

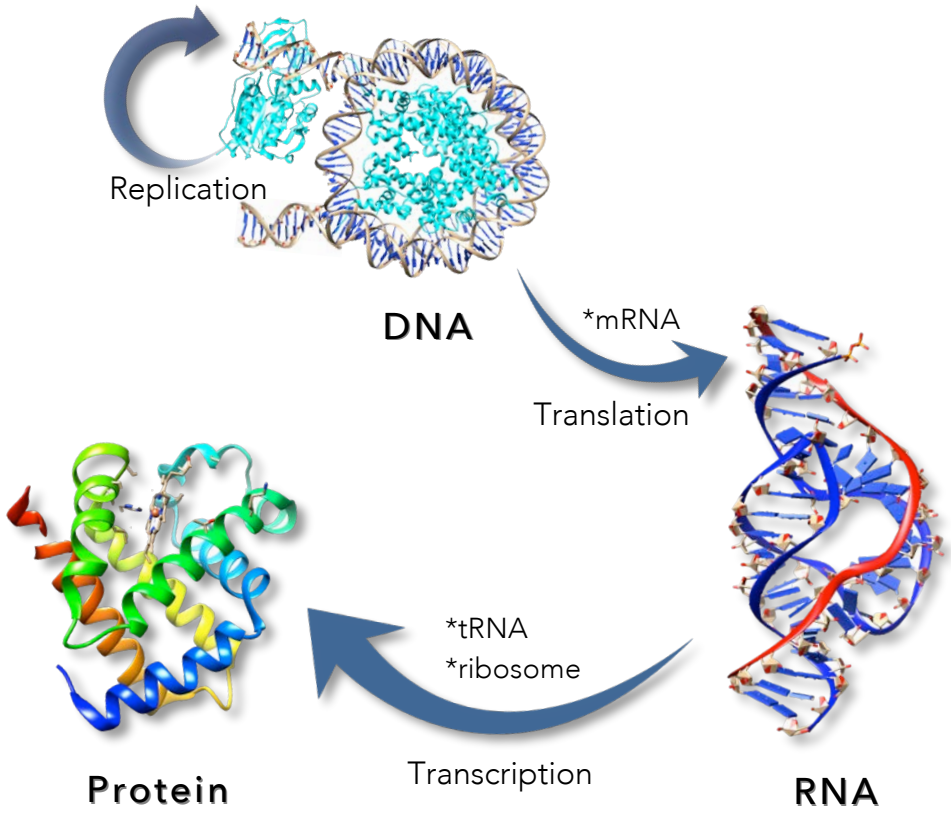



**Figure 1.11. Central dogma of molecular biology.** Information flow in biological systems.

## RNA vs. DNA

Although globally very similar to DNA, RNA has two major chemical differences. The major one is in the ribose sugar, where RNA has a hydroxyl group at the 2'-position (see Figure 1.12.). Conformational wise, 2'-hydroxyl group forces the ribose into C3'-endo sugar conformation, unlike C2'-endo conformation in DNA, making a RNA double-helix change from a B-form to a A-form. As an effect of the extra hydroxyl group, RNA sugars are much more rigid than the DNA ones. Moreover, 2'-hydroxyl group gives another degree of freedom to the conformation of the molecule as it is a strong hydrogen bond partner and it is known to interact with many different groups in the molecule and solution too (Auffinger & Westhof 1997). Overall, this small change in the ribose leads RNA molecules to have a much bigger variety of conformations, thus giving them more functionality than DNA (Neidle 2008).

Second difference from DNA is that RNA uses uracil base instead of thymine. Chemically, these two bases are very similar, differing only by a methyl group on C5, but with no effect on base pairing (see Figure 1.12). However, uracil is one product of hydroxylation of cytosine, making RNA more susceptible to mutations than DNA. Overall, the chemical properties of RNA make large RNA molecules *inherently fragile* and suboptimal for storing genetic information in complex organisms.

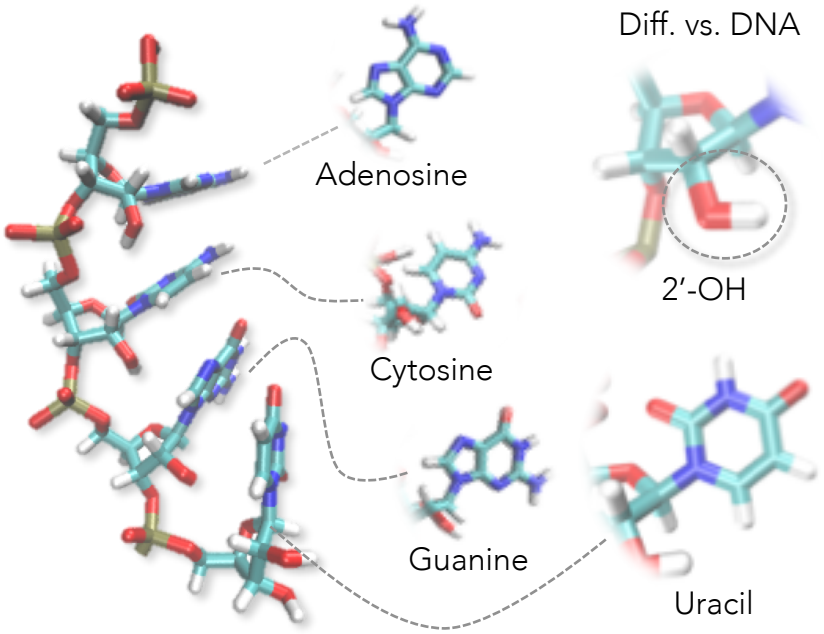



**Figure 1.12. Basics of RNA.** A RNA oligo with highlighted bases and differences from DNA (2'-OH group and Uracil).

Functionally speaking, RNAs can be divided into two groups, coding, also called messenger-RNA (mRNA), and non-coding, transfer-RNA (tRNA) and

ribosomal-RNA (rRNA). There are also non-coding RNAs involved in other roles like gene regulation and RNA processing, named interference RNAs, and others based on the nuclei, which might be involved in maintaining chromatin structure. Focusing in the three most abundant RNA, and summarizing a complex process we can say that mRNA carries the information to ribosome, which is made of rRNAs, that builds proteins using building blocks of tRNAs.

## RNA local structures

Unlike DNA, RNA is more often found in nature as single-strand folded onto itself, rather than in a duplex paired to a complementary strand. Single-stranded RNA can form many secondary structures by folding in complex motives stabilized by intramolecular hydrogen bonds between complementary bases and a myriad of stacking contacts. Most known secondary structures motifs of RNA are single and double stranded stems, and, more notably, extra-helical loops, namely **hairpin** and **bulge loops** (see Figure 1.13). Combinations of stems and loops form secondary structure like stem-loops, namely, *tri-*, *tetra-* and *hexa-loops*. An important example of complex RNA architectures is the tRNA, discovered by Robert Holley in 1965 (Apgar et al. 1965), which is defined as a combination of double stranded stems, a bulge loop and 3 hairpin loops (see Figure 1.13). The **Sarcin-Ricin domain** (SRD) motif forms a crucial site for the binding of elongation factors during protein synthesis and it is one of the longest and better conserved sequences found in the RNA of all large ribosomal subunits. SRD motif is a distorted hairpin consisting of a GAGA tetraloop, a G-bulged cross-strand A-stack, a flexible region and a terminal A-form duplex. It has been studies extensively both experimentally (Szewczak et al. 1993) and theoretically (Kruse et al. 2014). Finally, another biologically relevant RNA domain with well defined structure is HIV-1 **TAR RNA** which has a highly folded stem-bulge-loop structure, and serves as the binding site of the viral protein Tat, the trans-activator of the HIV-1 LTR. This very dynamic structure has been studied extensively in recent years (experimentally mainly by the group of Al-Hashimi (Merriman et al. 2016)).

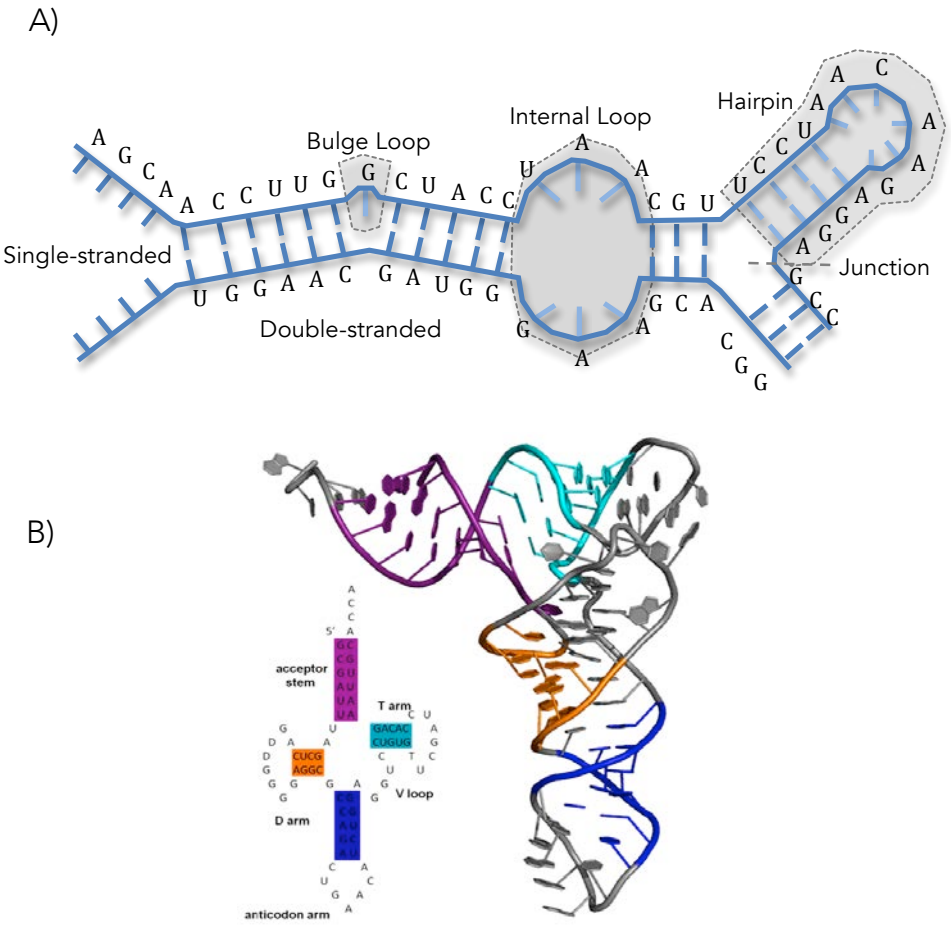



**Figure 1.13. Local structures of RNA.** (A) Representation of most common secondary structure motifs in RNA and (B) tRNA (double-stranded motifs in color and hairpins in grey) (right).

The functional form any RNA requires a specific tertiary structure, in which secondary structure elements are bonded within the molecule through short range interactions like hydrogen bonding or stacking. Recurrent structural local RNA motifs are kink-turn (**K-turn)** and **C-loop** (Lescoute et al. 2005). The K-turn is a common structural motif in RNA that introduces a tight kink into the helical axis. This internal loop occurs on the shallow/minor groove side and brings together the minor grooves of the two supporting helices (see Figure 1.14). It plays an important architectural role in RNA and serving as binding site for a number of proteins. C-loops, like K-turn are asymmetric internal loops with two bases in the longer strand form non-WC base pairs with bases in the shorter strand (see Figure 1.14). In the ribosome, C-loop are found in hairpin stem-loop structures that engage in tertiary interactions involving the hairpin loop. Another known RNA motifs are **pseudoknots** and **kissing loops** (see Figure 1.14). In pseudoknots motif a single stranded region of a hairpin loop base pairs with an upstream of a downstream sequences within the same RNA strand (like in Hepatits Delta virus ribozyme (Ferré-D'Amaré et al. 1998)), while kissing loops motif forms when the single-stranded loop regions of two hairpins interact through

base pairing. Kissing loops is one type of an Interacting loops motif, other being *ring RNA*. Finally, another small fold motif is the **three-way junction**, which can be categorized into three distinct topological families that depend on the size of the junction, and on the relative orientations of the three stems forming the arms of the junction. It plays key roles in the architecture of ribozymes and riboswitches (Lescoute et al. 2005).

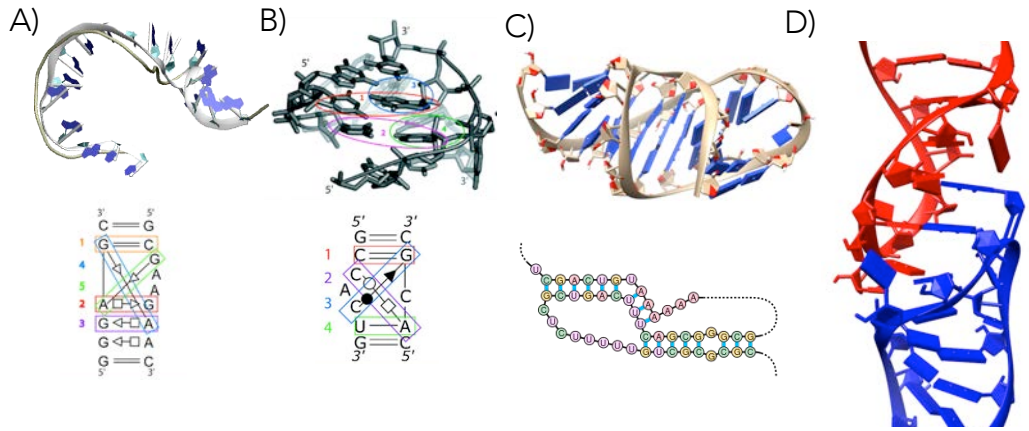



**Figure 1.14. RNA motifs.** (A) K-turn, (B) C-loop, (C) RNA pseudoknot [PDB:2N6Q], (D) RNA kissingloops with the two strands colored differently [PDB:2MI0]. (some taken from (Lescoute et al. 2005)).

## RNA architecture

Similarly to DNA, RNA can form double helices, triplex and quadruplex structures. Because of difference in sugar puckering conformation (see above), RNA **duplex** adopts more A-like form. This is characterized by right-handed helicity, slightly wider diameter (of 23 Å), deeper and narrower major groove, and steep and wide minor groove. RNA **triplexes**, unlike DNA triplexes, are often formed through minor groove binding. Most common example is the insertion of adenosine base into the minor groove, so called *A-minor motif* (see Figure 1.15). A minor groove triplex is present in *sarcin-ricin* motif, a highly conserved sequence found in the RNA of all large ribosomal subunits (Szewczak et al. 1993). Although rare, major groove RNA triplex interactions can be observed in several RNA structures. **Quadruple helices** in RNA can be formed in more ways than in DNA. Equivalently to DNA G-quadruplex, four consecutive guanine residues can form a quadruplex in RNA by Hoogsteen hydrogen bonds, a so-called “Hoogsteeen ring” (Cheong & Moore 1992) [PDB: 1RAU], but a different bonding pattern makes the core of malachite green aptamer (Baugh et al. 2000) [PDB: 1FIT] (see Figure 1.15). The unique structure of quadruplex regions in RNA may serve several biological functions like potential binding sites for ligands or proteins, regulators of gene expression (Oliver et al. 2000) or modulate transcription and replication (Arthanari & Bolton 2001).

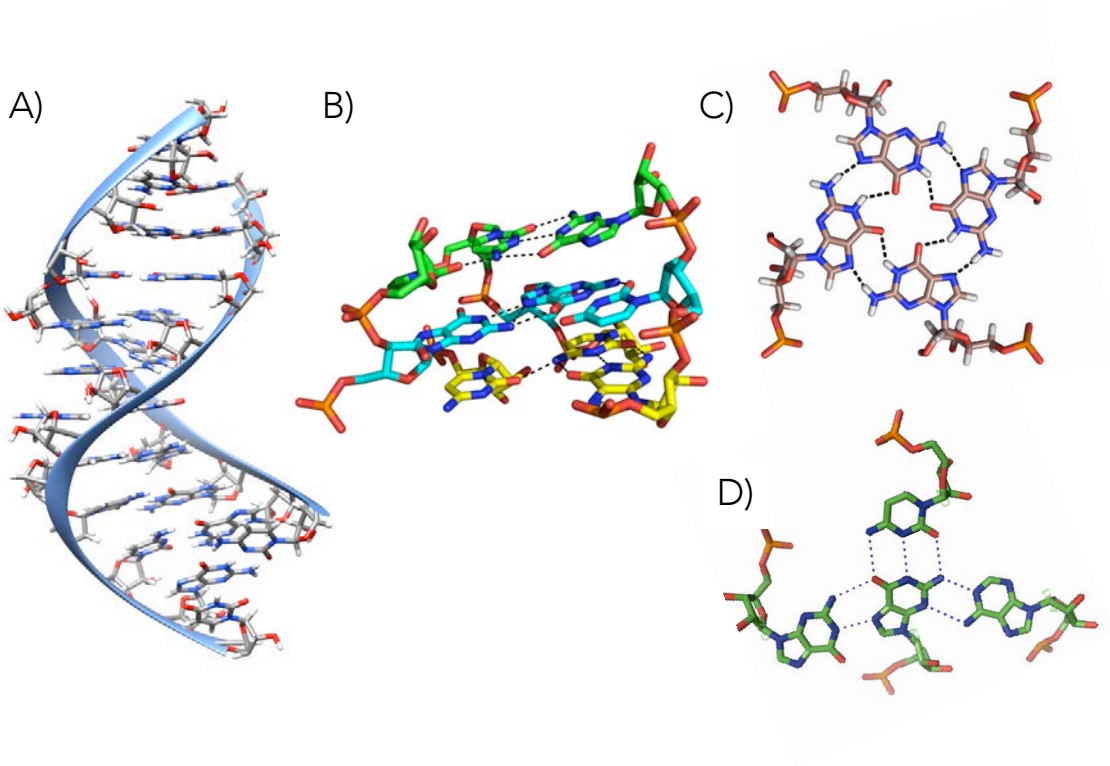



**Figure 1.15. Variety of RNA architecture.** (A) Double stranded RNA [PDB:2L8F], (B) Minor groove triplet, A-minor, [PDB ID: 5KSC], (C) RNA G-quartet [PDB ID: 1RAU], (D) Quadruplex seen in Malachite Green RNA aptamer [PDB ID: 1FIT].

## 1.4 Structural methods for the study of nucleic acids

### Experimental methods

Nucleic acids (NAs) are studied in multiple fields like biochemistry, genetics, microbiology, molecular biology, molecular evolution and structural biology with high importance. Just for experimental studies of RNA, 30 scientists have received a Nobel Prize. The core of all these fields lays in structural biology view of NAs. Structural biologists can study DNA and RNA in both experimental and theoretical ways. Likewise as already demonstrated by Watson and Crick experimental and theoretical studies are complementary and essential for better understanding of the structure of nucleic acids. Experimental structural studies of DNA are done in many different ways, but 2 most common atomistic-resolution techniques are X-ray crystallography and Nucleic magnetic resonance (NMR) spectroscopy. All the structures obtained so far with these two methods are stored in large databases; most popular being the Protein data bank (PDB; available online at www.rcsb.org).

**X-ray crystallography** determines the mean position of atoms and their chemical bonds based on the diffraction of a beam of X-ray light by a single crystal

(see Figure 1.16). Max Perutz (Perutz et al. 1960) (doctoral advisor of Francis Crick) was the first one to develop the X-ray crystallography method for biomolecules. The method consists of taking multiple two-dimensional images of the crystal at different rotations, which allows reconstructing the three-dimensional model of the electron density. Its drawbacks are the resolution dependence on the quality of the crystal, and that crystal packing forces can slightly change the native structure of the molecule, among other factors. On top of that, obtaining crystals of a given molecule (DNA or protein) is extremely challenging, especially for flexible systems.

Using X-ray structures for structural validation of a molecular dynamics simulation (see below) is not a completely justifiable way of doing the analysis since the experimental conditions don't perfectly agree with biological environment. Crystal packing can have an effect on the structure that can cause a deviation (Johansson et al. 2000), an artifact that can be spotted by visual inspection of the crystal structure (see Figure 1.16). A more accurate approach would be to simulate the structure in its crystal environment and then compare the results with experimental findings (see Figure 1.16). This technique has been used several times in the group of prof. David Case (Liu et al. 2014).

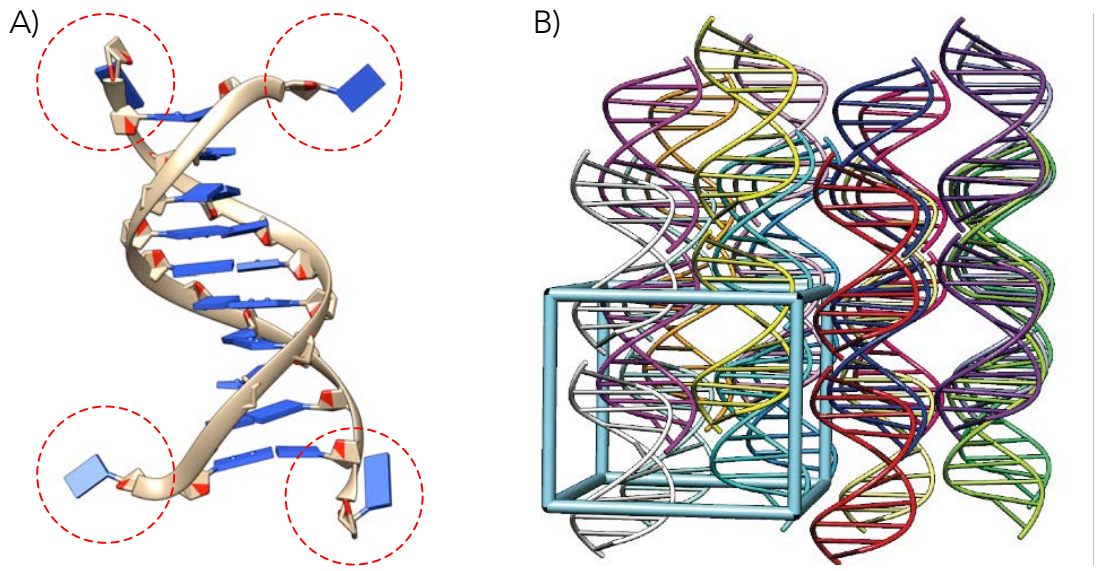



**Figure 1.16. Packing effects of X-ray structures.** (A) An example of packing effects on the structure of a Drew-Dickerson dodecamer [PDB ID: 1EHV] and (B) an illustration of a MD simulation unit of crystal. Figure taken from parmbsc1 publication (see Chapter 4.1).

Thanks to pioneers such as Kurt Wuthrich and technological advances in last 30 years, **NMR (nuclear magnetic resonance) spectroscopy** has gained much attention in structural determination of biomolecules (Palmer & Patel 2002). NMR method is based on exploiting the magnetic properties of certain atomic nuclei. Chemical environment of such nuclei gives rise to observables like the *Nuclear Overhauler*

*Effect* (**NOE**) and spin-spin coupling, mostly *J-coupling* (see Figure 1.16). NOEs occur through spatial cross-relaxation of nuclear spin polarization, quantitatively describing the proximity, or the distance, of protons. The strength of the NOE signal then depends only on the spatial proximity of protons and it can be used to originate distance restraints between atoms within 1.8 Å and 6 Å. Angle restraints instead can be obtained from J-coupling, arising from the interaction of different spin states through the chemical bonds of atoms linked by 2-3 covalent bonds. It contains information about bond distance and angles between active spin nucleic (usually 13C and 15N).

The information on the dynamics became accessible with the development of special NMR techniques like **residual dipolar coupling** (**RDC**), the dipolar coupling between two nuclei that depends on the distance between them and their angle relative to the external magnetic field, providing both structural information at the long distance and dynamical information on the time scales slower than a nanosecond. RDCs are obtained in special field-oriented NMR and provide spatially and temporally averaged information about the angle between the external magnetic field and a bond vector in a molecule. Rather than distance restraints (as NOEs) they can provide orientational constraints about the relative orientation of parts of the molecules, even when they lye far apart. However, two issues complicate the data interpretation: first, the definition of the ever-changing tensor that describes the alignment of a flexible molecule with respect to the laboratory field; and secondly, the decoding of the information packed into ensemble averages, which often requires the support of theoretical models to be transformed into atomic positions. RDCs have been proven as very useful technique specially in the RNA world for understanding, for example the dynamics of complex structural motives (Zhang et al. 2007).

Overall, NMR spectroscopy is very useful in providing detailed information about the structure, dynamics, reaction state, and environment of biomolecules, and has the additional advantage that the experiments are done *under physiological conditions*. The shortcomings of the technique are related to the size of the system to be studied and to a generally poor resolution in the detail than that obtained from X-Ray crystallography.

In addition to high resolution techniques other experimental approaches have been used to study nucleic acids, such as electron microscopy (Beer et al. 1966; Douglas et al. 2009) and ultra-resolution fluorescence (Weiss 1999; Lakowicz 2013), which have provided information, for example on the packing of chromatin in the nuclei (Solovei et al. 2002), or a variety of biophysical methods providing information on macroscopic properties of DNA and RNA structures, such as *small-angle X-ray scattering* (SAXS) technique (Lipfert & Doniach 2007), circular dichroism (Kypr et al. 2009) or hydrodynamic measures (Eimer & Pecora 1991), or on subtle details of the interactions, such as UV, RAMAN or Infrared spectroscopies

(Mergny et al. 1998; Thamann et al. 1981; Gabelica et al. 2008) .

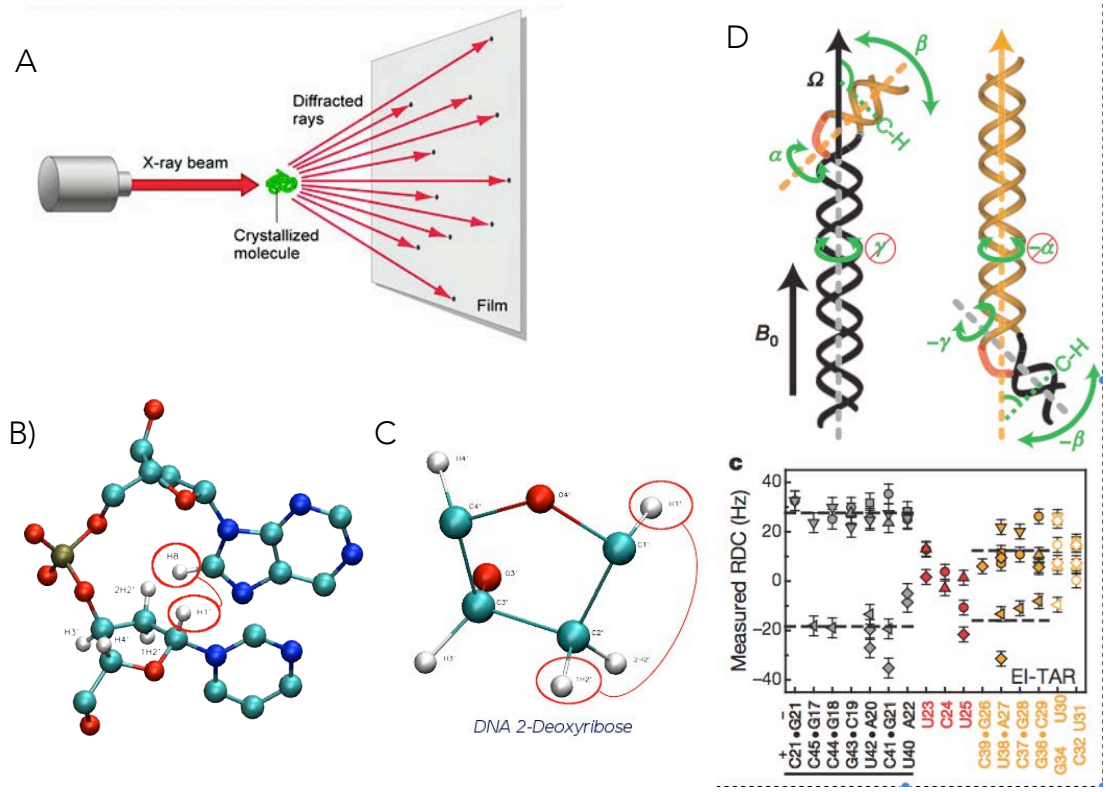



**Figure 1.17. Experimental techniques schemes.** (A) X-ray crystallography, (B) NOE, (C) J-coupling, (D) RDC (taken from (Zhang et al. 2007)).

## Theoretical modeling

Theoretical models are based on computational modeling of a given system using principles of physics, at different levels of simplifications, which allows us to expand the range of systems from small molecules to biomolecules and cells. All these modeling, also called *in-silico* models, are performed on highly sophisticated computers, called *supercomputers*. As computer power is limited, simplifications are essential to theoretical modeling in order to extend the magnitude of system size and simulation time. The focus of this thesis is precisely the development of computational models for a simplified representation of DNA. While an entire DNA molecule is immensely big (an extended human chromosome measure around 5 cm), its structural unit, a base-pair interaction, is happening in the $10^{-10}$ m scale. This multi-scale nature of DNA molecule makes it extremely challenging for any theoretical framework to properly study it. DNA is studied on multi-scale level ranging from atomic effect to nucleosome packing. For that reason theoretical framework includes several multi-level techniques moving from high-accurate methods able to deal only

with nucleobases to rough approximations able to manage entire chromosomes (Dans et al. 2016).

## Ab-initio methods

In principle, everything in Nature can be explained using Schrödinger's equation. Exact Schrödinger equation can be solved only for very small systems, so simplifications are needed for any system of interest. **Quantum mechanics** (QM) methods comprise then a variety of approaches of different accuracy and computer cost, but that provide the most accurate results on any chemical system. These methods are typically known as *ab initio* ('from first principle of quantum mechanics') and include, in practice, a big number of approaches aimed to simplify the complexity of the calculation. Most commonly they 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. An alternative to these computationally demanding methods are the DFT (density functional theory) approaches, which are based on the assumption that the spatial distribution of electrons at any point of space is sufficient to be able to describe any property of a system. QM methods are strictly necessary to study processes depending on the electronic structure, including catalytic, photophysical or spectroscopic properties that cannot be described with classical Molecular Mechancis (MM). In any case, even for the highest levels of simplification, QM methods are very costly, so they are limited to small model systems (around one nucleotide unit) and very short time scales. A closer look on QM methods used in this study will be given in the following chapter.

Big and complex in its conformational space, nucleic acids demand for methods capable of threating, in a dynamical way, up to several thousand atoms and allowing for simulations over time scales that cannot be achieved with any QM methods. In that sense, hybrid approaches, like *QM/MM* (Warshel & Levitt 1976), that combine the strengths of QM method for the chemically active region and the MM method for the surroundings, provide good alternatives to studies systems where the region requiring QM level of theory can be precisely localized, like in the case of the study of intra-strand oxidative crosslink lesions in DNA (Garrec et al. 2012) (see Figure 1.17). Pure QM and QM/MM methods are providing invaluable information on nucleic acids properties, impossible to obtain from other theoretical or experimental approach. For a better view on QM and QM/MM methods used in study of nucleic acids see (Sponer et al. 2008; Banáš et al. 2009).

## Classical approaches

Macromolecular systems consisting of hundreds to millions of atoms, like DNA and proteins, can be approximated to classical mechanics, which is less computationally demanding, by using the *"balls and strings" model*. This approach is based on representing atoms as deformable and charged balls of a given radius and chemical bonds as strings, described by Hook's law and integrated through Newton's laws of motion. This is the basis of molecular dynamics (**MD)** simulations. The core of MD simulations is the *force field*, a set of classical expressions that allows us to get a simple energy functional correlating the energy of a given configuration of the system with its energy. Within the force field approximation the electronic degrees of freedom are not specifically considered, but introduced parametrically in a simplified way. The accuracy of MD simulations is tightly connected with the development in nucleic acid force fields. Overall, with its various approaches, computational simulations have had an important impact on our understanding of structure and dynamics of nucleic acids. More details on MD simulations will be given in the following chapter.

More efficient approximations to represent large segments of nucleic acids imply the representation of sets of atoms as single particle, or grain, (thus the name *coarse-grain* (CG)). These methods loss then atomic resolution, but allow the study much bigger systems like long sequences of DNA consisting of thousands of base-pairs. For DNA, depending on the length scale of interest, quite different resolution in CG modeling can be applied. Usually, the number of grains per nucleotide varies from 8 to 3 grains, given the model. For a review on MD and CG methods used in study of nucleic acids see (Orozco et al. 2003; Cheatham & Case 2013; de Pablo 2011; Dans et al. 2016).

## Mesoscopic models

Given the DNA's unique properties (a long polymer made of repetitive base-pair building blocks that form a double-helix), an obvious choice of studying DNA would be through mathematical models. Several different mathematical models are based on each of DNA's characteristics. The *elastic rod model* (Landau & Lifshitz 1986), based on Kirchhoff elastic rod model, uses the fact that DNA is a long polymer which can be macroscopically represented by average properties. The model has been used in many studies involving long sequences of DNA, such as DNA loops (Balaeff et al. 1999), supercoiled DNAs (Bouchiat & Mezard 2000) or DNA mini-circles (Swigon et al. 1998). Another popular approach that describes the behavior of semi-flexible polymers is **worm-like chain model** (Bustamante et al. 1994; Marko & Siggia 1995), which is frequently used to characterize the average elastic properties of long sequences of DNA (Baumann et al. 1997). Other approaches have been developed to treat not ideal fibers, but real chromatin, introducing the nucleosomes as

specific particles in the model (Arya et al. 2006). This type of models have been used to determine the impact of linker histones or core histone modifications on the chromatin compaction (Collepardo-Guevara et al. 2015).

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 (Sim et al. 2012; Dans et al. 2016).

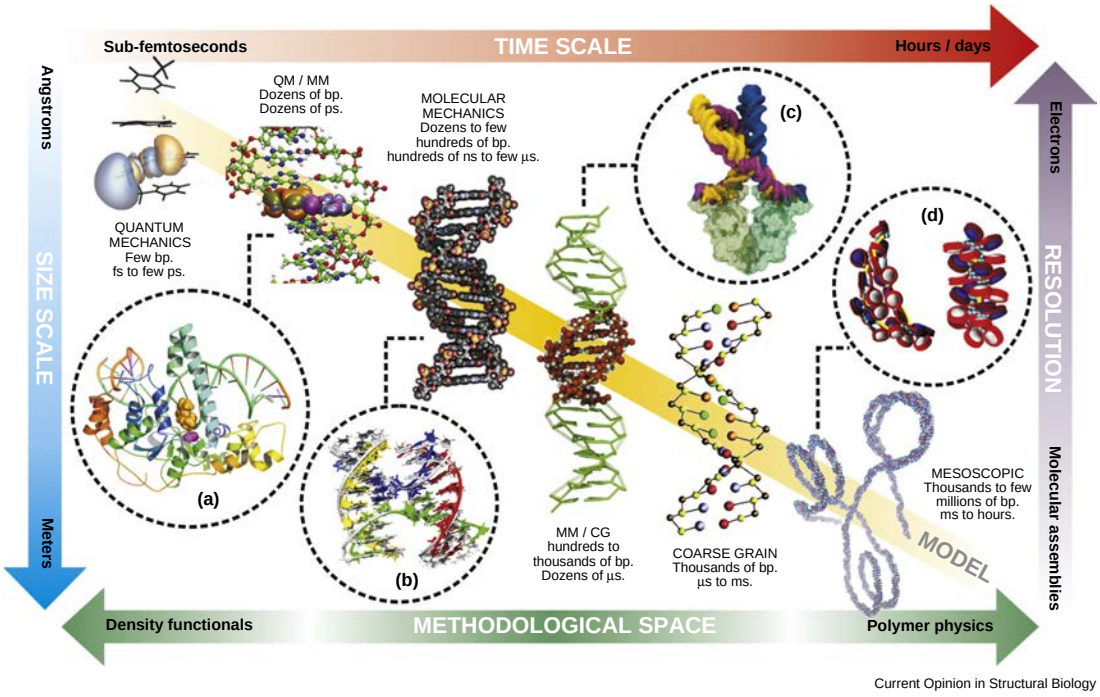



**Figure 1.17. Multi-scale techniques schemes.** Techniques used in studying DNA systems represented with its representative system size, time scale and resolution (taken from (Dans et al. 2016)).

# OBJECTIVES

The main aim of this thesis is to improve theoretical representation of DNA and RNA molecules in atomistic MD simulations by addressing the known caveats of current force fields. We divided this process in several steps with specific objectives:

1. **Reparameterization** of the current state-of-the-art DNA force field by fitting it to the high-level QM data, focusing mainly on parameterizing $\varepsilon/\zeta$ backbone torsions, sugar puckering and the glycosidic torsion, $\chi$.

2. **Validation** of the new force field. Testing it on big variety of DNA systems under various conditions, comparing the results with the experimental findings, as well as direct experimental observable like RDCs and NOEs.

3. Extension of the study to the **Drew-Dickerson sequence** testing the robustness of the new force field in respect to the choice of model of solvent and ions. Performing an extended simulation to check if the performance quality of the new force field remains, verify the convergence in longer timescales, and characterize the long-time dynamics of DNA.

4. The confirmation the lack of "overtraining artifacts" by **benchmarking** the new force field comparing its performance with other recent DNA force fields, in front of *de novo* obtained high quality NMR measures.

5. Understanding the conformational preference and **role of 2'-OH** in RNA conformation, and its potential role as a molecular switch controlling RNA-protein interactions.

6. Explore the possibility of current RNA modeling tools to design new RNA motifs, in particular a new **RNA dumbbell** structure, mimicking the behavior of synthetic siRNAs, but with higher resistance to degradation by nucleases.

# BIBLIOGRAPHY TO CHAPTER 1

Abrescia, N.G.A. et al., 2002. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proceedings of the National Academy of Sciences*, 99(5), pp.2806–2811.

Aishima, J. et al., 2002. A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Research*, 30(23), pp.5244–5252.

Altona, C. t & Sundaralingam, M., 1972. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *Journal of the American Chemical Society*, 94(23), pp.8205–8212.

Apgar, J., Everett, G.A. & Holley, R.W., 1965. Isolation of Large Oligonucleotide Fragments From the Alanine RNA. *Proceedings of the National Academy of Sciences*, 53(3), pp.546–548.

Arnott, S., Chandrasekaran, R. & Marttila, C.M., 1974. Structures for polyinosinic acid and polyguanylic acid. *Biochemical Journal*, 141(2), pp.537–543.

Arnott, S. & Hukins, D.W.L., 1972. Optimised parameters for A-DNA and B-DNA. *Biochemical and Biophysical Research Communications*, 47(6), pp.1504–1509.

Arthanari, H. & Bolton, P.H., 2001. Functional and dysfunctional roles of quadruplex DNA in cells. *Chemistry and Biology*, 8(3), pp.221–230.

Arya, G., Zhang, Q. & Schlick, T., 2006. Flexible histone tails in a new mesoscopic oligonucleosome model. *Biophysical journal*, 91(1), pp.133–50.

Auffinger, P. & Westhof, E., 1997. Rules governing the orientation of the 2'-hydroxyl group in RNA. *Journal of Molecular Biology*, 274(1), pp.54–63.

Balaeff, A., Mahadevan, L. & Schulten, K., 1999. Elastic Rod Model of a DNA Loop in the Lac Operon. *Physical Review Letters*, 83(23), p.4900.

Banáš, P. et al., 2009. Theoretical studies of RNA catalysis: Hybrid QM/MM methods and their comparison with MD and QM. *Methods*, 49(2), pp.202–216.

Baugh, C., Grate, D. & Wilson, C., 2000. 2.8 Å Crystal Structure of the Malachite Green Aptamer. *Journal of Molecular Biology*, 301(1), pp.117–128.

Baumann, C.G. et al., 1997. Ionic effects on the elasticity of single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 94(12), pp.6185–90.

Beer, M. et al., 1966. 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*, 5(7), pp.2283–2288.

Bouchiat, C. & Mezard, M., 2000. Elastic Rod Model of a Supercoiled DNA Molecule. *Eur. Phys. J. E*, 2, p.377.

Boyle, J., 2008. Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Biochemistry and Molecular Biology Education*, 36(4), pp.317–318.

Brahms, J. et al., 1973. Direct evidence of the C-like form of sodium deoxyribonucleate. *Proceedings of the National Academy of Sciences*, 70(12), pp.3352–3355.

Burge, S. et al., 2006. Quadruplex DNA: sequence, topology and structure. *Nucleic acids research*, 34(19), pp.5402–5415.

Bustamante, C. et al., 1994. Entropic elasticity of lambda-phage DNA. *Science (New York, N.Y.)*, 265(1994), pp.1599–1600.

Campos, J.L. et al., 2005. DNA coiled coils. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10), pp.3663–3666.
Chandrasekaran, R., Giacometti, A. & Arnott, S., 2000. Structure of Poly (dT)· Poly (dA)· Poly (dT). *Journal of Biomolecular Structure and Dynamics*, 17(6), pp.1011–1022.
Cheatham, T.E. & Case, D.A., 2013. Twenty-five years of nucleic acid simulations. *Biopolymers*, 99(12), pp.969–977.
Cheong, C. & Moore, P.B., 1992. Solution structure of an unusually stable RNA tetraplex containing G- and U-quartet structures. *Biochemistry*, 31(36), pp.8406–8414.
Collepardo-Guevara, R. et al., 2015. Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: A multiscale computational study. *Journal of the American Chemical Society*, 137(32), pp.10205–10215.
Crick, F., 1970. Central dogma of molecular biology. *Nature*, 227(5258), pp.561–563.
Cubero, E., Luque, F.J. & Orozco, M., 2006. Theoretical Study of the Hoogsteen–Watson-Crick Junctions in DNA. *Biophysical journal*, 90(3), pp.1000–1008.
Dans, P.D. et al., 2016. Multiscale simulation of DNA. *Current Opinion in Structural Biology*, 37, pp.29–45.
Dickerson, R.E. et al., 1982. The anatomy of a-, b-, and z-dna. *Science*, 216(4545), pp.475–485.
Douglas, S.M. et al., 2009. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature*, 459(7245), pp.414–418.
Drew, H. et al., 1980. High-salt d (CpGpCpG), a left-handed Z′ DNA double helix.
Eimer, W. & Pecora, R., 1991. Rotational and Translational Diffusion of Short Rodlike Molecules in Solution - Oligonucleotides. *J Chem Phys*, 94(3), pp.2324–2329.
Ferré-D'Amaré, a R., Zhou, K. & Doudna, J. a, 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395(6702), pp.567–574.
Gabelica, V. et al., 2008. Infrared signature of DNA G-quadruplexes in the gas phase. *Journal of the American Chemical Society*, 130(6), pp.1810–1811.
Garrec, J. et al., 2012. Insights into intrastrand cross-link lesions of DNA from QM/MM molecular dynamics simulations. *Journal of the American Chemical Society*, 134(4), pp.2111–2119.
Goñi, J.R., De La Cruz, X. & Orozco, M., 2004. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic acids research*, 32(1), pp.354–360.
Harris, S.A. et al., 2001. Cooperativity in drug-DNA recognition: a molecular dynamics study. *Journal of the American Chemical Society*, 123(50), pp.12658–12663.
Hashimoto, H. et al., 2012. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Research*, 40(11), pp.4841–4849.
Heddi, B. et al., 2006. Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *Journal of the American Chemical Society*, 128(28), pp.9170–9177.
Hélène, C., 1991. The anti-gene strategy: control of gene expression by triplex-forming-oligonucleotides. *Anti-cancer drug design*, 6(6), pp.569–584.
Hoogsteen, K., 1963. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta*

*Crystallographica*, 16(9), pp.907–916.
Johansson, E., Parkinson, G. & Neidle, S., 2000. A new crystal form for the dodecamer CGCGAATTCGCG: symmetry effects on sequence-dependent DNA structure. *Journal of molecular biology*, 300(3), pp.551–561.
Keshet, I., Lieman-Hurwitz, J. & Cedar, H., 1986. DNA methylation affects the formation of active chromatin. *Cell*, 44(4), pp.535–543.
Kornberg, R.D. & Lorch, Y., 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3), pp.285–294.
Kriaucionis, S. & Heintz, N., 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (New York, N.Y.)*, 324(5929), pp.929–30.
Kruse, H., Havrila, M. & Šponer, J., 2014. QM computations on complete nucleic acids building blocks: Analysis of the Sarcin-Ricin RNA motif using DFT-D3, HF-3c, PM6-D3H, and MM approaches. *Journal of Chemical Theory and Computation*, 10(6), pp.2615–2629.
Kypr, J. et al., 2009. Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Research*, 37(6), pp.1713–1725.
Lakowicz, J.R., 2013. *Principles of fluorescence spectroscopy*, Springer Science & Business Media.
Landau, L.D. & Lifshitz, E.M., 1986. *Theory of Elasticity 3rd edition*,
Lescoute, A. et al., 2005. Recurrent structural RNA motifs, Isostericity matrices and sequence alignments. *Nucleic Acids Research*, 33(8), pp.2395–2409.
Levitt, M. & Warshel, A., 1978. Extreme conformational flexibility of the furanose ring in DNA and RNA. *Journal of the American Chemical Society*, 100(9), pp.2607–2613.
Lipfert, J. & Doniach, S., 2007. Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annual review of biophysics and biomolecular structure*, 36, pp.307–27.
Liu, C., Janowski, P.A. & Case, D.A., 2014. All-atom crystal simulations of DNA and RNA duplexes. *Biochimica et Biophysica Acta (BBA)-General Subjects*.
Lu, X.J. & Olson, W.K., 2003. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17), pp.5108–5121.
Marko, J.F. & Siggia, E.D., 1995. Stretching DNA. *Macromolecules*, 28(26), pp.8759–8770.
Mergny, J.-L., Phan, A.-T. & Lacroix, L., 1998. Following G-quartet formation by UV-spectroscopy. *FEBS letters*, 435(1), pp.74–78.
Merriman, D.K. et al., 2016. Shortening the HIV-1 TAR RNA Bulge by a Single Nucleotide Preserves Motional Modes Over a Broad Range of Timescales. *Biochemistry*.
Neidle, S., 2008. *Principles of nucleic acid structure*, Amsterdam: Elsevier.
Nikolova, E.N. et al., 2011. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, 470(7335), pp.498–502.
Nugent, C.I. & Lundblad, V., 1998. The telomerase reverse transcriptase: components and regulation. *Genes & development*, 12(8), pp.1073–1085.
Oh, D.-B., Kim, Y.-G. & Rich, A., 2002. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proceedings of the National Academy of Sciences*, 99(26), pp.16666–16671.
Oliver, A.W. et al., 2000. Preferential binding of fd gene 5 protein to tetraplex nucleic

acid structures. *Journal of Molecular Biology*, 301(3), pp.575–84.
Orozco, M. et al., 2003. Theoretical methods for the simulation of nucleic acids. *Chemical Society Reviews*, 32(6), pp.350–364.
Otto, C. et al., 1991. The hydrogen-bonding structure in parallel-stranded duplex DNA is reverse Watson-Crick. *Biochemistry*, 30(12), pp.3062–3069.
de Pablo, J.J., 2011. Coarse-grained simulations of macromolecules: from DNA to nanocomposites. *Annual review of physical chemistry*, 62, pp.555–74.
Palmer, A.G. & Patel, D.J., 2002. Kurt Wuthrich and NMR of biological macromolecules. *Structure*, 10(12), pp.1603–1604.
Pérez, A. et al., 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11), pp.3817–3829.
Perutz, M.F. et al., 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*, 185(4711), pp.416–422.
Radhakrishnan, I. & Patel, D.J., 1993. Solution structure of a purine· purine· pyrimidine DNA triplex containing G· GC and T· AT triples. *Structure*, 1(2), pp.135–152.
Saenger, W., 1984. Principles of nucleic acid structure.
Shields, G.C., Laughton, C.A. & Orozco, M., 1997. Molecular Dynamics Simulations of the d (T⊙ A⊙ T) Triple Helix. *Journal of the American Chemical Society*, 119(32), pp.7463–7469.
Sim, A.Y.L., Minary, P. & Levitt, M., 2012. Modeling nucleic acids. *Current Opinion in Structural Biology*, 22(3), pp.273–278.
Solovei, I. et al., 2002. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Experimental cell research*, 276(1), pp.10–23.
Soyfer, V.N., 1996. *Triple-helical nucleic acids*, Springer Science & Business Media.
Sponer, J., Riley, K.E. & Hobza, P., 2008. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys Chem Chem Phys*, 10(19), pp.2595–2610.
Stofer, E., Chipot, C. & Lavery, R., 1999. Free energy calculations of watson-crick base pairing in aqueous solution. *Journal of the American Chemical Society*, 121(41), pp.9503–9508.
Sun, D. et al., 1997. Inhibition of human telomerase by a G-quadruplex-interactive compound. *Journal of medicinal chemistry*, 40(14), pp.2113–2116.
Svozil, D. et al., 2008. DNA conformations and their sequence preferences. *Nucleic Acids Research*, 36(11), pp.3690–3706.
Swigon, D., Coleman, B.D. & Tobias, I., 1998. The elastic rod model for DNA and its application to the tertiary structure of DNA minicircles in mononucleosomes. *Biophysical journal*, 74(5), pp.2515–30.
Szewczak, A.A. et al., 1993. The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 90(20), pp.9581–5.
Thamann, T.J. et al., 1981. The high salt form of poly (dG-dC)· poly (dG-dC) is left-handed Z-DNA: Raman spectra of crystals and solutions. *Nucleic acids research*, 9(20), pp.5443–5458.
Várnai, P. & Zakrzewska, K., 2004. DNA and its counterions: a molecular dynamics study. *Nucleic acids research*, 32(14), pp.4269–4280.
Wang, Y. & Patel, D.J., 1993. Solution structure of a parallel-stranded G-quadruplex DNA. *Journal of molecular biology*, 234(4), pp.1171–1183.

Warshel, A. & Levitt, M., 1976. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2), pp.227–249.

Watson, J. & Crick, F.H.F., 1953. Molecular structure of nucleic acids. *Nature.*, 171(4356), pp.737–8.

Watson, J.D. et al., 2003. Molecular biology of the gene. *Pearson Education, Inc.*, p.768.

Weiss, S., 1999. Fluorescence Spectroscopy of Single Biomolecules. *Science*, 283(5408), pp.1676–1683.

Wu, Z. et al., 2003. Overall structure and sugar dynamics of a DNA dodecamer from homo-and heteronuclear dipolar couplings and 31P chemical shift anisotropy. *Journal of biomolecular NMR*, 26(4), pp.297–315.

Zhang, Q. et al., 2007. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, 450(7173), pp.1263–1267.

*"The secret to modeling is not being perfect. You have to be given what's needed by nature, and what's needed is to bring something new."*
*Karl Lagerfeld*

# 2| MOLECULAR MODELING

Just as my grandmother uses different glasses for reading recipes, watching TV shows or talking with her friends, choosing a molecular modeling method depends on the interest and the size of the system. For small systems, up to 50-100 atoms, in which we are mainly interested in finding the correct geometry and properties derived from the electron distribution, QM calculations are the clear choice. Bigger systems consisting of up to a million atoms, with no changes in electron density, are ideal for classical MD simulations. QM/MM methods are reserved for large systems where the change in electron density can be located in a small section of the entire system, which can be then treated quantum mechanically (QM), while the rest is treated classically (MM).

In this chapter, I will introduce basic concepts of QM and MM that are required for a correct understanding of the thesis. For more detailed view on QM and MD methods reader is addressed to references (Szabo & Ostlund 1996; Cramer 2004; Monticelli & Salonen 2013; Rapaport 2004).

# 2.1 Quantum chemistry

According to quantum mechanics, the exact description of any physical system is represented by Schrödinger's equation. Thus, quantum chemistry is a branch of chemistry that focuses on the application of quantum mechanics, or solving Schrödinger's equation, for chemical systems. In principle, the Schrödinger equation is able to represent the *time-dependent* evolution of a QM system. However, for computational reasons, most times we use a *time-independent* equation, that is useful for the characterization of stationary states. In practice the exact solution of even the time-independent Schrödinger equation is limited to very small systems, very far away from those of interest in biology. Luckily several approximations to the general Schrödinger formalism can be used to reduce the computational cost of the QM calculations.

## Basic QM formalisms

The **Born-Oppenheimer approximation (BO)** assumes that the motion of atomic nuclei and electrons in a molecule can be separated. It exploits the fact that atomic nuclei are much heavier than electrons and move then very slowly compared to the electrons. This allows the Schrödinger's equation to be separated into two parts, the *electrons in the field of fixed nuclei* and *nuclei in the field of effective electron cloud*.

The Hamiltonian of a quantum system consisting of nuclei (denominates with capital latters) and electrons (denominated with lowercase letters) can be written as

$$\mathcal{H} = -\sum_i \frac{1}{2}\nabla_i^2 - \sum_A \frac{1}{2M_A}\nabla_A^2 - \sum_i \sum_A \frac{Z_A}{r_{i,A}} + \sum_i \sum_j \frac{1}{r_{i,j}} + \sum_A \sum_B \frac{Z_A Z_B}{r_{A,B}}$$

Eq. 2.1.

where $M_A$ is the ration of the mass of nucleus *A* to the mass of an electron, and $Z_A$ is the atomic number of nucleus *A.* The first two terms in Eq.2.1 are the operators for the kinetic energy of the electrons and the nuclei, respectively, while the last three terms are coulomb attraction between the electrons and nuclei, the repulsion between electrons and the repulsion between nuclei, respectively. Within BO approximation, the kinetic energy of the nuclei can be neglected and the repulsion between the nuclei can be considered to be constant, naming the remaining terms from Eq.2.1, the electronic Hamiltonian. The wave function that solves the electronic Hamiltonian

$$\Phi = \Phi_{elec}(\{\vec{r_i}\}; \{\vec{r_A}\}) \qquad \text{Eq. 2.2.}$$

describes the motion of the electrons and explicitly depends on the electronic coordinates but depends parametrically on the nuclear coordinates.

The BO approximation is fundamental to quantum chemistry, and is present in most of the QM methods used today.

The **Hartree-Fock approximation (HF)**, one of the most commonly used QM methods, is an approximation for the determination of the wave function and the energy in a stationary system. It assumes that the exact *N*-body wave function of the system can be approximated as a collection of *N spin-orbitals* $\chi$ (the electron wave function describing both its spatial distribution and its spin), also know as *Slater determinant* $|\Psi_0\rangle$.

$$|\Psi_0\rangle = |\chi_1\chi_2 \cdots \chi_a\chi_b \cdots \chi_N\rangle \qquad \text{Eq. 2.3.}$$

The best wave function of this functional form is the one that gives the lowest possible energy

$$E_0 = \langle\Psi_0|\mathcal{H}|\Psi_0\rangle \qquad \text{Eq. 2.4.}$$

where $\mathcal{H}$ is the full electronic Hamiltonian. By minimizing $E_0$ with respect to the choice of spin orbitals, one derives an eigenvalue equation, which determines the optimal spin orbitals. This equation, also known as the Hartree-Fock equation, is of the form

$$f(i)\chi(\vec{x_i}) = \varepsilon\chi(\vec{x_i}) \qquad \text{Eq. 2.5.}$$

where $f(i)$ is an effective one-electron operator, called the *Fock* operator.

$$f(i) = -\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{r_{i,A}} + \upsilon^{HF}(i) \qquad \text{Eq. 2.6.}$$

where $\upsilon^{HF}(i)$ is the average potential experienced by the *i*-th electron due to the presence of the other electrons, making the method *non-linear* that must be solved iteratively. The solutions to the non-linear Hartree–Fock equations is solved by the **self-consistent-field (SCF)** method, in which from an initial guess at the spin orbitals, one can calculate the average field electron and then solve the eigenvalue equation for a new set of spin orbitals. Using these new spin orbitals, one can repeat the procedure until self-consistency is reached (when the initial guess matches the calculated spin orbitals within a certain threshold). The solution of the Hartree-Fock eigenvalue problem

$$E_0 = \sum_a \langle a|h|a\rangle + \frac{1}{2}\sum_{a,b}[aa|bb] - [ab|ba] \qquad \text{Eq. 2.7.}$$

yields a set of orthonormal ($\langle\chi_a|\chi_b\rangle = \delta_{ab}$) Hatree-Fock spin orbitals $\{\chi_a\}$ with orbital energies, $(\varepsilon_a \ldots)$. The Hartree-Fock equation consists of classical *coulomb* term (second term in Eq.2.7) and non-classical *exchange* term (third term in Eq.2.7). As the Slater determinant is asymmetrical in respect of the exchange of the spin-orbitals (to fulfill the Pauli exclusion principle), the exchange operator acts on any identical spin-orbitals and it does not have any classical interpretation.

As an effect of mean-field approximation, electron correlation (around 1% of the overall energy) is neglected in HF approximation, which represents a problem in certain type of studies. This imprecision can be treated by many post-HF methods. Most popular is **Møller-Plesset perturbation theory (MP)** (sometimes called Rayleigh-Schrödinger), which can be applied at several orders of perturbation, the most popular one being the second one (MP2). This approach partitions the total Hamiltonian of the system into two pieces: a zeroth-order part, $H_0$, and a perturbation, $V$.

$$\mathcal{H}|\Phi_i\rangle = (\mathcal{H}_0 + V)|\Phi_i\rangle = \mathcal{E}_i|\Phi_i\rangle \qquad \text{Eq. 2.8.}$$

where we know the eigenfunctions and eigenvalues of $\mathcal{H}_0$.

$$\mathcal{H}_0\left|\Psi_i^0\right\rangle = E_i^0\left|\Psi_i^0\right\rangle \qquad \text{Eq. 2.9.}$$

If we would want to expend the perturbation in a way in which systematically improved eigenfunctions and eigenvalues of $\mathcal{H}_0$ would become closer and closer to the eigenvalues and eigenfunction of the total Hamiltonian, $\mathcal{H}$, we can introduce an ordering parameter $\lambda$ (which we will later set to unity),

$$\mathcal{H} = \mathcal{H}_0 + \lambda V \qquad \text{Eq. 2.10.}$$

and expend the exact eigenfunction and eigenvalues in a Taylor series in $\lambda$. Eventually, it can be shown that

$$\mathcal{E}_1 = E_1^{(0)} + \lambda E_1^{(1)} + \lambda^2 E_1^{(2)} + \lambda^3 E_1^{(3)} + \cdots \qquad \text{Eq. 2.11.}$$

where

$$E_1^{(1)} = V_{11}$$

$$E_1^{(2)} = \frac{V_{12}V_{21}}{E_1^{(0)} - E_2^{(0)}}$$

$$E_1^{(3)} = \frac{V_{12}V_{22}V_{21}}{\left(E_1^{(0)} - E_2^{(0)}\right)^2} - \frac{V_{12}V_{11}V_{21}}{\left(E_1^{(0)} - E_2^{(0)}\right)^2}$$

where $E_1^{(1)}$ is the *first-*, $E_1^{(2)}$ *second-*, and $E_1^{(3)}$ *third-order energies*, and $V_{12} = \langle 1|V|2\rangle$ is the correlation between states $|1\rangle$ and $|2\rangle$, where $|1\rangle$ is the state with lower eigenvalue.

Another popular post-HF method is the **Coupled-cluster (CC)** one, known for producing highly accurate results in comparison to experiments (Kümmel 2003). Developed by Čížek and Paldus (Čížek 1966; Cizek & Paldus 1980), CC method addresses the *size-consistency* problem that truncated *Configurational Interaction* (CI) method has (see (Szabo & Ostlund 1996)), for which the larger the system, the smaller the fraction of correlation energy. Instead of linearly expending the multi-

electron wavefunction from Slater determinant (obtained from HF method), like in the case of CI, CC method expends the wavefunction exponentially,

$$|\Psi_{CC}\rangle = \mathrm{e}^{\mathbf{T}}|\Psi_{HF}\rangle \qquad \text{Eq. 2.12.}$$

using exponential cluster operator, **T**, defined as

$$\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \cdots + \mathbf{T}_n \qquad \text{Eq. 2.13.}$$

where *n* is the total number of electrons. $\mathbf{T}_i$ operators generate a linear combination of all possible determinants, having *i* excitations from the reference. For example,

$$\mathbf{T}_2 = \sum_{i<j}^{\boldsymbol{occ.}} \sum_{a<b}^{\boldsymbol{vir.}} t_{ij}^{ab} |\Psi_{ij}^{ab}\rangle \qquad \text{Eq. 2.14}$$

CC is typically defined by the highest number of excitations allowed, from which *CCSD(T)* is most commonly used (includes single, double and triple excitations, the later ones as a perturbative approximation). CC energies are obtained by expanding Eq.2.12. into a Taylor expansion of the exponential function and computing

$$E_{CC} = \langle \Psi_{HF}|\mathcal{H}|\mathrm{e}^{\mathbf{T}}\Psi_{HF}\rangle \qquad \text{Eq. 2.15.}$$

For its high accuracy, CCSD(T) is considered *the "gold-standard" of QM calculation* for closed-shell systems, but the accuracy is obtained at a cost of high computational demand, which limits its applicability to relatively small systems.

On the contrary to HF methods, a computationally less demanding method, **Density functional theory (DFT)** determines the properties of the many-body system with *N* electrons using the functional of the spatially dependent electron density, avoiding the resolution of the Schrödinger equation, and assuming that the energy of a system (for a nuclear configuration) can be expressed as a function of the electron density. DFT methods reduce the problem from 3*N* spatial coordinates to 3 spatial coordinates, but have difficulties in expressing the exchange part of the energy (last term in the Eq. 2.16)

$$E_{DFT}[\rho] = T[\rho] + E_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \qquad \text{Eq. 2.16.}$$

where *T* is the kinetic energy of the electrons, $E_{ne}$ is the nuclear-electron attraction energy, *J* is the electron-electron repulsive energy, and $E_{xc}$ is the electron-electron exchange-correlation energy; all of which are the function of the electron density, $\rho$, which itself is a function of the three positional coordinates, making each term in Eq. 2.16 a functional. The difficulty of expressing the exchange part of the energy can be relieved by including a component of the exact exchange energy calculated from Hartree–Fock theory. Functionals of this type are known as *hybrid functionals*. One of the most commonly used versions of these hybrid functionals is B3LYP (stands for

Becke, 3-parameter, Lee-Yang-Parr) (Becke 1988; Lee et al. 1988)[1]. The hybrid form of the exchange functional of B3LYP is expressed as following

$$E_{xc} = (1 - a_0)E_x^{LDA} + a_0E_x^{HF} + a_xE_x^{B88} + a_cE_c^{LYP88} + (1 - a_c)E_c^{VWN80}$$

Eq. 2.17.

where $a_0 = 0.2$, $a_x = 0.72$, $a_c = 0.8$, $E_x$ is the Becke-1988 exchange functional, and $E_c$ is the correlation functional from Lee-Yang-Parr for B3LYP. Note that various approximations – local density approximation (*LDA*), Hartree-Fock (*HF*), Becke-1988 (*B88*) (Becke 1988), Lee-Yang-Parr-1988 (*LYP88*) (Lee et al. 1988), and Vosko, Wilks, Nusair 1980 (*VWN80*) (Vosko et al. 1980) are part of this hybrid functional. The three parameters defining B3LYP have been taken without modification from Becke's original fitting of the analogous B3PW91 function (Becke 1993). Since B3LYP, many other hybrid functions have been developed and tuned to deal with specific systems, or interactions (Zhao & Truhlar 2008; Hobza et al. 1995), but B3LYP still remains the 'industry standard'.

In general, DFT provides quality results with fast computation, but in general it is not as accurate in describing the energetics as MP2 calculations is, and is far from the performance of the "gold standard" CCSD(T). Thus, DFT functionals like B3LYP, are used as good initial methods in structure optimization, which can be later used as a initial guess for higher-level calculations.

## Basis sets

QM methods are always coupled with a specific **basis set**, a set of mathematical functions from which the wave function is constructed. For example, each molecular orbital (MO) in HF theory is expressed as a linear combination of basis functions, the coefficients for which are determined from the iterative solution of the HF SCF equations. The full HF wave function is expressed as a Slater determinant formed from the individual occupied MOs. The HF limit (the best results that can be obtained at the HF level) is achieved by the use of an infinite basis set, which necessarily permits an optimal description of the electron probability density. In practice, however, one cannot make use of infinite basis set. Much work has been done into identifying mathematical functions that allow wave functions to approach the HF limit arbitrarily closely in as efficient a manner as possible, meaning keeping the total number of basis functions to a minimum, permitting various integrals, and having a large amplitude or small amplitude depending on the probability density.

A good description of atomic orbitals (AO) is achieved using Slater-type orbitals (STOs), the basis functions used in an semiempirical method called extended

[1] Becke's original paper is one the most cited papers of all time, indicating how popular and well established in literature the B3LYP method is.

Hückel theory. Even though, having a number of advantages (like proper radial shape), STOs suffer a fairly significant limitation, as there is no analytical solution available for the general four-index integral (like the one being solved from Eq.2.7). Boys proposed an alternative to STOs, approximating it as linear combinations of Gaussian functions (Boys 1950). The STO-3G basis set is what is known as a 'single-ζ' basis set, where there is only one basis function defined for each type of orbital core through valence. One way to increase the flexibility of a basis set is to 'decontract' it, constructing two basis functions for each AO where the second would be normalized third primitive. A basis set with two functions for each AO is called 'double-ζ' basis set. Coming from the fact that the valence orbitals carry more importance in chemical bonding, **'split-valence'** basis sets represent core orbitals by a single (contracted) basis function, while splitting valence orbitals into arbitrary many function. Most notable used split-valence basis sets are those of Pople *et al.*, which include 3-21G, 6-31G and 6-311G (Krishnan et al. 1980). Split-valence basis sets, which have single primitives used in all contracted basis functions, but with different coefficients, are called **'correlation-consistent'** basis sets. Examples are the cc-pV*n*Z (cc-pVDZ, cc-pVTZ, etc.) sets, where the acronym stands for 'correlation-consistent polarized Valence (Double/Triple/etc.) Zeta' (Woon & Dunning Jr 1993). Correlation-consistent means that the exponents and contraction coefficients were variationally optimized not only for HF calculations, but also for calculations including electron correlation.

**Polarization functions** are one of the most common additions to the minimal basis sets, as MOs require more mathematical flexibility than do the AO. Polarization functions make use of higher quantum number functions that in principle required, but are absolutely needed in order to make reasonable geometry predictions of molecules that include such atoms, like phosphates. In split-valence basis sets, '*' implies a set of d functions added to polarize the p functions (second '*' implies light atoms as well), but nowadays the standard notation for Pople basis sets typically includes an explicit enumeration of those functions instead of the star nomenclature. Dunning's type of basis set have the polarization already included (small p in the name) in form of d functions on heavy atoms and p functions on H.

Significant errors in energies are produced when a basis set does not have the flexibility necessary to allow a weakly bond electron to localize far from the remaining density, like is the case of anions. To address this, standard basis sets are 'augmented' with **diffuse functions**. In the Pople family of basis sets diffusion is annotated with '+' sign for heavy atoms ('++' for H as well), while for Dunning family by prefixing with 'aug', which diffuses f, d, p, and s functions on heavy atoms and diffuse d, p, and s functions on H.

By definition the HF limit cannot be achieved, but a reasonable approach can be obtaine by using **complete basis set** (**CBS**) calculations. The cc-pV*n*Z basis sets were

specifically designed for this purpose, as for consistent increase in *n* would allow the extrapolation to infinite, giving the CBS value. Two most common extrapolation schemes, based on extrapolating to the infinite the results obtained with the increasingly large basis set, were proposed by Halkier and co-workers (Halkier et al. 1998; Halkier et al. 1999), and Truhlar (Truhlar 1998).

## Solvent effects

As biological systems are rarely found in vacuum, including solvent corrections into QM calculations is of high interest for biomolecular systems. A common approach is to model the solvent as a polarizable continuum, also known as **polarizable continuum model** (PCM). The molecular free energy of solvation is computed as the sum of *electrostatic*, *cavitation* and *dispersion-repulsion* (van der Waals) terms. Two types of PCM models have been popularly used nowadays, with dielectric-like (or polarizable; D-PCM) and with conductor-like continuum (C-PCM). The most popular C-PCM is COSMO solvation model (Cossi et al. 2003). In this thesis we have made extensive use of the PCM solvation model from Miertus, Scrocco and Tomasi (**MST**), refined in the group (Miertuš et al. 1981; Miertus & Tomasi 1982; Cances et al. 1997; Bachs et al. 1994). PCM is a *self-consistent reaction field* method (**SCRF**) method in which the solute resides in a cavity carved into a continuum polarizable medium that simulates the solvent. Free energy terms of MST model are defined as following: electrostatic term is determined by using a set of imaginary charges spread over the cavity surface and generated as a reaction to the solute's charge distribution. Cavitation free energy is computed from an atom-scaled version of Pierotti's formalism, and finally, the van der Waals term is determined as the sum of products of atomic surface tension of an atom and SESA. The average error of the MST free energy of solvation of neutral molecules in the variety of solvents is below 1 kcal/mol (Klamt et al. 2009). Different implementation of the PCM model have been incorporated in computer programs such as Gaussian (Marenich et al. 2009) (see Section 3.1.).

The main use of QM methods in this thesis was to determine conformational energy profiles that can be used as reference for the parameterization of classical force fields. We created then QM **energy scans** along given coordinates, typically torsion angles (see below). In that purpose, geometries of a set of structures along the coordinate would need to be optimized, sequentially calculating the energy. With QM methods varying in accuracy and speed, the choice of the method for each step is crucial for such a task. A reasonable approach for geometry optimizations is the B3LYP method as its functionals perform well in predicting minimum energy structures, especially for organic molecules (Tirado-Rives & Jorgensen 2008). The accuracies in bond angles averages about 1º, the same as is found for more computationally demanding MP2 method (Cramer 2004). 6-31++G(d,p) basis set provides accurate geometry predictions, while for bigger basis set very small

improvements are expected. Moreover, its fast computation makes B3LYP more suitable method for such a demanding task, as geometry optimization is. On the energetics perspective, even though, B3LYP does perform relatively accuracy decrease with increasing system size and underestimation of weak interactions. Thus, it is necessary to perform energy calculations with some post-HF method, namely MP2 or CCSD(T), in an sufficiently big basis set. From our experience, MP2 method with a Dunning type 'double-ζ' basis set, preferably augmented (MP2/aug-cc-pVDZ), produces precise results in reasonable computational time. Ideally, one should perform "the gold-standard" CCSD(T)/CBS calculation (see Chapter 3.3) for all points of the profile, especially when stacking interaction plays a significant role in the system (Hobza & Šponer 2002), but due to the high computational overload this is not always possible, and Following Riley et al (Riley et al. 2012), we limit these calculations to crucial points in the energy landscape.

## 2.2 Molecular dynamics

While "electrons in field of fixed nuclei" part of Born-Oppenheimer approximation is mainly reserved for QM methods, "nuclei in effective field of electrons", can be further simplified and treated classically. In the **classical approximation**, atoms are represented as *spheres* (also called "balls") of a given radius, mass and point charge. In a molecule, these balls are connected with *springs*, resembling a chemical bond, transforming a chemical system into a *mechanical body* (Lifson 1968).[2]

This representation allows the time-dependent Schrödinger equation to be replaced with *Newton's equations of motions*, bringing it into the world of classical mechanics. In this configuration atoms stretch, bend and rotate about their bonds as a response to inter- and intra-molecular forces. The energy of the entire system can be summed up with an energy function calculated for each pair of atoms in the system (also known as **pair-wise additive approximation**). These assumptions are the basis of the classical pair-additive force field Molecular Dynamics (MD).

[2] Long before the first successful molecular dynamics (MD) simulation was performed on biomolecule (McCammon 1977), chemists have used balls-and-sticks models to better understand 3-dimensional aspects of molecules by applying the laws of biophysics (similarly to Watson and Crick's DNA model; see Chapter 1.1.).

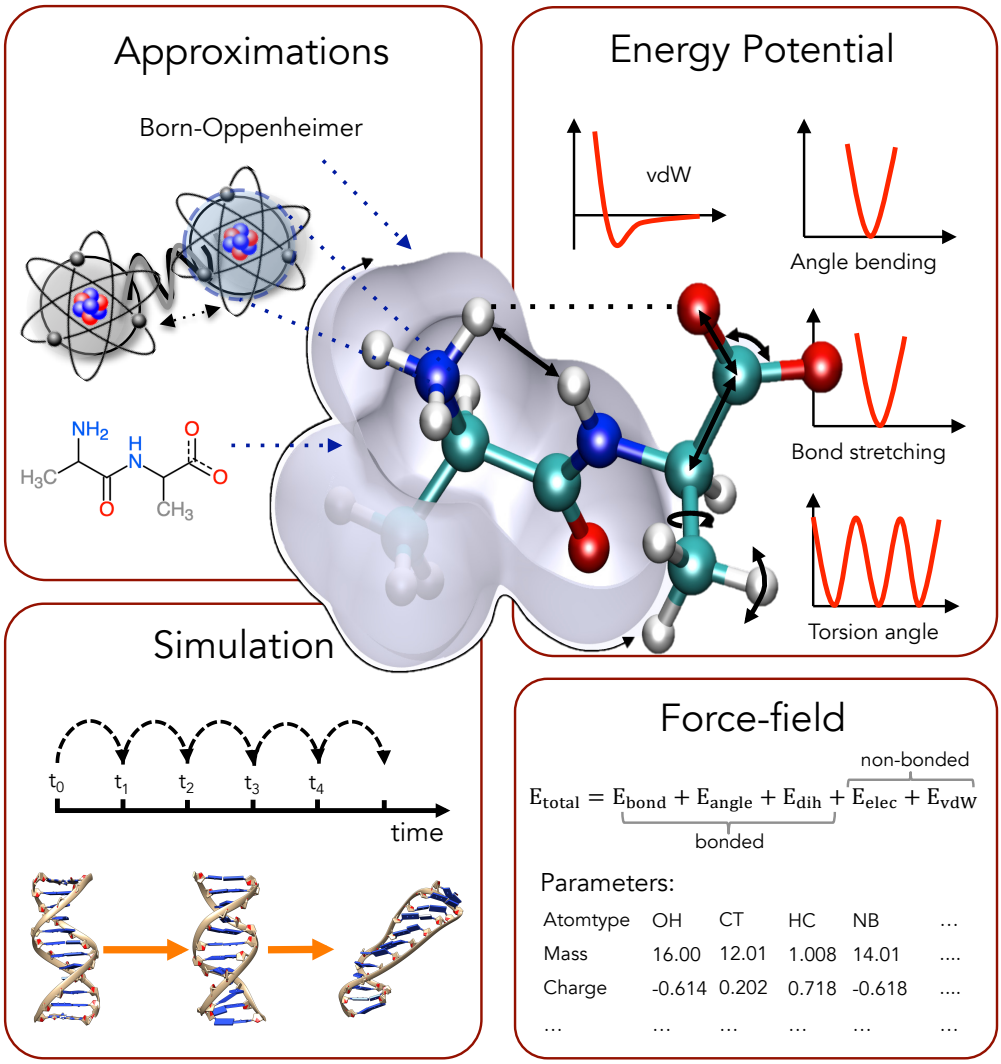



**Figure 2.1. Principles of Molecular Dynamics**. Illustration of basic principles of Molecular Dynamics.

Taking this setup, the potential energy function can be written as a sum of forces acting on each particle. These forces can be classified into two categories, bonded and non-bonded:

- **bond stretching** between two adjacent atoms of the bond length $r$ can be expressed in a form of the Hooke's law of elasticity:

$$E_{bond} = \sum_{all\ bonds} k_r \cdot (r - r_0)^2 \qquad \text{Eq. 2.17.}$$

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

- **angle bending** between two atoms adjacent to the third atom can be expressed also using Hooke's law as:

$$E_{angle} = \sum_{all\ angles} k_a \cdot (a - a_0)^2 \qquad \text{Eq. 2.18.}$$

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

- **torsion twisting** of a dihedral $\omega$ formed by two adjacent atoms bonded to two bonded atoms. Because of its periodicity, dihedral terms cannot be described by a harmonic term. The definition of dihedral angle varies depending on the MD software, but most commonly accepted form (also the form used in AMBER package (see **Chapter 2.5.**)) is represented like such:

$$E_{dih} = \sum_{all\ dihedrals} \sum_{n=1}^{4} \frac{V_{dih}}{2} \cdot (1 - \cos(n \cdot \omega - \phi)) \qquad \text{Eq. 2.19.}$$

where $V_{dih}$ represents the torsional barrier, $n$ the periodicity (typically in the range of 1 to 3 or 4) and $\phi$ the phase angle.

Non-bonded interactions between two non-covalently bonded atoms (as well as bonded atoms) are:

- **electrostatic interaction** between particles *i* and *j*, described by Coulomb's potential:

$$E_{elec} = \sum_{i,j} \frac{1}{4\pi\varepsilon} \cdot \frac{q_i q_j}{r_{ij}} \qquad \text{Eq. 2.20.}$$

where $\varepsilon$ is the dielectric constant of the medium and $r_{ij}$ the distance between two partial charges $q_i$ and $q_j$, of particle *i* and *j*, respectively.

- **van der Waals interaction**, the residual attractive and repulsive forces that count for intermolecular forces. The short-range repulsive term, also called Pauli repulsion, is due to overlapping of election clouds, while attraction at long range is due to the London dispersion forces. A good approximation of van der Waals interactions is Lennard-Jones potential:

$$E_{vdW} = \sum_{i,j} \varepsilon^* \left[ \left( \frac{r_m}{r_{ij}} \right)^{12} - 2 \left( \frac{r_m}{r_{ij}} \right)^{6} \right] \qquad \text{Eq. 2.21.}$$

where $\varepsilon^*$ 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 Lennard-Jones potential the $r^{-12}$ term counts for the short distance repulsion, while $r^{-6}$ term is used to represent dispersion interactions.

This potential function with fixed atom charges does not explicitly account for

the effects of induced electronic **polarization** between atoms, an effect of mutual relaxation of the electron clouds. Polarization leads to non-additivity, since any two molecules will interact differently when polarized by a third molecule than if the third molecule was not present. In classical MD, polarization can be introduced into the total energy of the system following two alternative models. *Induced point dipoles* represent the charge relaxation produced by polarization by means of atom-centered induced dipoles, $\vec{\mu}$, defined as the product of the total electric field, $\vec{E}$, and a scalar atomic polarizability, $\alpha$. The total electric field is composed of the external electric field from permanent charges $\vec{E}_0$ and the contribution from other induced dipoles. Due to the interdependence between total field and the induced dipole the system of equations need to be solved iteratively. The induce dipole approach is in practice difficult to incorporate, as under certain conditions, two inducible dipoles at short distances can cause a polarization catastrophe (when two interacting dipoles diverge at finite distance leading to nonphysical forces and velocities causing the simulation to fail) (Lindan 1995). *Drude oscillator* model (also known as shell model), models polarizability by adding an auxiliary massless particle (called Drude particle) with charge $q_{i,D}$harmonically attached to the fixed charged core. In Drude model, atomic polarizability, $\alpha_i$, is related to force constant *k* of the harmonic spring connecting the core and shell, determined by $\alpha_i = q_{i,D}^2/$. By fitting molecular polarizability data and experimental intermolecular interaction energies and other properties, charge magnitudes and harmonic force constants for the Drude particle may be obtained. Drude models are less sensate to polarization catastrophe since the shell interaction is associated with fairly steep repulsion function modeled by Lennard-Jones potential. In the field, induced point dipoles model approach is incorporated in AMBER and OPLS force fields (Wang et al. 2011; Schyman & Jorgensen 2013), while Drude oscillator models are more dominant models in CHARMM and GROMOS family of polarizable force fields (Savelyev & MacKerell 2014; Geerke & Van Gunsteren 2007). For a more detailed view on polarized force fields, see (Cieplak et al. 2009; Luque et al. 2011).

Since a typical atom is bonded to only a few of its neighbors, but interacts with every other atom in the molecule, the non-bonded terms are much more computationally costly to calculate in full, specially as the size of the system increases. The van der Waals term falls off rapidly with distance, polarization effects (if included) decays also reasonably fast (with $r^{-3}$), justifying the use of "cut-offs", but the Coulomb potential decays only with $r^{-1}$ and therefore converge in electrostatic happens only at very long distances. To overcome this problem, methods for long-range correction of the electrostatic potential have been suggested. The most used of these approaches is named **particle mesh Ewald (PME)** (Darden et al. 1993) and shows a perfect balance between accuracy and computational efficiency. PME divides the electrostatic energy into two terms, a short-range potential, calculated in the real space, and a long-range potential, which is calculated in the Fourier space. The advantage of PME is that both terms converge rapidly in their respective spaces thus

an introduction of a *cut-off distance* will not sacrifice the accuracy. Another advantage of PME is *scalability* with the number of atoms: $O(N \cdot logN)$, $N$ being the number of atoms in the system, compared to $O(N^2)$ that gives the direct calculation, which facilitates the calculations of larger systems.

PME assumes a periodicity of the system in order to be able to perform Fourier transformation. In MD, this is achieved by introducing **Periodic boundary conditions (PBC)**, a method in which the entire system is placed in a unit cell infinitely replicated in every direction around the unit cell to fill up the entire space. In this configuration, the periodic image of each atom moves the same way as the original one in the unit cell and if an atom leaves the unit cell into a periodic image, it would just enter to the unit cell from the same side from which it would enter into the periodic image (see Figure 2.2). Thus, the unit cell will maintain all the atoms during the whole simulation, while not giving it the sensation of a closed system. The unit cell can be in a shape of a *cube*, a *dodecahedron* or a *truncated octahedron* with a size big enough to avoid the biomolecule, being too close to the edge (usually at least 12 Å away). Even though the most common choice, truncated octahedron, has the smallest volume per radius, the choice of the shape of the box will depend on the shape of the molecule, especially in the case of long sequence of DNA, where the molecule can move in such way to interact with its periodic copy creating crystal-like artifacts.

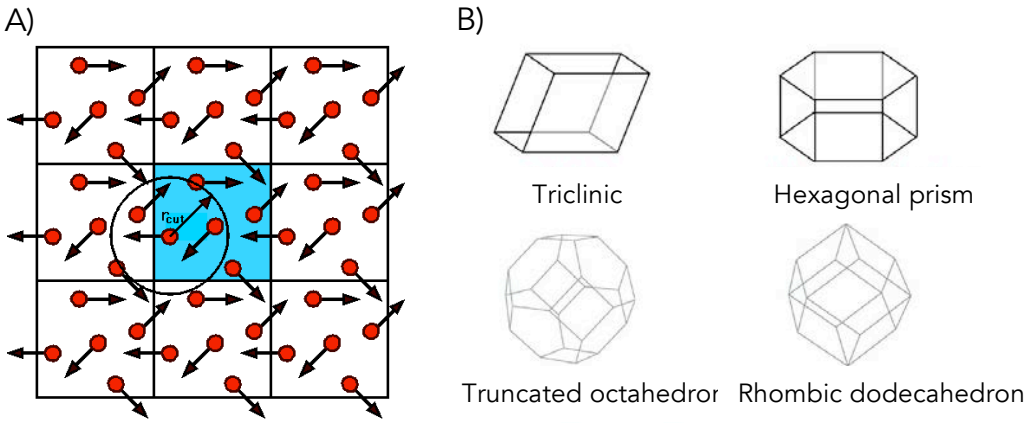



**Figure 2.2. Periodic boundary conditions.** An illustration of the PBC with the unit cell colored in blue (A) and an illustration of 3D unit cells most commonly used in MD simulations.

## 2.3 Force fields

Within a generic force field,

$$E_{total} = E_{bond} + E_{angle} + E_{dih} + E_{elec} + E_{vdW} \qquad \text{Eq. 2.22.}$$

we have to categorize the atoms and define the parameters for each atom type in order

to solve the function for a given biomolecule. The set of atom types and their parameters (together with its energy function) is called a **force field**. Atom types assignment depends on the functional group that the atom is part of (molecular environment) and/or hybridization state. For example, a carbon atom that is part of the DNA backbone, like $C_{5'}$ (connected to an oxygen, 2 hydrogens and a carbon; with a *sp3* hybridization), will behave differently, thus have different parameters, than a carbon in a nucleic base ring, like $C_5$ from Adenosine (connected to two carbons and a nitrogen; with a *sp2* hybridization).

There is no ideal universal force field for all biomolecules, rather a set of force fields developed for groups of biomolecules, like nucleic acids or proteins. The pioneer work of Lifson and coworkers (Lifson 1968) set the basis for most force fields, which have been extensively improved ever since. After years of evolution two families of force fields are the mostly used in the nucleic acids world:

- **AMBER** (Assisted Model Building with Energy Refinement) - originally developed by prof. Peter Kollman at UCSF, and maintained mainly by groups of prof. David Case at Rutgers University and prof. Thomas Cheatham at University of Utah, among many other active contributors. Version of *parm99* force field (Cheatham III et al. 1999) made big improvements at its prime and served as the basis for most used AMBER force fields nowadays, like *parm99-ILDN* (Lindorff‐Larsen et al. 2010) for proteins or *parmbsc0* (Pérez, Marchán, et al. 2007) for nucleic acids.

- **CHARMM** (Chemistry at HARvard Macromolecular Mechanics) - developed primarily at Harvard University in the group of prof. Martin Karplus, but prominently improved as well by other groups. Group of prof. Alex MacKerell have developed several versions of CHARMM force fields based on high-level QM data, most relevant being *CHARMM27* (MacKerell et al. 1998) or *CHARMM36* (Klauda et al. 2010), and group of prof. Benoit Roux with force fields for ions (Beglov et al. 1994); among other significant contributions the authors have made to the field (Lamoureux et al. 2003; Roux 1995).

Besides the two most prominent families, I should briefly mention **OPLS** family of force fields coming mainly from prof. William Jorgensen's group at Yale University, optimized to fit experimental properties of liquids, such as density and heat of vaporization, in addition to fitting gas-phase torsional profiles (called OPSL-AA) (Jorgensen et al. 1996), recent improvements of OPLS-AA makes this force field very powerful for proteins and a variety of ligands, it being very used in drug-design studies (Harder et al. 2015). Finally, it is worth to cite the **GROMOS** family of force fields developed by Berendsen and Van Gunsteren, based on reproducing the free enthalpies of hydration and apolar solvation for a range of compounds (Oostenbrink

et al. 2004). GROMOS is a very popular force field in the representation of dynamics of proteins, but its use to nucleic acids has been very limited.

Environment of most biomolecules MD simulations is made of water and ions, thus its representation is equally important as the representation of the molecule being studied. Most commonly used water type, *TIP3P* (Jorgensen et al. 1983), has been used in almost 20 thousand publications, while second most commonly used model, *SPC/E* (Berendsen et al. 1987) in over 7 thousand studies. From our experience both models produce similar results for nucleic acids simulations (see **Results** section and (Dans et al. 2016)), and represent a good compromise between accuracy and computational cost. Note that these water models are explicit representation of the solvent and that implicit representation (as continuous medium) can also be used (most common being MM/PBSA (Srinivasan et al. 1998; Kollman et al. 2000) or MM/GBSA (Born 1920; Hawkins et al. 1995)), thought it is often used as a post-processing tool for atomistic MD simulations, where it is applied to estimate free energy associated to solute-solvent interaction in structural and chemical processes, such as folding (Zhou & Doctor 2003; Benedix et al. 2009) or conformational transitions (Brice & Dominy 2011; Fogolari et al. 2005). The impact on the results of using different ion models in simulations of nucleic acids has been previously reported, with two most common monovalent ion models for AMBER family of force field (Smith & Dang 1994; Joung & Cheatham III 2008) giving similar results for DNAs, with a slight advantage of those from Smith & Dang (Noy et al. 2009).

Every system is unique at its own way and it is hard to recommend the perfect force field, but from in our experience, while CHARMM produces good results for proteins, its nucleic acids' force field do not produce as good results as the AMBER ones (see **Results** section and (Pérez, Marchán, et al. 2007; Dans et al. 2016; Ivani et al. 2015)).

## Force field evolution

From the first simulation of a small protein (McCammon et al. 1977) and the first simulation of a DNA molecule (Levitt 1983) in the '70s and '80s, MD simulations had seen dramatic improvements up to recent years with the first microsecond simulation of a Drew-Dickerson dodecamer (Pérez, Luque, et al. 2007) and a stunning full atomist protein folding on a millisecond time scale (Shaw et al. 2010). MD simulations have clearly shown its predicting power and precision for it to be used as a computational microscope for molecular biology (Dror et al. 2012). Fundamental to this progress is force field development, tightly coupled with computational advances. Faster processors, bigger supercomputers and advances in GPU technologies have allowed an increase in system size and longer simulation times, which sometimes yields problems never seen on smaller time scales (Perez et

al. 2008; Fadrná et al. 2009; Krepl et al. 2012; Drsňata et al. 2012). These inaccuracies are connected with glitches in force fields, glitches that are ought to be improved.

The basis of force field development relies in the **transferability approximation**, which implies that energy function developed on a small set of molecules applies to a wider range of molecules with similar chemical groups, given that parameters are not dependent on local environment. Thus, a small amount of atom types derived from small systems should be sufficient for describing any macromolecule. By parameterizing a residue, i.e. nucleotide unit, the overall results of a simulation of a larger DNA system would also be improved. Current parametrization efforts are "reactive" i.e. a new force fields appears as a response to errors detected, in a systematic and reproducible way in simulations. Once these errors have been spotted two strategies of correction can be applied. The most common approach implies a QM study on a small system with a direct comparison of the same system with MD results and eventual fitting of the parameters. This is the approach behind AMBER or CHARMM families of force fields (MacKerell et al. 1998; Huang & Mackerell 2013; Cheatham III et al. 1999; Pérez, Marchán, et al. 2007; Zgarbová et al. 2011; Zgarbová́ et al. 2013; Krepl et al. 2012). A big advantage of this approach is the precision of high-level QM data, and a potential disadvantage could be negligence of some neighboring effect; if the system is too small to capture all the effects that could influence the energy profile; thus one has to be careful when designing a system to do the study. Complementary approaches involves using experimental data as additional restrains in the fitting of the force field the reference, for example by biasing simulations to reproduce experimental structures (like CHARMM22/CMAP version for proteins (Mackerell 2004) or a recent RNA force field correction (Gil-Ley et al. 2016)), macroscopic liquid properties as in the case of the OPLS force field, or direct NMR observables (like in case of $\chi$ corrections from Turner's group (Yildirim et al. 2010)). These approaches have the advantage that guarantee the reproducibility of the experimental observable, but at the expense of potential problems for the description of other properties, as fitting to a given experimental observable, does not guarantee quality of the classical Hamiltonian.

## DNA force field problems

Being flexible and highly charged polymer, DNA is a difficult molecule to simulate. DNA is balanced by two opposite forces: strong **electrostatic repulsion** between the phosphate in the backbone, and **stacking** and **hydrogen bonding** between nucleobases. Additionally, solvent interactions tune these two forces and indirectly affect the shape of DNA double helix, transforming it from B- to Z- or A-form by changes in solvent environment. For that reason first DNA force fields always have had difficulties to properly simulate its structure.

In our opinion, the first reliable DNA force field was **parm99** force field (Cheatham III et al. 1999), which enabled the correct simulation of DNA on nanosecond time scale (up to 50 ns). However, in an extended run of a B-DNA duplex, parm99 introduced big distortion in the structure with massive $\alpha/\gamma$ transition to *gauche+/trans* geometry, away from canonical *gauche-/gauche+* state (Várnai & Zakrzewska 2004). These problems were corrected with **parmbsc0** (Pérez, Marchán, et al. 2007), which brought a big improvement for both DNA and RNA simulations, and enabled a first stable microsecond simulation of a B-DNA double helix (Pérez, Luque, et al. 2007). Parmbsc0 has been the "gold-standard" for DNA simulation. Nevertheless, some errors emerge as computer power increases allowing to explore more DNA structure for more extended periods of time.

- Parmbsc0 and all other nucleic acids' force fields until now *lack the ability to reproduce experimental values of helical parameters*, especially of *twist* and *roll*. Parmbsc0, undertwists the structure by, in average, 3º (Pérez, Luque, et al. 2007), which can produce major structural changes in long DNA polymers.

- Twist distribution profile is known to be bimodal for some base-pair steps, especially for CG step. This bimodality is connected with biologically relevant $B_I$ and $B_{II}$ state. Most force field, including parmbsc0 fail to reproduce properly this *bimodality* and *underestimate* $B_{II}$ *population* (Pérez, Luque, et al. 2007; Dans et al. 2012; Drsňata et al. 2012).

- DNA duplexes simulated with parmbsc0 show *excessive terminal fraying* (Zgarbová et al. 2014), which generate some distortions in neighboring base pairs and the formation of unrealistic hydrogen bond patterns at terminal base pairs.

- Given that the DNA conformational space is quite diverse, DNA force fields, such as parmbsc0, have *difficulties in reproducing non-canonical structures* like quadruplex loops, or Z-DNA (Krepl et al. 2012) which are far from the canonical B-DNA which was used for calibration of the force field.

These problems can be tracked down to wrong description of several backbone dihedrals, sugar puckering and $\chi$ torsion. Solving these problems is one of the main topics of this thesis.

## 2.6. Molecular dynamics algorithm

Now that we have the energy function completely defined for a given system we can apply the laws of classical mechanics (Newton's laws of motion) and see the time evolution of the system. Forces acting on a particle with coordinates $X$ are proportional to the negative gradient of its potential energy $U$,

$$\vec{F}(X) = -\nabla U(X) \qquad \text{Eq. 2.23.}$$

as well as are proportional to its mass *m* and acceleration *a*,

$$\vec{F}(X) = m\vec{a}(X) = m\frac{d\vec{v}}{dt} = m\frac{d^2\vec{x}}{dt^2} \qquad \text{Eq. 2.24.}$$

As time dependence of force fields is complex, this equation needs to be solved numerically. In short, starting from a set of initial coordinates $\overrightarrow{X_0}$ and velocities $\overrightarrow{V_0}$, the sequential coordinates, $\overrightarrow{X_1}$would be obtained as following

$$\overrightarrow{X_1} = \overrightarrow{X_0} + \overrightarrow{V_0}\Delta t + \frac{1}{2}\vec{a}(X_0)\Delta t^2 \qquad \text{Eq. 2.25.}$$

where $\Delta t$ is small period of time (ideally infinitesimally small). Given the initial coordinates, which come from experimental structures, and velocities, typically assigned from the Maxwell-Boltzmann distribution given the temperature of the system, we can calculate all future positions and velocities.

This procedure is called *Verlet-leapfrog integration* and it is done in an iterative way with a given **time step**, The integration time step ($\Delta t$) is chosen to be not bigger than the shortest motion in the system, which for biological system is the bond stretching of a hydrogen atom, happening on *1 femtosecond* timescale. As these vibrations are irrelevant for the final result at biological level, they can be constrained by means of special algorithms, most notably SHAKE (Ryckaert et al. 1977) (used in AMBER simulation packages (see bellow)), LINCS (Hess et al. 1997) (used in GROMACS package) or RATTLE (Andersen 1983) (used in NAMD package), which allow longer integration step (2 femtoseconds).

Classical Newtonian setup assumes conservation of energy of the system, also known as *microcanonical ensemble*, where the number of particles N, the volume V and the total energy E are conserved (**NVE** ensemble). To capture more experimental-like conditions, with constant temperature and pressure, ensembles like canonical or isothermal-isobaric give a more appropriate description. In the canonical ensemble (**NVT** ensemble), the energy is allowed to vary but temperature is conserved by means of a thermostat. Most popular techniques to control temperature include from weaker ones like velocity rescaling to the *Nose-Hoover thermostat* (Nosé 1984; Hoover 1985), *Berendsen thermostat* (Berendsen et al. 1984) and *Langevin dynamics* (Brooks et al. 1985), which adds friction and random forces. The most similar conditions to the experimental ones are reproduced with isothermal-isobaric ensemble (**NPT**), where both temperature and pressure are kept constant, allowing the volume to change. Similarly to thermostats pressure can be controlled by Berendsen or Nose-Hoover bath, or *Parrinello-Rahman barostat* (Parrinello & Rahman 1981). Each bath has its own advantages and disadvantages; weakly coupled ones have a stable ensemble average but not good perturbation, while stochastic methods disturb

dynamics.

Most of the technical details (small difference in energy function and various algorithms) depend on the software used to perform the MD simulation. In this work, I used extensively software package **AMBER** (Case et al. 2012). As mentioned in the force field section (see Section 2.4) AMBER was developed by Kollman's group and is now maintained by his former group members. AMBER software package consists of two main parts: *AMBER simulation package* and *Ambertools*. Recently, AMBER engine has been re-written to take advantage of the new and popular GPU architecture (Case et al. 2014), showing extreme computer efficiency. Another MD software that I used is **GROMACS** (GROningen Machine for Chemical Simulation) (Hess et al. 2008), originated from Berendsen's group and maintained and improved now by groups of prof. Berk Hess and prof. Erik Lindahl. Their free MD simulation package consisting of tools to prepare, run and analyze MD simulations is probably the fastest CPU implementation of a MD algorithm. I should also mention **CHARMM** (Brooks et al. 2009) (mentioned in Section 2.4) and **NAMD** (Phillips et al. 2005) (NAno scale Molecular Dynamics), which I had a brief chance of using, but are also very popular in the field, specially the latest which provide excellent performance for huge systems in large supercomputers.

## 2.7 Enhanced sampling and free energy

Ideally, in an infinite simulation time, one should be able to obtain the free energy associated to any state of the system by just counting the population of this state over the others. Unfortunately this "pure-force" approach is not useful when the state of interest has a low population and is not likely to be sampled spontaneously in a finite-time MD simulation. This forces the use of biasing strategies which guarantee sufficient sampling of the required states (Van Gunsteren 1989; Sitkoff et al. 1994; Kumar et al. 1992). This methods are based on the concept of the **potential of mean force** (PMF; Kirkwood 1935). The PMF of a system with *N* molecules is strictly the potential that gives the average force over all the configurations of all the *n+1*...*N* molecules acting on a particle *j* at any fixed configuration keeping fixed a set of molecules *1*...*n*. In a more practical way, the PMF can be used to know how the free energy changes as a function of a coordinate of the system. PMF simulations are often used in conjunction with enhanced sampling method that guarantee a smooth and continuous sampling of the conformational transition, whose associated free energy we want to determine.

**Umbrella Sampling** (US) (Torrie & Valleau 1977) is a commonly used computational technique used in MD to improve sampling of a system. The basic principle of US lays in improving the sampling along an arbitrary coordinate, also called *collective variable* (CV), by adding an artificial biasing potential ($V_b(\chi)_i$),

mostly in a form of a harmonic restrain (which resamples an inverted umbrella; see Eq. 2.26 and Figure 2.3) that is moved systematically along the conformational coordinate to force a complete sampling in all the transition path.

$$V_b(\chi)_i = \frac{1}{2}k(\chi - \chi_i)^2 \qquad \text{Eq. 2.26.}$$

where $\chi$ is the coordinate of the collective variable, $\chi_i$ is the reference point of the coordinate, where the bias is zero and $k$ is the strength of the harmonic bias.

By doing so, the method is able to determine the probability for the system to be in a given conformation and from there obtain the energy landscape by inverting the Boltzmann distribution, once the bias in the population introduced by the umbrella potential has been removed (see below). The CV ($\chi$) can be a given angle or dihedral (an easy choice) or a combination of coordinates. Two crucial points of US are a proper sampling of the entire configurational space (i.e. the histograms of all the neighboring umbrellas should overlap) and a correct choice of CV (a hidden coordinate can produce inconsistency in final energy profile; mostly the case of choosing a reaction coordinate as CV). Overall, US is a powerful and easy-to-use technique as it implies adding a simple harmonic potential (a feature that is available in almost any MD code).

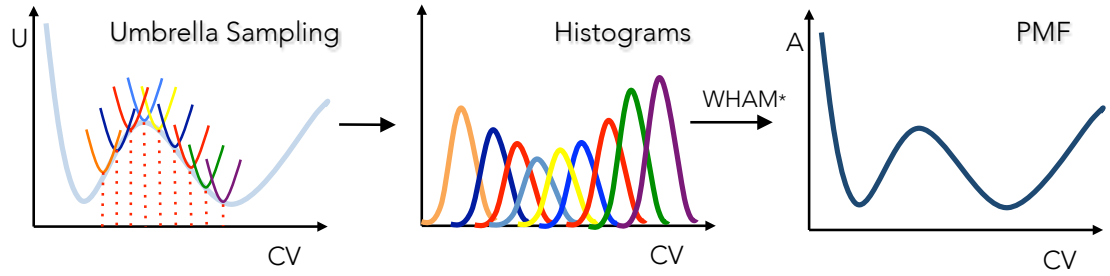



**Figure 2.3. The Principles of Umbrella Sampling and Potential of Mean Force** (PMF). Histograms from Umbrella sampling along the coordinate are used to reconstruct PMF profiles using WHAM (see Chapter 3.5 for more details).

**Metadynamics** is **a**n algorithm developed in Parrinello's group (Laio & Parrinello 2002), which inserts memory in the sampling by adding a positive Gaussian potential along the simulation based on the location of the system in terms of collective variable (see Figure 2.4.). The technique has been improved in recent years to solve problems with convergence (Barducci et al. 2008), and has the advantage over US that allows the use of more complex collective variables and the disadvantage that its use is less straightforward.

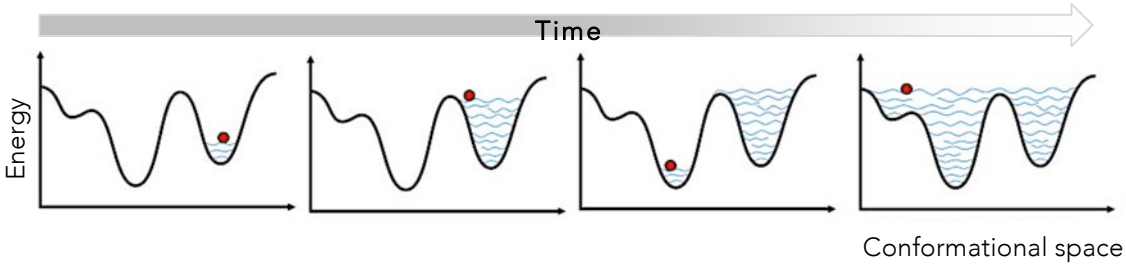



**Figure 2.4. Metadynamics.** An illustration of the PBC with the unit cell colored in blue (A) and an illustration of 3D unit cells used in MD simulations.

**Replica-exchange MD** (REMD) is another very popular enhanced sampling method (Sugita & Okamoto 1999), in which a set of non-interacting replicas running at different values of an exchange variable, usually temperature (T-REMD), are swapped at neighboring replicas at specific intervals, based on a Monte-Carlo acceptance criterion. REMD can be a very efficient method as no additional external potential is added and properties at the given temperature (even though discontinuous, but still follow a proper Boltzmann distribution) can be extracted from one replica. REMD is not hypothesis-driven, which makes them a perfect choice when no clear idea of the reaction coordinate exist, but this is also the disadvantage as the technique is very costly and often produce little information on the process of interest.

The study of biomolecular systems by MD is particularly challenging given the timescales relevant to biomolecular processes (for example nucleosome rearrangement) are often in microseconds to seconds, enormously long compared to the time-step required for stable integration (2 fs). Even with expensive hardware such as Anton computer (Shaw et al. 2009), MD simulations can barely reach biomolecular timescales. One solution to the timescale problem consists of extracting stochastic kinetic information from multiple simulations that are shorter than the timescales of interest to build a discrete-state stochastic model, called **Markov state model (MSM)**, capable of describing long-time statistical dynamics (Chodera & Noé 2014). MSM approach is similar to biochemists' approach for describing a chemical reaction by means of "states and rates". In contrary to just a handful of states used by experimentalists, MSM builds a model defined by *N* states (thousands to potentially millions) and a transition matrix (from the rates). The idea behind having many states is that it allows one to construct a very high resolution model of the intrinsic dynamics from relatively short MD trajectories, since the kinetic distance between adjacent states is small. The requirement for a big quantity of short trajectories can take an advantage of using computer architectures with many less-powerful cores. Initiatives like Folding@Home (Shirts & Pande 2000) or GPUGRID (Buch et al. 2010) have been exploiting this advantage to study in more details complex processes like protein folding (Snow et al. 2002) or kinetics of disordered proteins (Stanley et al. 2014).

Many more enhanced sampling techniques have been developed, but we limit ourselves the explanation to those used along this thesis. For a better perspective on enhanced sampling technique see for example (Bernardi et al. 2014).

# BIBLIOGRAPHY TO CHAPTER 2

Andersen, H.C., 1983. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52(1), pp.24–34.

Bachs, M., Luque, F.J. & Orozco, M., 1994. Optimization of solute cavities and van der Waals parameters in ab initio MST-SCRF calculations of neutral molecules. *Journal of computational chemistry*, 15(4), pp.446–454.

Barducci, A., Bussi, G. & Parrinello, M., 2008. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2).

Becke, A.D., 1988. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6), pp.3098–3100.

Becke, A.D., 1993. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7), p.5648.

Beglov, D., Roux, B. & Hc, C., 1994. Finite representation of an infinite for computer simulations bulk system : Solvent boundary potential. *Journal of Medical Physics*, 100(June), pp.9050–9063.

Benedix, A. et al., 2009. Predicting free energy changes using structural ensembles. *Nature methods*, 6(1), pp.3–4.

Berendsen, H.J.C. et al., 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8), pp.3684–3690.

Berendsen, H.J.C., Grigera, J.R. & Straatsma, T.P., 1987. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24), pp.6269–6271.

Bernardi, R.C., Melo, M.C.R. & Schulten, K., 2014. Enhanced sampling techniques in molecular dynamics simulations of biological systems ☆. *BBA - General Subjects*, 1850, pp.872–877.

Born, M., 1920. Volumen und hydratationswärme der ionen. *Zeitschrift für Physik A Hadrons and Nuclei*, 1(1), pp.45–48.

Boys, S.F., 1950. Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 200(1063), pp.542–554.

Brice, A.R. & Dominy, B.N., 2011. Analyzing the robustness of the MM/PBSA free energy calculation method: application to DNA conformational transitions. *Journal of computational chemistry*, 32(7), pp.1431–1440.

Brooks, B.R. et al., 2009. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), pp.1545–1614.

Brooks, C.L., Brünger, a & Karplus, M., 1985. Active site dynamics in protein molecules: a stochastic boundary molecular-dynamics approach. *Biopolymers*, 24(5), pp.843–865.

Buch, I. et al., 2010. High-throughput all-atom molecular dynamics simulations using distributed computing. *Journal of chemical information and modeling*, 50(3), pp.397–403.

Cances, E., Mennucci, B. & Tomasi, J., 1997. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of Chemical Physics*, 107(8), pp.3032–3041.

Case, D.A. et al., 2012. AMBER 12. *University of California, San Francisco*, 142.

Case, D.A. et al., 2014. Amber 14.

Cheatham III, T.E., Cieplak, P. & Kollman, P.A., 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 16(4), pp.845–862.

Chodera, J.D. & Noé, F., 2014. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25, pp.135–144.

Cieplak, P. et al., 2009. Polarization effects in molecular mechanical force fields. *Journal of physics. Condensed matter : an Institute of Physics journal*, 21(33), p.333102.

Cĺžek, J., 1966. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *The Journal of Chemical Physics*, 45(11), p.4256.

Cizek, J. & Paldus, J., 1980. Coupled Cluster Approach. *Phys. Scr.*, 21(3–4), p.251.

Cossi, M. et al., 2003. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *Journal of Computational Chemistry*, 24(6), pp.669–681.

Cramer, C.J., 2004. *Essentials of Computational Chemistry Theories and Models*,

Dans, P.D. et al., 2012. Exploring polymorphisms in B-DNA helical conformations. *Nucleic acids research*, 40(21), pp.10668–10678.

Dans, P.D. et al., 2016. Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Research*, (April), p.gkw264.

Darden, T., York, D. & Pedersen, L., 1993. Particle mesh Ewald: An N · log (N) method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12), pp.10089–10092.

Dror, R.O. et al., 2012. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41, pp.429–52.

Drsňata, T. et al., 2012. Structure, stiffness and substates of the Dickerson-Drew dodecamer. *Journal of chemical theory and computation*, 9(1), pp.707–721.

Fadrná, E. et al., 2009. Single stranded loops of quadruplex DNA as key benchmark for testing nucleic acids force fields. *Journal of Chemical Theory and Computation*, 5(9), pp.2514–2530.

Fogolari, F. et al., 2005. MM/PBSA analysis of molecular dynamics simulations of bovine β-lactoglobulin: Free energy gradients in conformational transitions? *Proteins: Structure, Function, and Bioinformatics*, 59(1), pp.91–103.

Geerke, D.P. & Van Gunsteren, W.F., 2007. On the calculation of atomic forces in classical simulation using the charge-on-spring method to explicitly treat electronic polarization. *Journal of Chemical Theory and Computation*, 3(6), pp.2128–2137.

Gil-Ley, A., Bottaro, S. & Bussi, G., 2016. RNA Conformational Ensembles: Narrowing the GAP between Experiments and Simulations with Metadynamics. *Biophysical Journal*, 110(3), p.522a–523a.

Van Gunsteren, W.F., 1989. Methods for calculation of free energies and binding constants: Successes and problems. *Computer simulation of biomolecular systems: Theoretical and experimental applications*, pp.27–59.

Halkier, A. et al., 1998. Basis-set convergence in correlated calculations on Ne, N 2, and H 2 O. *Chemical Physics Letters*, 286(3), pp.243–252.

Halkier, A. et al., 1999. Basis-set convergence of the energy in molecular Hartree–Fock calculations. *Chemical physics letters*, 302(5), pp.437–446.

Harder, E. et al., 2015. OPLS3: a force field providing broad coverage of drug-like

small molecules and proteins. *Journal of Chemical Theory and Computation*, 12(1), pp.281–296.
Hawkins, G.D., Cramer, C.J. & Truhlar, D.G., 1995. Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters*, 246(1–2), pp.122–129.
Hess, B. et al., 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3), pp.435–447.
Hess, B. et al., 1997. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12), pp.1463–1472.
Hobza, P. & Šponer, J., 2002. Toward true DNA base-stacking energies: MP2, CCSD (T), and complete basis set calculations. *Journal of the American Chemical Society*, 124(39), pp.11802–11808.
Hobza, P., Sponer, J. & Reschel, T., 1995. Density Functional Theory and Molecular Clusters. *Journal of Computational Chemistry*, 16(11), pp.1315–1325.
Hoover, W.G., 1985. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3), pp.1695–1697.
Huang, J. & Mackerell, A.D., 2013. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25), pp.2135–2145.
Ivani, I. et al., 2015. Parmbsc1: a refined force field for DNA simulations. *Nature methods*, 13(1), pp.55–58.
Jorgensen, W.L. et al., 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2), pp.926–935.
Jorgensen, W.L., Maxwell, D.S. & Tirado-Rives, J., 1996. Development and Testing of the OLPS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(15), pp.11225–11236.
Joung, I.S. & Cheatham III, T.E., 2008. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B*, 112(30), pp.9020–9041.
Kirkwood, J.G., 1935. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5), pp.300–313.
Klamt, A. et al., 2009. On the performance of continuum solvation methods. A comment on "Universal approaches to solvation modeling." *Accounts of chemical research*, 42(4), pp.489–492.
Klauda, J.B. et al., 2010. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *Journal of Physical Chemistry B*, 114(23), pp.7830–7843.
Kollman, P.A. et al., 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, 33(12), pp.889–897.
Krepl, M. et al., 2012. Reference simulations of noncanonical nucleic acids with different χ variants of the amber force field: Quadruplex dna, quadruplex rna, and z-dna. *Journal of chemical theory and computation*, 8(7), pp.2506–2520.
Krishnan, R. et al., 1980. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *The Journal of Chemical Physics*, 72(1), pp.650–654.
Kumar, S. et al., 1992. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8), pp.1011–1021.

Kümmel, H.G., 2003. a Biography of the Coupled Cluster Method. *International Journal of Modern Physics B*, 17(28), pp.5311–5325.

Laio, A. & Parrinello, M., 2002. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20), pp.12562–12566.

Lamoureux, G., MacKerell Jr, A.D. & Roux, B., 2003. A simple polarizable model of water based on classical drude oscillators. *The Journal of chemical physics*, 119(10), pp.5185–5197.

Lee, C., Yang, W. & Parr, R.G., 1988. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2), pp.785–789.

Levitt, M., 1983. Computer simulation of DNA double-helix dynamics. In *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, pp. 251–262.

Lifson, S., 1968. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics*, 49(11), p.5116.

Lindan, P.J.D.J.D., 1995. Dynamics with the Shell Model. *Molecular Simulation*, 14(4–5), pp.303–312.

Lindorff-Larsen, K. et al., 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8), pp.1950–1958.

Luque, F.J. et al., 2011. Polarization effects in molecular interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5), pp.844–854.

MacKerell, A.D. et al., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18), pp.3586–3616.

Mackerell, A.D., 2004. Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry*, 25(13), pp.1584–1604.

Marenich, A. V, Cramer, C.J. & Truhlar, D.G., 2009. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B*, 113(18), pp.6378–6396.

McCammon, J.A. et al., 1977. Dynamics of folded proteins. *Nature*, 267(5612), p.16.

Miertuš, S., Scrocco, E. & Tomasi, J., 1981. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics*, 55(1), pp.117–129.

Miertus, S. & Tomasi, J., 1982. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chemical physics*, 65(2), pp.239–245.

Monticelli, L. & Salonen, E., 2013. *Biomolecular simulations*, Humana Press;

Nosé, S., 1984. A unified formulation of the constant temperature molecular dynamics methods. *Journal of Chemical Physics*, 81(1), pp.511–519.

Noy, A. et al., 2009. The impact of monovalent ion force field model in nucleic acids simulations. *Physical Chemistry Chemical Physics*, 11(45), pp.10596–10607.

Oostenbrink, C. et al., 2004. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25(13), pp.1656–1676.

Parrinello, M. & Rahman, A., 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12), pp.7182–7190.

Perez, A. et al., 2008. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic acids research*, 36(7), pp.2379–2394.

Pérez, A., Marchán, I., et al., 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11), pp.3817–3829.

Pérez, A., Luque, F.J. & Orozco, M., 2007. Dynamics of B-DNA on the microsecond time scale. *Journal of the American Chemical Society*, 129(47), pp.14739–14745.

Phillips, J.C. et al., 2005. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16), pp.1781–1802.

Rapaport, D.C.C., 2004. *The Art of Molecular Dynamics Simulation*,

Riley, K.E. et al., 2012. Assessment of the performance of MP2 and MP2 variants for the treatment of noncovalent interactions. *Journal of Physical Chemistry A*, 116(16), pp.4159–4169.

Roux, B., 1995. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1–3), pp.275–282.

Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.C., 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3), pp.327–341.

Savelyev, A. & MacKerell, A.D., 2014. All-atom polarizable force field for DNA based on the classical drude oscillator model. *Journal of computational chemistry*, 35(16), pp.1219–1239.

Schyman, P. & Jorgensen, W.L., 2013. Exploring Adsorption of Water and Ions on Carbon Surfaces using a Polarizable Force Field. *The journal of physical chemistry letters*, 4(3), pp.468–474.

Shaw, D.E. et al., 2010. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002), pp.341–346.

Shaw, D.E. et al., 2009. Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking Storage and Analysis SC 09*, (c), p.1.

Shirts, M. & Pande, V.S., 2000. Screen savers of the world unite! *Science*, 290(5498), pp.1903–1904.

Sitkoff, D., Sharp, K.A. & Honig, B., 1994. Accurate calculation of hydration free energies using macroscopic solvent models. *The Journal of Physical Chemistry*, 98(7), pp.1978–1988.

Smith, D.E. & Dang, L.X., 1994. Computer simulations of NaCl association in polarizable water. *The Journal of Chemical Physics*, 100(5), pp.3757–3766.

Snow, C.D. et al., 2002. Absolute comparison of simulated and experimental protein-folding dynamics. *nature*, 420(6911), pp.102–106.

Srinivasan, J. et al., 1998. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *Journal of the American Chemical Society*, 120(37), pp.9401–9409.

Stanley, N., Esteban-Martín, S. & De Fabritiis, G., 2014. Kinetic modulation of a disordered protein domain by phosphorylation. *Nature communications*, 5.

Sugita, Y. & Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1), pp.141–151.

Szabo, A. & Ostlund, N.S., 1996. Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory. *Introduction to Advanced Electronic Structure Theory*, p.480.

Tirado-Rives, J. & Jorgensen, W.L., 2008. Performance of B3LYP density functional

methods for a large set of organic molecules. *Journal of Chemical Theory and Computation*, 4(2), pp.297–306.
Torrie, G.M. & Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2), pp.187–199.
Truhlar, D.G., 1998. Basis-set extrapolation. *Chemical Physics Letters*, 294(1–3), pp.45–48.
Várnai, P. & Zakrzewska, K., 2004. DNA and its counterions: a molecular dynamics study. *Nucleic acids research*, 32(14), pp.4269–4280.
Vosko, S.H., Wilk, L. & Nusair, M., 1980. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics*, 58(8), pp.1200–1211.
Wang, J. et al., 2011. Development of polarizable models for molecular mechanical calculations II: Induced dipole models significantly improve accuracy of intermolecular interaction energies. *Journal of Physical Chemistry B*, 115(12), pp.3100–3111.
Woon, D.E. & Dunning Jr, T.H., 1993. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *The Journal of chemical physics*, 98(2), pp.1358–1371.
Yildirim, I. et al., 2010. Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine. *Journal of chemical theory and computation*, 6(5), pp.1520–1531.
Zgarbová, M. et al., 2013. Toward improved description of DNA backbone: revisiting epsilon and zeta torsion force field parameters. *Journal of chemical theory and computation*, 9(5), pp.2339–2354.
Zgarbová, M. et al., 2014. Base pair fraying in molecular dynamics simulations of DNA and RNA. *Journal of Chemical Theory and Computation*, 10(8), pp.3177–3189.
Zgarbová, M. et al., 2011. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation*, 7(9), pp.2886–2902.
Zhao, Y. & Truhlar, D.G., 2008. Density functionals with broad applicability in chemistry. *Accounts of Chemical Research*, 41(2), pp.157–167.
Zhou, G.P. & Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. *Proteins: Structure, Function and Genetics*, 50(1), pp.44–48.

*"Man is a tool-using animal. Without tools he is nothing, with tools he is all."*
*Thomas Carlyle*

# 3| METHODS

The focus of this thesis is set on force field parameterization and validation of its MD results. This chapter describes general force field parameterization scheme using high-level QM data as input, obtaining QM profiles, and comparing them with MM profiles obtained from Umbrella sampling. Once the force field is done, validation is performed through comparison of MD simulations with experimental finding. Later, a description of the protocol to prepare a system for MD simulation and methods to analyze MD results, from global to local measures, with the focus on helical analysis, as well as stiffness, PCA and entropy analysis is described. Lastly, the method for direct comparison of MD results with experimental observables like NOEs and RDCs is explained.

# 3.1. Force field parameterization scheme

**Finding the problem**

Force field refinement is mostly “reactive”, i.e. it is done as a response of the presence of errors in the available force field. As simulation time increases with computational power increase, discrepancies from experimental data (hidden in shorter trajectories) emerge. Finding the exact force field term that causes inaccuracy is a complex process considering the number of terms that play a role in a given bimolecular system and the coupling between them. Assuming that Coulomb and van der Walls terms are reproduced accurately enough for most nowadays timescales’ simulations, with additional in-depth analysis, one can reduce the problem to a set of dihedral term that is directly connected with the aberration, given the case.

**Isolating the difference**

After choosing the term (in our case a dihedral) that we suspect is the responsible of a bad behavior of the simulation one has to determine the discrepancy between correct energy landscape and the one reproduced by MD simulations, for the given dihedral. The aim of this step is to isolate the contribution of the term to the overall energy profile (or to exclude it from the overall profile) in order to be able to fit that contribution to the correct profile. The choice of “correct” energy profile can come typically from high-level QM profile, which can be ideally corrected if experimental conformational populations are available.

Two approaches are most common in obtaining MM profiles. A clear choice n vacuum simulations is doing single point MD calculations of QM optimized geometries. However, for most purposes it is preferable to reproduce the behavior in solution. In our work we fit force field to reproduce conformation of nucleic acids in water. For this purpose, we perform PMF calculations in solution (see Chapter 2.7.) to obtain comparable profiles with those obtained at the QM level when explicit or implicit solvent effects are included. Once the QM and the MM profiles are obtained the fitting term will be the difference between QM obtained profile and the MM energy profile with the specific term set to zero (see Eq. 3.1).

$$E_{diff}(x) = E_{QM} - E_{MM(x=0)} \qquad \text{Eq. 3.1.}$$

where QM calculations need to consider solvent, either by QM/MM calculations with explicit solvent molecules, or by SCRF calculations (see previous chapter).

**Fitting procedure**

For dihedral terms fitting we used an in-lab developed algorithm based on a locally developed Metropolis-Monte Carlo criteria with weights on fitting points, allowing us to reinforce the fitting at given points, or introduce if required

experimental information on conformer population. The difference $E_{diff}$ (see Eq.3.1) was fitted to the Fourier series term of the 3rd order (see Eq.3.2).[1]

$$E_n(x) = \sum_{1}^{n} \frac{V_n}{2}(1 + \cos(nx - \phi) \qquad \text{Eq. 3.2.}$$

The algorithm is adapted for 2-dimensional fitting as well, especially convenient for coupled dihedrals like $\varepsilon/\zeta$. As perfect fitting is not always possible, data weighting allows us to be focused on areas of higher interest or putting more importance to certain data points, such as those coming from very high level QM calculations (see below).

## 3.2. QM calculations

In Section 2.1, I briefly described the theory of the QM methods that I used in this thesis. There are several programs available to do these calculations, from which we choose the *Gaussian 03* (Frisch et al. 2004) and *Gaussian 09* (Frisch et al. 2009) ones, due to their flexibility and the large variety of calculations that allowed us to perform. .

To obtain a QM profiles I first build a system that will capture the dihedral of interest and chemically relevant surrounding, but with a reduced size that allows the access to high level QM calculations. This is an important step as small systems can neglect some important neighboring effects while big system can be very computationally demanding and hard to converge. During geometry optimization other backbone and sugar dihedrals are constrained to typical values obtained from a survey of crystal structure, to reduce noise in the parameterization process. My general strategy for obtaining QM profiles was that I firstly optimized geometries on DFT level using B3LYP (Becke 1988; Lee et al. 1988) functional with a split-valence Pople-type basis set (Krishnan et al. 1980) with polarization functions (d,p) and diffusion functions (++) added to both heavy and light atoms, denoted as **B3LYP/6-31++G(d,p)**. Once the geometry optimization is done, I performed single point calculations at the MP2 level using a basis set of Dunning-type (Woon & Dunning Jr 1993) correlation-consistent polarized (cc-p) on first- and second- row atoms (VDZ) with diffuse functions (aug), denoted as **MP2/aug-cc-pVDZ**. 'Correlation-consistent polarized' basis sets are designed to converge to complete-basis-set following well-defined protocols by Halkier or Truhlar (CBS; see Chapter 2.1.). For solvent correction I used polarized continuum model (PCM). In calculation done with Gaussian 03 I used the PCM-MST model developed in the group (Miertuš et al. 1981;

[1] *In case of OL-family of force field, they usually extended the function to the 4th order. In our experience, this is not needed for the quality of the fit, especially in case of DNA.*

Miertus & Tomasi 1982; Cances et al. 1997; Bachs et al. 1994; Orozco & Luque 2000), and for calculations done with Gaussian 09 I used Cramer and Truhlar (SMD) (Marenich et al. 2009) implementation with standard integral equation formalism IEF-PCM (Cances et al. 1997). Both default MST and SMD implementations produce nearly identical results (see Figure 3.1.).

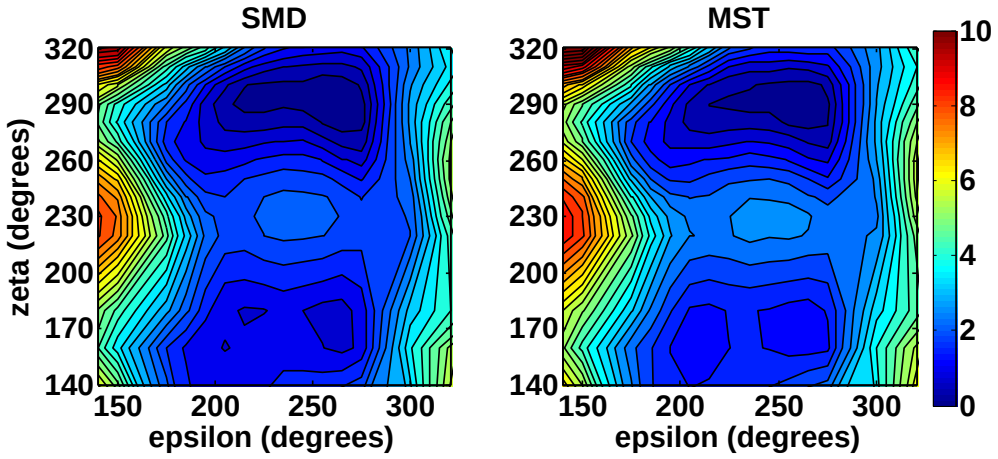



**Figure 3.1. Comparison of QM profiles of $\varepsilon/\zeta$ with 2 PCM solvent corrections, MST and SMD.** Contour plots show $\varepsilon/\zeta$ energy landscape represented in kcal/mol units.

On key points of the QM profile (like minima and maxima) I ran a CCSD(T)-complete-basis-set calculation in order to get more precise energy values.

## 3.3. The state-of-the-art CCSD(T)/CBS calculations

Complete Basis-Set (CBS) approach is based on extrapolating the results to those obtained if all the space was saturated with Gaussian functions. If coupled to a high level QM calculation (ideally CCSD(T)) they represent the "gold-standard" in quantum chemistry. The idea is based on extrapolating to infinite the results obtained with increasingly large basis set for a medium/high level calculation, followed by increase from in the level of calculation for a medium-size basis set. As noted above, the two most common extrapolation schemes to move to CBS limit come from Helgaker and co-workers (Halkier et al. 1998; Halkier et al. 1999), and Truhlar (Truhlar 1998). Both assume that the Hartree-Fock (HF) energy term and the correlation term of the total energy behave differently when the CBS limit is approached, and as a result, they are treated differently. I have used Helgaker's approach, presented as follows:

$$E_X^{HF} = E_{CBS}^{HF} + Ae^{-\alpha X} \qquad \text{Eq. 3.3.}$$

$$E_X^{corr} = E_{CBS}^{corr} + BX^{-3} \qquad \text{Eq. 3.4.}$$

where $\alpha$ is the fitting parameter obtained in the original work ($\alpha$ = 1.43 for X=2,3 and $\alpha$ = 1.54 for X=3,4), while $A$ and $B$ are fitting parameters, and $E^{corr}$ is the correlation energy. As this method is a two-point extrapolation, it is necessary to perform calculations at two subsequent levels, meaning that for X=2 the double-*zeta* basis set (aug-cc-pVDZ) is used, for X=3 the triple-*zeta* (aug-cc-pVTZ), etc. In this studies I used X=2,3 and X=3,4 (for DNA $\varepsilon/\zeta$) combinations.

As noted above, the finite size of basis set is a source of uncertainty in QM calculations, but another equally important source of problems is related to the incorrect representation of correlation effects. As the cost of introducing better correlation scales dramatically with the basis set we perform a divide-and-conquer approach by doing CBS extrapolation at the MP2 level, and introduce a MP2→ CCSD(T) correction with a smaller basis set (like that used in DFT calculations). Thus, taking the assumption that the difference between the CCSD(T) and MP2 interaction energies depends only negligibly on the basis set size (provided that the smallest basis set is already large enough), the CBS limit CCSD(T) interaction energies can be obtained as following:

$$E_{CBS}^{CCSD(T)} = \Delta E_{CBS}^{MP2} + (\Delta E^{CCSD(T)} - \Delta E^{MP2})_{small\ basis\ set} \qquad \text{Eq. 3.5.}$$

*CCSD(T)* standing for Coupled Cluster calculation with Single and Double excitations consider and perturbative treatment of Triple ones. It has been described (Sponer et al. 2008) that this protocol provides energies very close to the exact value in the case of closed shell systems.

## 3.4 RESP

As AMBER force fields charge model is transferable, individual residues can be used as building blocks to build larger protein and nucleic acid systems without the need to refit the charges for each model. This makes the charge derivation scheme an integral part of the force field parameterization. In the AMBER force field the method used for charge derivation is called the *Restrained Electrostatic Potential* (RESP) method. RESP is based on fitting an electrostatic potential (ESP), obtained from QM, while introducing restraints in form of a penalty function into the fitting procedure (Bayly et al. 1993). This multistage approach is very convenient for charge parameterization of modified residues as one can fit a specific region while keeping the rest restrained to AMBER default values. RESP calculations are typically performed using QM ESP computed from a quite modest level of theory (HF/6-31G(d)) due to the fortuitous fact that these calculations overestimate polarity in the ESP, which helps the pair-additive force field to correct part of the polarization effects, granting a good interface with currently used water models such as TIP3P or SPC/E.

## 3.5 Potential of mean force calculations

Potential of Mean Force (Kirkwood 1935) (explained in Chapter 2.7.) is a good technique for obtaining an energy profiles as a function of a coordinate of the system. It can be a geometrical coordinate or a more general energetic (solvent) coordinate. As sampling is crucial for correct PMF profile, PMF simulations are often used in conjunction with techniques that guarantee a smooth and continuous sampling of the conformational transition, in our case umbrella sampling.

Umbrella Sampling (Torrie & Valleau 1977) (see Chapter 2.7.) adds an artificial biasing potential ($V_b(\chi)_i$), mostly in a form of a harmonic restrain, that is moved systematically along the conformational coordinate. In that manner, US is defined by the strength of the added potential, *k*, length of the sampling, *t*, and number of reference point along the CV where the potentials will be places, *n*, also called windows. The strength of the potential directly effects the sampling time and number of windows, as a very strong potential will sample just a small area around the reference point, while a weak one will have problems converging and sampling some unfavorable areas. Thus, the choice of *k*, *n* and *t* differs depending on the system. As mentioned in Chapter 2.7, crucial points of US are proper sampling and the overlap between neighboring umbrellas, as that will be important in further analysis.

**WHAM**

As the umbrella potential introduces a bias in the population of different states along the CV, corrections need to be performed to generate unbiased populations from which free energies can be derived. From one US simulation, free energy can be expressed as

$$A(x) = -k_b T \, lnP'(x) - U'(x) + C \qquad \text{Eq. 3.6.}$$

where $U'(x)$ is the umbrella potential at a given point, $P'(x)$ is the probability of the state at that point, and *C* is an offset, a constant for up to which $P'(x)$ determines $A(x)$. Weighted Histogram Analysis Method (WHAM) (Kumar et al. 1992) method determines optimal *C* values for combining US simulations in an iterative self-consistent way yielding the free energy profile. In addition, WHAM can be easily extended to multi-dimensional PMFs (Boczko & Brooks 1993; Rajamani et al. 2003), which is quite useful for obtaining 2-dimensional PMF profiles (like in case of $\varepsilon/\zeta$ torsions' study).

## 3.6. MD preparation protocol

In order to properly simulate a system using MD one has to prepare the system using a protocol that guarantees the quality of the system to simulate and reduces the risks of equilibration problems. The MD protocol presented here was done using AMBER simulation package (i.e. Ambertools and AMBER MD engine). Following ABC recommendations (Lavery et al. 2010) and our own experience, we designed a multi-step procedure which provided good performance across all the studied systems:

i) The 3D structure of a given sequence was obtained experimentally, or generated using *NAB*, a software implemented into Ambertools (Roe & Cheatham 2013) that can generate a generic structure of a given sequence with any of the canonical helical forms.

ii) Once the desired structure is obtained and processed, the system has to be solvated and electroneutralized. Program called *tleap* is used for loading the necessary force field and library files, prior to **solvation and electroneutralization**. It also creates input topology (**.prmtop*) and coordinate (**.inpcrd*) files, necessary for running a MD simulation. To better mimic physiological condition, it is recommended to put an excessive salt concentration of 150 mM, which usually comes to an additional 12 pairs of ions for a system as big as a DNA dodecamer.

iii) The solvated system is then subjected to **energy minimization** to relax bad contacts. This process is done in two steps, first by minimizing just the solvent, keeping the DNA fixed, and secondly by minimizing the whole system.

iv) Following the minimization, the system needs to reach the desired ensemble conditions (temperature and pressure) by **heating** up the solvent while keeping the DNA fixed, usually from 0 K to 300 K in about 15000 steps, using NVT ensemble (controlling volume and temperature) with Langevin thermostat. Using NVT ensemble during heating is crucial because at low temperatures the calculation of pressure is inaccurate and can cause the barostat to overcorrect leading to instabilities (see previous chapter).

v) Next, the system needs to be equilibrated before data collection. The **equilibration** run is performed on the entire system prior to the actual production run in order to get stable temperature and pressure. In this step we use same conditions as in the production run, i.e. NPT ensemble with Berendsen bath. Equilibration is usually done for 1 to 10 ns, where after 1 ns most of the systems already reach stable temperature and pressure.

vi) Finally upon analysis of trajectories prove that equilibration is achieved, **production run** is performed, which will produce data that will later be analyzed. As mentioned previously, usual conditions include default temperature and pressure at 300 K and 1 atm, respectively, in a Berendsen bath, using 2 fs time step in conjunction with SHAKE (Ryckaert et al. 1977) (in case of AMBER) to constrain X-H bonds to the default values, and Ewald method (Darden et al. 1993) with default grid settings and tolerance. Conformational snapshots are usually saved every 1 or 10 ps depending on the length and the objective of the simulation.

## 3.7. Analysis

### RMSd – Root Mean Square Deviation

*RMSd* is a value that describes the structural difference between two structures, given that the structures are superimposed. In other words, RMSd quantifies the minimum deviation of a structure *X* from the reference structure *Y*:

$$RMSd = \min\left(\sqrt{\sum_{i=1}^{N}(\overrightarrow{x_i} - \overrightarrow{y_i})^2}\right) \qquad Eq.\,3.7.$$

where $\overrightarrow{x_i}$ and $\overrightarrow{y_i}$ are the coordinates of each of the $N$ selected atoms in structure *X* and in the reference structure *Y*, respectively, where minima is obtained from a simple least-squares fitting algorithm. RMSd is primarily used as the first indicator of structural stability along the simulation time, where for the reference point, we take an experimental structure or the first frame of the simulation (if there is no experimental structure). RMSd along time can indicate common thermal fluctuation (small *RMSd*) or conformational changes (for large RMSd values), which might be real, a problem in the starting structure, or an artifact of the simulation.

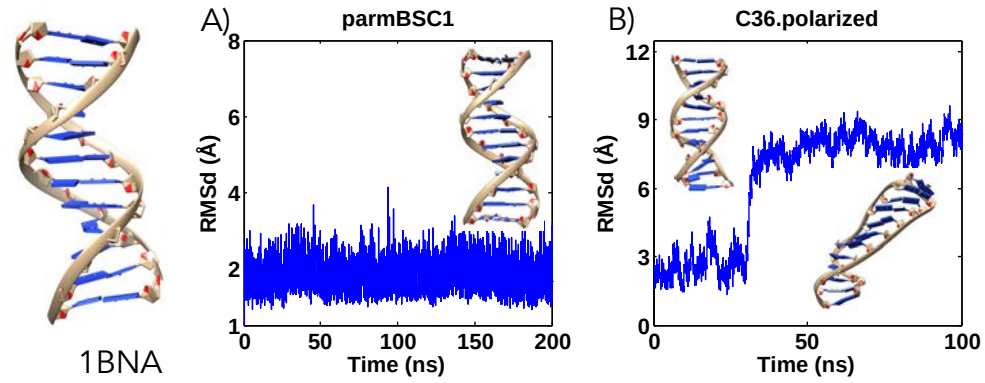



**Figure 3.2. Examples of the RMSd plot** in the reference to an experimental structure (PDB ID: 1BNA) showing stability (A) and conformational change (B), in this particular case a force field artifact.

### RMSF – Root Mean Square Fluctuation

*RMSF* quantifies local changes along the DNA backbone. In contrary to RMSd, RMSd gives the average fluctuations over time for each atom *i*:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\vec{x_i}(t) - \langle\vec{x_i}\rangle)^2} \qquad Eq.\,3.8.$$

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

### Hydrogen bonds

Hydrogen bonds are particular type of attractive electrostatic interaction occurring when a hydrogen atom H- covalently bound to an electronegative atom (often oxygen or nitrogen atoms), called hydrogen donor (D), approach another electronegative atom, called hydrogen acceptor (A). The hydrogen bond free energy content is between 1 – 5 kcal/mol, making it stronger than van der Waals interaction, but weaker than covalent or ionic bond. As the strength of the hydrogen bond is sensitive to atoms' orientation and distance, MD analysis of hydrogen bonds are usually counted by a defined cut-off distance between donor and acceptor, in our case 3.5 Å, and cutoff angle, in our case 120º. The analysis includes extracting distance between potential donor and acceptor and checking if they fulfill the hydrogen bond criteria described above. Furthermore, we counted hydrogen bond break if the bond was lost for consecutive 10 picoseconds (Dillon 2012).

## PCA – Principle component analysis

Since the majority of motion in a MD simulation comes from local thermal fluctuations it is not an easy task to understand the dynamics of a bimolecular system just from looking at the trajectory. **Principle component analysis (PCA)** is a statistical procedure that converts a set of data of possibly correlated variables into a set of linearly uncorrelated variables called *principle components*. In MD, PCA helps extracting *essential* motions from irrelevant thermal fluctuation, thus in MD world PCA is also known as *Essential Dynamics* (Amadei et al. 1993; Orozco et al. 2003). This process is done by means of diagonalization of the Cartesian covariance matrix *C* containing atomic positional fluctuations in all 3 coordinate axes

$$C = \langle \Delta X \Delta X^T \rangle \qquad \text{Eq. 3.9.}$$

where $\Delta X$ is the difference between the position and a reference value (usually a MD average structure). Therefore, the transformation matrix $A$ for the diagonalization of the covariance provides the diagonalized correlation matrix $\Lambda$

$$\Lambda = A^T C\, A \qquad \text{Eq. 3.10.}$$

which contains eigenvalues $\lambda$, where *n-th* column of the transformation matrix $A$ corresponds to the eigenvector with the eigenvalue $\lambda_n$. The eigenvector provides information on the nature of the essential movement, and the eigenvalue on its impact in explaining the sampled variance.

Essential dynamics has been deeply used to characterize the low-frequency movements of DNA (Pérez et al. 2005), and to determine similarity between trajectories. For this particular purpose, I used standard Hess metrics (based on accumulated dot products, (Hess 2000)) as well as energy-corrected Hess metrics (Pérez et al. 2005) which correct the metrics for the different energy of the essential movements under comparison.

## Entropy

*Molecular entropy* can be extracted by using pseudo-harmonic models following either Andricioaei & Karplus' (Andricioaei & Karplus 2001) (Eq. 3.11.) or Schlitter's (Schlitter 1993) (Eq.3.12.) methods. Two approaches rely on the principle of the quantum harmonic oscillator and both require the diagonalization of the mass-weighted covariance matrix to obtain the frequencies associated with the essential deformations.

$$S = k \sum_i \frac{\alpha_i}{e^{\alpha_i} - 1} - ln(1 - e^{-\alpha_i}) \qquad \text{Eq. 3.11.}$$

$$S \approx \frac{1}{2}k\sum_{i} ln\left(1+\frac{e^2}{\alpha^2}\right) \qquad \text{Eq. 3.12.}$$

where $\alpha_i = \hbar\omega/kT$, with eigenvalues, $\omega$, obtained by diagonalization of the mass-weighted covariance matrix.

Both methods produce very similar results and share the same shortcoming; they can be used in trajectories near the equilibrium but not in those that sample irreversible conformational transitions, i.e., when macromolecular movement is very far from the harmonic regime. Time dependence is intrinsic to entropy calculations, since the number of visited microstates increase by the length of the trajectory. This problem can be avoided by using the an exponential correction formula developed by (Harris et al. 2001).

$$S(t) = S_\infty - \frac{\alpha}{t^\beta} \qquad \text{Eq. 3.13.}$$

where $\alpha$ and $\beta$ are fitted parameters and $t$ is the simulation time.

With nowadays simulation times, time-dependence problem is almost neglectable (see Figure 3.3) if macromolecule moves in the (pseudo)harmonic regime.

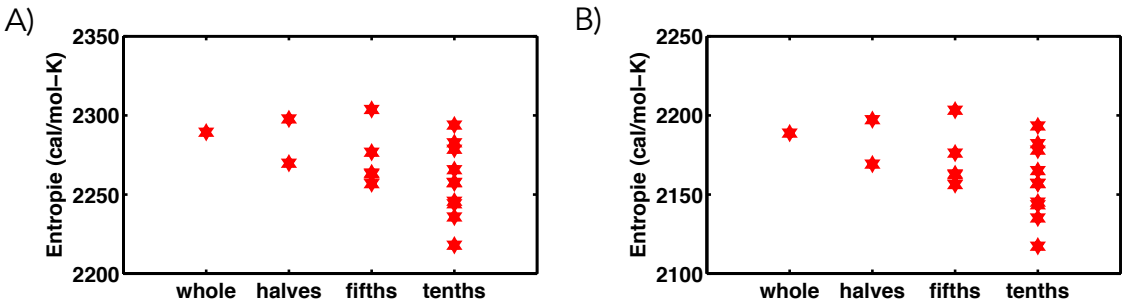



**Figure 3.3. Entropy calculations** from 10 μs simulation of DDD (see Publication #2 in the Results) using (A) Schlitter's and (B) Andricioaei & Karplus' methods, showing the convergence of entropy over time. The trajectory was divided into tenths, fifths and halves.

## 3.8. Helical analysis

DNA is easier to study in the helical space (see Chapter 1.2., Olson et al. 1998; Lankaš et al. 2000; Noy et al. 2004). The helix axis is defined in the direction of propagation of the helix (not necessarily a straight line). A base-pair is defined by a set of 10 base-pair parameters (for example opening, describing the fraying of terminal residues) affecting more local properties, and 6 base-pair parameters (from which twist was studied in greater extent in this thesis), which affect more global

properties. In MD analysis, helical parameters, as well as backbone and puckering analysis was done using Curves+ program (Lavery et al. 2009). Curves+ provides statistical averages with corresponding standard deviations, distributions and evolution of the helical parameters along the trajectory or set of structures. Additionally, Curves+ newest CANION module calculates the position of each cation in curvilinear helicoidal coordinates with respect to the helical axis. This feature is very useful in a study of ion distribution or environmental impacts on the base pair step helical characteristics.

When the oscillation of helical parameters sampled during a MD simulation is plotted, a clear Gaussian shape is typically obtained, which suggest that DNA often behaves following a stiffness model, where harmonic variables are helical base pair step parameters. Thus, by inverting the covariance matrix in helical space, one obtains the stiffness matrix $\Xi$ defining the flexibility of DNA with the respect to the deformation along the helical inter-base pair coordinates is obtained:

$$\Xi = k_B T C^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{rw} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{tw} & k_{sr} & k_t & k_{ts} & k_{tl} & k_{tf} \\ k_{sw} & k_{sr} & k_{st} & k_s & k_{sl} & k_{sf} \\ k_{lw} & k_{lr} & k_{lt} & k_{ls} & k_l & k_{lf} \\ k_{fw} & k_{fr} & k_{ft} & k_{fs} & k_{fl} & k_f \end{pmatrix} \qquad \text{Eq. 3.14.}$$

where *k* stands for stiffness constant with diagonal terms ($k_i$) corresponding to pure helical stiffness constants, whereas non-diagonal ones ($k_{ij}$) correspond to coupled terms. Note that *w* stands for *twist, r* for *roll, t* for *tilt, s* for *rise, l* for *slide* and *f* for *shift* (for the description of helical parameters see **Chapter 1.2**).

## 3.9. Experimental observables

Direct comparison between experimental and simulated structures is a risky approach as “experimental structures” are really just models that reproduce some experimental observables obtained, in some cases, under conditions far from the physiological ones. For example, extreme caution is needed when comparing simulations in solution with X-Ray structures as they can be contaminated by severe lattice artifacts. For the case of NMR, the problem is that often, experimental restrains do not define in a unique way the structure. So, a good approach to evaluate the quality of a trajectory is to estimate “observables” from it and compare them with those really determined experimentally. For the case of NMR it is possible to compare directly *J-couplings*, *NOEs* and *RDCs* from simulation with those collected in the NMR equipment.

NOEs occur through spatial cross-relaxation of nuclear spin polarization, quantitatively describing the distance of protons between atoms within 1.8 to 6 Å threshold with the strength of the NOE signal depending only on the spatial proximity of protons. A set of NOE distances (usually between 1H - 1H) can be compared with MD-averaged distances to obtain an average violation, largest violation and number of violations (calculated by defining a cut-off, usually 0.5 Å). Furthermore, the analysis can be divided based on the strength of NOE signals, with strong NOEs defined at < 3.5 Å and weak NOEs at > 5 Å. This can eliminate experimental bias, as weaker NOEs are typically of a lower resolution. To calculate NOE violation a very useful tool called *g_disre* from the GROMACS package can be used.

RDCs (see Chapter 1.4) describe the orientation of internuclear vectors with respect to the external magnetic field, providing spatially and temporally averaged information. RDCs are observed in the solution when a molecule is aligned with the external magnetic field. The complication in assigning the experimental observables lies in defining the alignment tensor. In order to obtain the alignment tensor that best fit the observed RDCs (typically between C-H and N-H), we used the program PALES (Zweckstetter 2008), which accurately predicts molecular alignment tensor, and thus RDCs, based purely on the molecular structure. This comes useful for direct comparison of MD obtained structures with experimental observables, where RDCs are computed for the structures obtained from MD and then averaged for comparison. For quantification of the comparison between experimentally observed and calculated RDCs, we can define a quality factor (Q-factor) like,

$$Q = \frac{\sqrt{\sum (RDC_{calc} - RDC_{exp})^2}}{\sqrt{\sum RDC_{exp}^2}} \qquad \text{Eq. 3.15.}$$

Lower Q-factor indicates better agreement with experimentally observed residual dipolar couplings. See Supplementary Figure 7 of Publication #1 in the **Results** section for more details on this topic.

# BIBLIOGRAPHY TO CHAPTER 3

Amadei, A., Linssen, A. & Berendsen, H.J.C., 1993. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4), pp.412–425.

Andricioaei, I. & Karplus, M., 2001. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics*, 115(14), pp.6289–6292.

Bachs, M., Luque, F.J. & Orozco, M., 1994. Optimization of solute cavities and van der Waals parameters in ab initio MST-SCRF calculations of neutral molecules. *Journal of computational chemistry*, 15(4), pp.446–454.

Bayly, C.I. et al., 1993. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40), pp.10269–10280.

Becke, A.D., 1988. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6), pp.3098–3100.

Boczko, E.M. & Brooks, C.L., 1993. Constant-Temperature Free-Energy Surfaces for Physical and Chemical Processes. *Journal of Physical Chemistry*, 97(17), pp.4509–4513.

Cances, E., Mennucci, B. & Tomasi, J., 1997. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of Chemical Physics*, 107(8), pp.3032–3041.

Darden, T., York, D. & Pedersen, L., 1993. Particle mesh Ewald: An N · log (N) method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12), pp.10089–10092.

Dillon, P.F., 2012. *Biophysics: A physiological approach*,

Frisch, M.J. et al., 2004. Gaussian 03, revision c. 02; Gaussian. *Inc., Wallingford, CT*, 4.

Frisch, Mj. et al., 2009. Gaussian 09, revision A. 02; Gaussian, Inc. *Wallingford, CT*, 19, pp.227–238.

Halkier, A. et al., 1998. Basis-set convergence in correlated calculations on Ne, N 2, and H 2 O. *Chemical Physics Letters*, 286(3), pp.243–252.

Halkier, A. et al., 1999. Basis-set convergence of the energy in molecular Hartree–Fock calculations. *Chemical physics letters*, 302(5), pp.437–446.

Harris, S.A. et al., 2001. Cooperativity in drug-DNA recognition: a molecular dynamics study. *Journal of the American Chemical Society*, 123(50), pp.12658–12663.

Hess, B., 2000. Similarities between principal components of protein dynamics and random diffusion. *Physical Review E*, 62(6), p.8438.

Kirkwood, J.G., 1935. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5), pp.300–313.

Krishnan, R. et al., 1980. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *The Journal of Chemical Physics*, 72(1), pp.650–654.

Kumar, S. et al., 1992. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8), pp.1011–1021.

Lankaš, F. et al., 2000. Sequence-dependent elastic properties of DNA. *Journal of*

*molecular biology*, 299(3), pp.695–709.
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 research*, 38(1), pp.299–313.
Lavery, R. et al., 2009. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic acids research*, 37(17), pp.5917–5929.
Lee, C., Yang, W. & Parr, R.G., 1988. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2), pp.785–789.
Marenich, A. V, Cramer, C.J. & Truhlar, D.G., 2009. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B*, 113(18), pp.6378–6396.
Miertuš, S., Scrocco, E. & Tomasi, J., 1981. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics*, 55(1), pp.117–129.
Miertus, S. & Tomasi, J., 1982. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chemical physics*, 65(2), pp.239–245.
Noy, A. et al., 2004. Relative flexibility of DNA and RNA: a molecular dynamics study. *Journal of molecular biology*, 343(3), pp.627–638.
Olson, W.K. et al., 1998. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19), pp.11163–11168.
Orozco, M. et al., 2003. Theoretical methods for the simulation of nucleic acids. *Chemical Society Reviews*, 32(6), pp.350–364.
Orozco, M. & Luque, F.J., 2000. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chemical Reviews*, 100(11), pp.4187–4226.
Pérez, A. et al., 2005. Exploring the essential dynamics of B-DNA. *Journal of Chemical Theory and Computation*, 1(5), pp.790–800.
Rajamani, R., Naidoo, K.J. & Gao, J., 2003. Implementation of an Adaptive Umbrella Sampling Method for the Calculation of Multidimensional Potential of Mean Force of Chemical Reactions in Solution. *Journal of Computational Chemistry*, 24(14), pp.1775–1781.
Roe, D.R. & Cheatham, T.E., 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, 9(7), pp.3084–3095.
Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.C., 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3), pp.327–341.
Schlitter, J., 1993. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chemical Physics Letters*, 215(6), pp.617–621.
Sponer, J., Riley, K.E. & Hobza, P., 2008. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys Chem Chem Phys*, 10(19), pp.2595–2610.
Torrie, G.M. & Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2), pp.187–199.
Truhlar, D.G., 1998. Basis-set extrapolation. *Chemical Physics Letters*, 294(1–3), pp.45–48.

Woon, D.E. & Dunning Jr, T.H., 1993. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *The Journal of chemical physics*, 98(2), pp.1358–1371.

Zweckstetter, M., 2008. NMR: prediction of molecular alignment from structure using the PALES software. *Nature protocols*, 3(4), pp.679–690.

Barcelona, 29 Sep 2015

Comisión de doctorado.

El Sr. Ivan Ivani ha estado realizando su tesis doctoral durante los últimos años en el IRB bajo mi dirección. El trabajo de su tesis doctoral se verá reflejado en una serie de publicaciones científicas algunas aún en fase de redacción, dos que se ha sometido a revisión muy recientemente y tres ya publicadas.

1. Ivani, Ivan, Pablo D. Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther et al. "Parmbsc1: a refined force field for DNA simulations." Nature methods 13, no. 1 (2016): 55-58.
2. Pablo D., Dans, Linda Danilāne, Ivan Ivani, Tomáš Dršata, Filip Lankaš, Jürgen Walther, Ricard Illa Pujagut et al. "Long-timescale dynamics of the Drew–Dickerson dodecamer." Nucleic acids research (2016): gkw264.
3. Montserrat Terazzas, Ivan Ivani, Nu´ria Villegas, Clement Paris, Candida Salvans, Isabelle Brun-Heath and Modesto Orozco. “Rational design of novel N-alkyl-N capped biostable RNA nanostructures for efficient long-term inhibition of gene expression”. Nucleic acids research (2016): gkw169

Nature Methods es una de las revistas multidisciplinares de más impacto con un IF de 32. El artículo en concreto, publicado en este año tiene ya más de 10 citaciones, por encima de la media de la revista. Nucleic acids research, es la revista de referencia en biología de acidos nucleicos, con un IF cercano al 10. Los artículos enviados o en proceso de escritura apuntan también a revistas de alto impacto.

En dos artículos el Sr. Ivani es el primer firmante y responsable de haber realizado y discutido la gran mayoría de trabajo, y en los otros tres el Sr. Ivani fue responsable de calculos importantes para el trebajo. Los trabajos de su tesis no forman parte de otra tesis doctoral.

Prof. Modesto Orozco
Modesto.orozco@irbbarcelona.org

*"Every time you understand something, religion becomes less likely. Only with the discovery of the double helix and the ensuing genetic revolution have we had grounds for thinking that the powers held traditionally to be the exclusive property of the gods might one day be ours. . . ."*
*James D. Watson*

# 4| PARMBSC1

DNA overall stability is defined by the equilibrium between two opposite forces: strong electrostatic repulsion between the phosphate in the backbone, and stacking and hydrogen bonding between nucleobases. Additionally, solvent interactions tune these two forces and indirectly affect the shape of DNA double helix. Representation of DNA requires then to balance strong opposite interactions, something very challenging for simple molecular mechanics force fields.

Even the first MD simulation of nucleic acids, which should be credited to Michael Levitt (Levitt 1983) show problems keeping the structure stable (in the ps-regime). Large magnitude of the nucleotide-nucleotide interactions, especially of the phosphate-phosphate repulsion, which generates strong forces yielding to instabilities in the structure, were quite difficult to address in early days of DNA simulation. The introduction of explicit solvent, an accurate description of the long-range electrostatic contribution (PME; see Chapter 2.2.) coupled to a new generation of force fields, yielded stable structures in the nanosecond time scale (for a historic view on the evolution of MD simulations of nucleic acids we address the reader to previous reviews (Orozco et al. 2003; Dans et al. 2016)), opening the possibility to use MD simulations to study aspects of nucleic acids difficult to access from experiments.

A clear breakthrough in MD simulations of nucleic acids were the force fields developed in the late 90's by the group of Peter Kollman (Cheatham et al. 1995; Cheatham III et al. 1999), as they allowed, for the first time, to collect equilibrated trajectories which sampled conformations of DNA and RNA not far from the experimental ones. However, later increase in computer power extended the MD simulation time, uncovering the existence of some force field errors which yielded to the corruption of structures in ~10 ns time scale (Beveridge et al. 2004; Pérez, Marchán, et al. 2007). Namely, big distortion in the structure were observed, with massive $\alpha/\gamma$ transition to *gauche+/trans* geometry, away from canonical *gauche-/gauche+* state (Várnai & Zakrzewska 2004), giving "ladder-like" structures of DNA duplexes. Reparameterization efforts yielded to force fields such as *parmbsc0* (Pérez, Marchán, et al. 2007), a force field developed in our group, which corrected these problems producing stable canonical structures even in a microsecond simulation run (Pérez, Luque, et al. 2007). Since then, parmbsc0 has been extensively used to simulate a variety of nucleic acids in the multi-nanosecond and sub-microsecond timescale (Pérez et al. 2012), producing more than 1000 citation up to date.

However, as simulation systematically allowed us to approach to the microsecond regime, some errors in the last generation force field become evident. For example, a possible underestimation of $B_{II}$ state and $B_I \leftrightarrow B_{II}$ equilibrium for B-DNA (Heddi et al. 2006; Heddi et al. 2008), a systematic underestimation of twist compared with NMR or X-Ray data (see Figure 4.1.), a slightly biased representation of puckering to the East state, problems with the glycosidic torsion (Perez et al. 2008), which mainly affects the representation of some exotic conformations of DNA (Fadrná et al. 2009; Krepl et al. 2012), and introduced excessive distortions at the ends of the canonical B-type duplexes in very long simulations, which generate severe end-effects (Drsata et al. 2012).

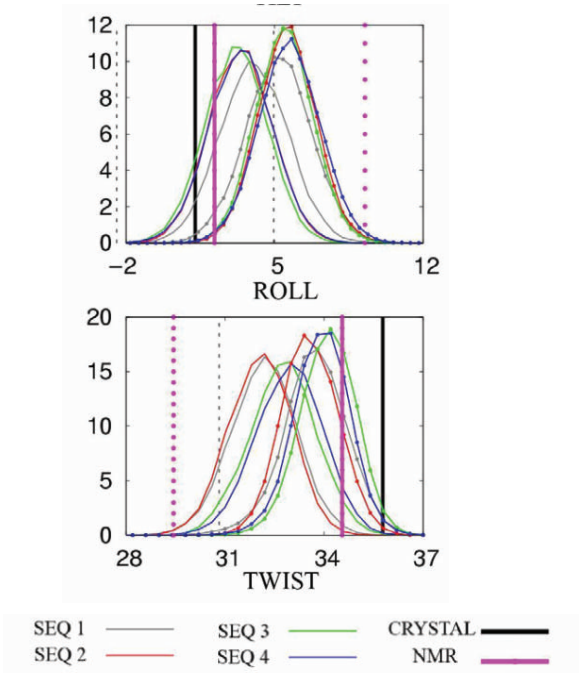



**Figure 4.1. Problems in helical parameters of parmbsc0.** Distribution of twist and roll values coming from MD simulations of 4 sequences using parmbsc0 force field showing clear undertwisting of structures [taken from (Perez et al. 2008)].

To address these problems, we started systematically reparameterizing parmbsc0 force field in aspects of sugar puckering, ε, ζ and χ torsions using high-level QM calculations both in gas phase and solution. The refined force field, called *parmbsc1*, has been tested for more than 4 years to an unprecedented level of detail, considering a large variety of DNAs, and analyzing structural, mechanical and dynamical properties of the DNAs resulting from the corresponding MD simulations (see **Chapter 4.1**). The force field was available for the community already in 2014, and being used (without major errors detected) for one year before it was sent for publication (appeared in January 2016).

To validate our new parameter set, we performed a systematic study of the long-timescale dynamics of the Drew–Dickerson dodecamer (DDD) a prototypical B-DNA duplex, describing the conformational landscape of DDD in a variety of ionic environments from minimal to high salt concentrations. We also explored the sensitivity of the simulations to the use of different solvent and ion models. Finally, an extended (10 μs) simulation is used to characterize slow and infrequent conformational changes in DDD. The analysis of terminal residues showed an almost complete absence of hydrogen bonding between the base the sugar, an artifact that caused a great deal of problem in simulations using parmbsc0 (see **Chapter 4.2**).

In the meantime, different patches have been developed to correct parmbsc0 errors, mostly from a Czech consortium. For example, OL1 (Zgarbová et al. 2013) was created to improve ε/ζ representation, and accordingly to reproduce better the $B_I/B_{II}$ representation, OL4 (Krepl et al. 2012) patch had the objective to correct χ conformation for DNA. Other authors used harmonic restraints derived from NMR measures to guarantee a good representation of the $B_I/B_{II}$ equilibrium (Heddi et al. 2008). It is worth to note, that all these patches are specific for DNA, and RNA-patches for parmbsc0 have been developed in parallel, by means of the reparametrization of χ angle (Zgarbová et al. 2011), or even by scaling down van der Waals interactions (Chen & García 2013), which will be discussed later. The introduction of patches has allowed the AMBER-community, to use MD to study exotic forms of DNA, but has produced also a notable confusion in the field, since it is not always clear when these patches should be used, or whether or not they can be combined. At the end of last year, the Czech group published the latest force field (called OL15), which incorporated all the previous OL corrections for DNA and included additional correction of the β torsion (Zgarbová et al. 2015). To give a better look into the difference between force fields' performance, we benchmarked all important corrections of parmbsc0 comparing the results with *de novo* NMR experiments (see **Chapter 4.3**).

## 4.1 Parmbsc1 – development of a "state of the art" DNA force field (Publication 1)

At the beginning of my Ph.D, *parmbsc0* force field, developed in the group, was already established as the gold-standard in the field of nucleic acids simulations. But, as mentioned previously, errors connected with the force field had become more evident, thus we decided to address these problems by reparameterizing parmbsc0 using high-level QM data. As described in Section 2.5, main problems coming from parmbsc0 simulations included: inability to reproduce experimental values of helical parameters (especially twist and roll), improper reproduction of bimodality for some base-pair steps and underestimation of $B_{II}$ population, excessive terminal fraying of DNA duplexes, and difficulties in reproducing non-canonical structure. We have found that wrong description of several torsions was the cause of these glitches.

Firstly, we addressed the problem of underestimation of $B_{II}$ state and $B_I \leftrightarrow B_{II}$ equilibrium, and under-twisting of the canonical structures. As I mentioned in Chapter 1.2, $B_I$ and $B_{II}$ are connected with *high-twist* and *low-twist* states of the base-pair step, thus by proper reproduction of the $B_I \leftrightarrow B_{II}$ equilibrium we would get a better description of twist, and consequently roll (as they are coupled). By the definition, the $B_I$ and $B_{II}$ states are defined in terms of $\varepsilon/\zeta$ torsions, more specifically $B_I$ being in $\varepsilon/\zeta$=*trans/gauche-* region and $B_{II}$ being in $\varepsilon/\zeta$=*gauche-/trans* region. Thus, our idea was that by reproducing the coupled $\varepsilon/\zeta$ profile, we would directly correct $B_I \leftrightarrow B_{II}$ equilibrium problems and indirectly affect twist and roll (as they are coupled). We performed QM scan of $\varepsilon/\zeta$ energy landscape on MP2/aug-cc-pVDZ level (see Chapter 3.1) with SCFR solvent corrections, with additional CCSDT(Q)/CBS calculations on three point of interest, $B_I$ and $B_{II}$ minima and the maxima in between them, incorporating those points into the energy profile with higher weight during fitting (see Supplementary Figure 26).

Additionally, we performed scans of $\chi$ torsion for 4 DNA bases. The goal was to check the correctness of MM profiles, especially *syn/anti* equilibrium. The idea in this step was that, as the *syn* conformation plays a significant role in non-canonical DNA structures (see Chapter 1.2), reproducing $\chi$ profiles would allow a better description of non-canonical DNA structures in MD simulations, and probably positively affects the excessive terminal fraying. Similarly like in case of $\varepsilon/\zeta$, we performed high-level QM scan with additional CBS points on *syn* and *anti* minima, and the barrier between the minima (see Supplementary Figure 24). The *syn/anti* equilibrium is now well reproduced, forbidding transitions between the two in the direction of *high-anti*, a flaw in profiles of parmbsc0.

We then incorporated the two correction and tested them on a small duplex $d(CGATCG)_2$. We noticed small imperfections with puckering profiles, most probably because the two corrections were done in an independent way, and the sugar ring is the direct connection between the three torsions. For that reason, we performed

a scan of the pseudorotational angle in similar approach as the previously parameterized torsions, doing CBS calculations for *North* and *South* minima, and *East* barrier. In comparison with parmbsc0, we increased the East barrier and displaced the minima into more realistic regions (see Supplementary Figure 25).

Incorporating these three corrections into one force field, that we called *parmbsc1*, we proceeded to assess its performance by testing it on more than a hundred DNA structures (see Supplementary Table 1) for a large variety of DNA motifs. From an accumulative of ~140 μs we obtained unprecedented results in diverse systems ranging from canonical B-DNA, various non-canonical forms, other unusual DNA configurations like triplexes and quadruplexes, to DNA in complexes with proteins and ligands, and in various conditions.

We performed an extensive study of the most known B-DNA structure, Drew-Dickerson dodecamer (DDD), where parmbsc1 yielded significant improvements in comparison with its predecessor, sampling a stable B-DNA duplex that remained close to the experimental structures, and preserving hydrogen bonds and helical parameters even at the terminal residues (see Figure 1 in the following publication). The average twist value is improved by 1.5º, with the average roll values dropping by 1.2º and an increase in $B_{II}$ population by 7% (see Supplementary Table 2 and Supplementary Figure 3 in the following publication). Moreover, the improvements of helical parameters did not defect the shape of helical parameter profiles, where the average sequence-dependent helical parameters (see Figure 1 and Supplementary Figures 1 and 2 in the following publication) matched experimental values, especially those derived in aqueous solution. Parmbsc1 was also able to capture some unique characteristics of A-track sequences, such as strong propeller twist, smaller slide, higher inclination and narrowing of the minor groove (see Supplementary Figures 4-6 in the following publication).

Besides the universal experimental validation by the means of structural comparison, we computed direct experimental observables, RDCs and NOEs, for several structures with the available data. In the case of DDD, the success metrics of the reproduced observable are similar to those obtained in the NMR-refined structures (see Supplementary Table 3 in the following publication). Additionally, our new force field was able to obtain NOE violations statistics equivalent to those determined from “de novo” NMR-derived ensembles collected by our collaborator, Carlos Gonzalez, after parmbsc1 was developed (see Supplementary Table 8 in the following publication). In collaboration with David Case’s group, we reproduced the structures of DNA in crystal environments with an improvement on previous simulations done with parmbsc0 force field (see Supplementary Figures 14 and 15 in the following publication).

Simulations of various non-canonical systems showed significant improvements

for all structures (see Figure 2 in the following publication). Thus, parmbsc1 was not only able to sample stable structure, but also recognized experimentally known unstable structures, like Z-DNA under 4M salt concentration, or antiparallel triplex under physiological conditions (see Figure 2 in the following publication). We also focused on DNA under stressed conditions, particularly on 4 DNA-protein complexes and 2 drug-DNA complexes, finding excellent agreement with experiments (see Figure 3 and Supplementary Figures 16 and 17 in the following publication).

Very recently, we also performed a small benchmark comparing parmbsc1 performance with other modern force fields in simulating DDD (see Supplementary Table 2 and Supplementary Figure 29 in the following publication). In the benchmark, parmbsc1 clearly out-performs all other force fields available (see Supplementary Figure 30 in the following publication). Just a force field developed after parmbsc1 was available (OL15) provides results of similar quality than parmbsc1.

Looking at DNA flexibility, we obtained persistence lengths values in the range of 40 to 57 nm (see Supplementary Table 11 in the following publication), close to the generally accepted value is 50 nm. On the dynamics part, we reproduced the spontaneous A-to-B DNA transition in water and stable A-DNA form of a duplex in 85%-15% ethanol-water mixture (see Supplementary Figure 21 in the following publication). Finally, parmbsc1 reproduced previous successful simulations of unfolding of a duplex in a 4 M pyridine solution (see Supplementary Figure 21 in the following publication), and folding a small DNA hairpin motif in water (see Supplementary Figure 22 in the following publication).

Overall, our conclusions from this extensive work are that undoubtedly parmbsc1 is a clear improvement of parmbsc0 force field providing a good representation of static and dynamic properties of DNA. We believe that parmbsc1 will become the reference force field for DNA simulations under various conditions. More details on the comparison with other force fields will be given in the following publications (see **Chapter 4.3**).

# 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[9,10], Sarah A Harris[3], David A Case[11] & Modesto Orozco[1,2,7]

**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 timescales: as MD trajectories are extended to longer timescales, errors previously undetected in short simulations emerge, creating the need to improve the force field[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], fueling the development of parmbsc0 (ref. 4), which in turn started to show deviations from experimental data in the microsecond regime (for example, underestimation of twist, deviations in sugar puckering, biases in $\varepsilon$ and $\zeta$ torsions, excessive terminal fraying[2,5] and severe problems in representing certain noncanonical DNAs[1,6]). Various force-field modifications have been proposed to address these problems, such as the Olomouc ones[5,6] designed to reproduce specific forms of DNA. Although 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 the parmbsc1 force field presented here to address these needs, with the aim of creating a general-purpose force field for DNA simulations. We assessed its performance by testing its ability to simulate a wide variety of DNA systems (**Supplementary Table 1**).

Parmbsc1 was able to fit quantum mechanical (QM) data well (**Supplementary Discussion**), improving on previous force-field results (Online Methods and **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 a benchmark in force-field development. 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 with other force fields[2,8] (Online Methods and **Supplementary Discussion**). The average sequence-dependent helical parameters (**Fig. 1** and **Supplementary Figs. 1** and **2**) and BI and BII conformational preferences (**Supplementary Table 2** and **Supplementary Fig. 3**) matched experimental values (comparisons with estimates obtained with other force fields are presented in the Online Methods). Furthermore, parmbsc1 reproduced residual dipolar couplings ($Q$-factor = 0.3) and the nuclear Overhauser effect (NOE; only two violations), yielding success metrics similar to those obtained in the NMR-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 (r.m.s. deviation 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 of 33.9° and 2.5°, respectively) matched values from analyses of the RCSB Protein Data Bank (PDB) (33.6° and 2.9°) (ref. 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, and we obtained 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**).

[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, IRB Barcelona, 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, New York, 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', Consejo Superior de Investigaciones Cientificas, Madrid, Spain. [9]School of Pharmacy, University of Nottingham, Nottingham, UK. [10]Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK. [11]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey, USA. Correspondence should be addressed to M.O. (modesto.orozco@irbbarcelona.org).

RECEIVED 5 MAY; ACCEPTED 22 SEPTEMBER; PUBLISHED ONLINE 16 NOVEMBER 2015; DOI:10.1038/NMETH.3658

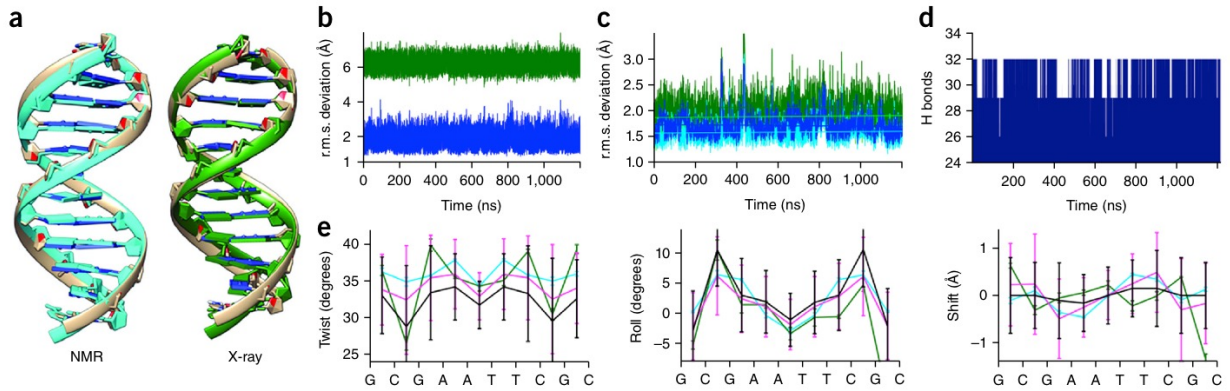



**Figure 1** | Analysis of the DDD. (**a**) Comparison of the MD average structure (light brown) with the NMR structure (light blue) (PDB ID 1NAJ) and the X-ray structure (green) (PDB ID 1BNA). (**b**) r.m.s. deviation of 1.2-µs trajectory of DDD compared with that for the B-DNA (blue) and A-DNA (green) forms (from the standard geometries derived from fiber diffraction (Online Methods)). (**c**) r.m.s. deviation of parmbsc1 data compared to experimental X-ray (green) and NMR (blue) structures (with (dark) and without (light) ending base pairs). Linear fits of all r.m.s. deviation curves are plotted on top. (**d**) Evolution of the total number of hydrogen bonds formed between base pairs in the whole duplex. (**e**) Comparison of average values of helical rotational parameters (twist, roll and shift) per base-pair step coming from NMR (cyan), X-ray (green), 1-µs parmbsc0 trajectory[2] (black) and 1.2-µs parmbsc1 trajectory (magenta) data. Error bars denote ±s.d.



We also explored the ability of parmbsc1 to represent unusual DNA configurations, such as a Holliday junction, a complex duplex-quadruplex structure, which was fully preserved in microsecond-long trajectories (**Supplementary Figs. 8** and **9**), or 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 concentrations[11]. For Hoogsteen 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[13] (**Supplementary Fig. 11**). Parmbsc1 simulations not only reproduced the known structure of parallel d(T-A·T) and d(G-G·C) triplexes (**Fig. 2c**,**d**) but also showed correctly that the equivalent antiparallel structures are unstable in normal conditions[14] (**Fig. 2e**). Finally, parmbsc1 was able to reproduce experimental structures of both parallel and antiparallel DNA quadruplexes with r.m.s. deviation of <2 Å (**Fig. 2f**,**g**).

We also explored the ability of parmbsc1 to reproduce the complex conformations of hairpins and loops, exceptionally challenging structures for force fields[15]. We performed microsecond simulations of the d(GCGAAGC) hairpin (PDB ID 1PQT), the 4T-tetraloop in *Oxytricha nova* quadruplex d$(G_4T_4G_4)_2$ (OxyQ; PDB ID 1JRN) and the junction loops in the human telomeric quadruplex (HTQ; PDB ID 1KF1). Parmbsc1 provided excellent representations (r.m.s. deviation of ~1 Å) of the d(GCGAAGC) hairpin (**Fig. 2h**) and OxyQ (**Fig. 2i**). For the very challenging HTQ structure, parmbsc1 maintained the stem structure 20 times longer than in previous

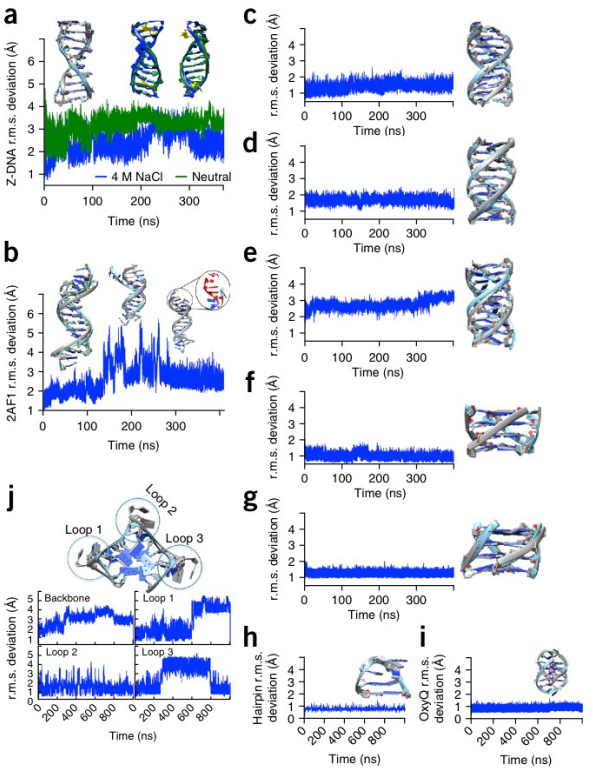



**Figure 2** | Analysis of noncanonical DNA structures. (**a**) Comparison of Z-DNA (PDB ID 1I0T) simulations in neutral conditions and in 4 M NaCl. Structural comparisons at different time points are shown above the r.m.s. deviation curves. (**b**) Simulation of anti-parallel Hoogsteen DNA (PDB ID 2AF1) showing deviation of the structure over time (highlighted in red). (**c**–**e**) r.m.s. deviation of (**c**) parallel d$(T\text{-}A\cdot T)_{10}$, (**d**) parallel d$(G\text{-}G\cdot C)_{10}$ and (**e**) antiparallel d$(G\text{-}G\cdot C)_{10}$ triplexes. (**f**,**g**) Parallel (**f**) (PDB ID 352D) and anti-parallel (**g**) (PDB ID 156D) quadruplexes showed stable structures over time. (**h**) Structural stability of d(GCGAAGC) hairpin (PDB ID 1PQT) and (**i**) OxyQ (PDB ID 1JRN) with ions over time. (**j**) HTQ (PDB ID 1KF1) with highlighted loops. r.m.s. deviations of HTQ backbone, loop 1, loop 2 and loop 3 regions are shown below. In all panels, parmbsc1 structures (light blue; final, averaged or at a given trajectory point) overlap the experimental structure (gray) for comparison. Green shading in structures denotes Z-DNA. **Supplementary Table 1** presents information on the PDB structures.

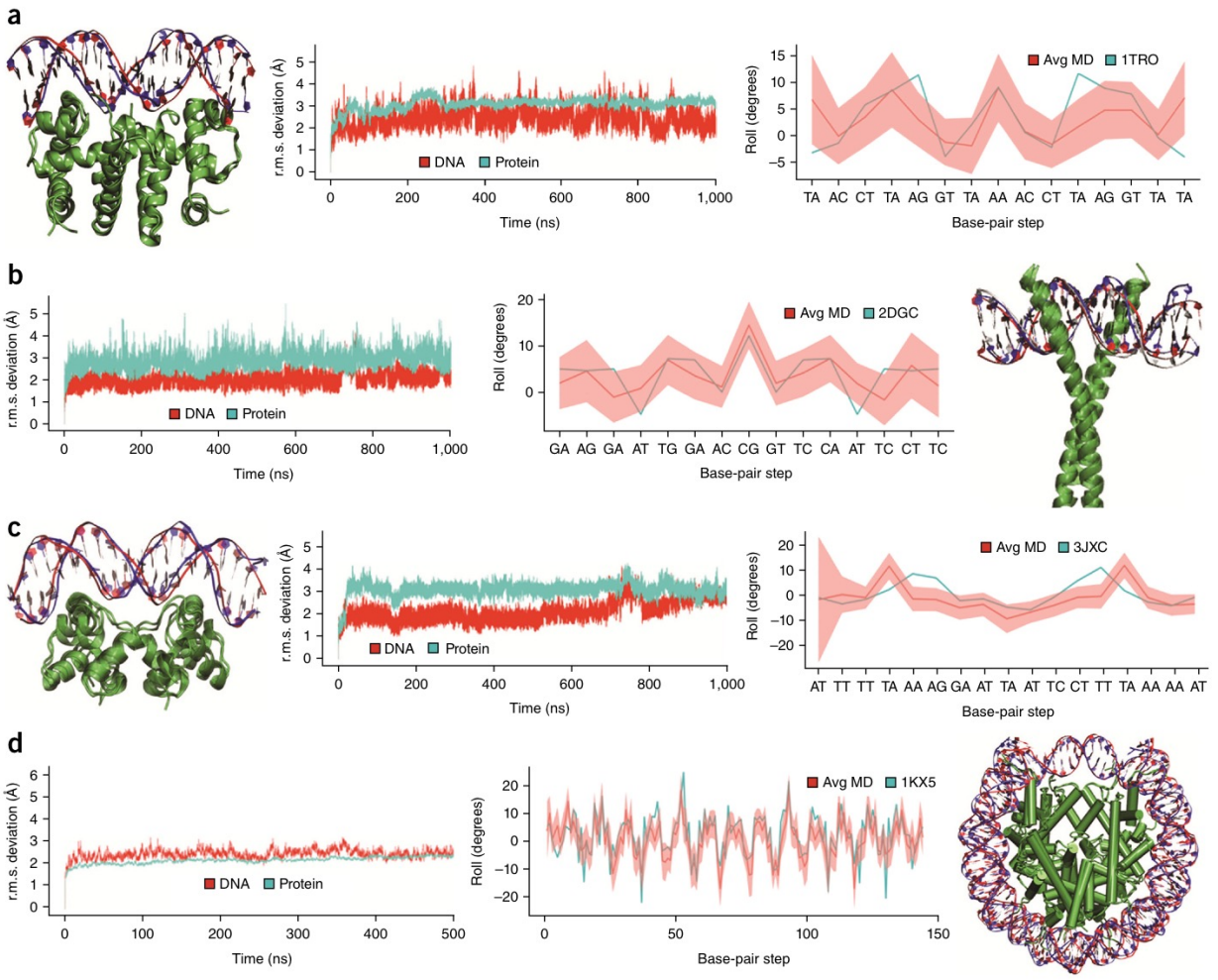



**Figure 3** | Analysis of DNA-protein complexes. (**a**–**d**) Structural details of microsecond trajectories of four complexes: PDB IDs 1TRO (**a**), 2DGC (**b**), 3JXC (**c**) and 1KX5 (**d**) (500-ns trajectory). Each panel shows the overlap of the MD starting (red) and final (blue) structures, the protein secondary structure (green), the time-dependent mass-weighted r.m.s. deviation of all DNA and protein heavy atoms, and a comparison of the rotational helical parameter roll at each base-pair step calculated from the X-ray crystal structure and averaged along the MD simulation (Avg MD) (the s.d. envelope is shown in light red). For clarity, in the roll plot for 1KX5 (**d**), the base-pair steps are defined by the position along the DNA strand, and not by the base pair.

simulations[15] and recognized the considerable 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 NOE violation statistics equivalent 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 an r.m.s. deviation between the simulated and crystal structures of only 0.7 Å and average twist differences of <1°, representing improvements 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 IDs 1TRO, 2DGC, 3JXC and 1KX5) and two prototypical drug-DNA complexes. In all cases, we found excellent agreement with experiments (r.m.s. deviation for DNA of about 2–3 Å in protein-DNA complexes and 1–2 Å in drug-DNA complexes) (**Fig. 3** and **Supplementary Figs. 16** and **17**).

A force field should reproduce not only the structure of DNA but also its mechanical properties[1]. To evaluate the performance of parmbsc1, we first evaluated the microsecond-scale dynamics of the central 10 bp of the DDD. The agreement between parmbsc0 and parmbsc1 normal modes and entropy estimates (Online Methods and **Supplementary Table 9**) demonstrated that parmbsc1 did not 'freeze' the DNA structure, a risk for a force field reproducing 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; **Supplementary Fig. 18**) and the cooperative binding (~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 were substantially smaller than those suggested by Olson *et al.*[17] (**Supplementary Table 10** and **Supplementary Fig. 19**); the dependence of the stiffness parameters on sequence was similar for parmbsc1 and parmbsc0 (ref. 17).

The persistence length and the torsional and 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 about 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 ($\sigma = 0$) and therefore no denatured regions[19,20] (**Supplementary Fig. 20**). We observed a kink in only a single replica for one of the register angles, and in the remaining simulations the DNA remained intact (**Supplementary Fig. 20**). In contrast, negatively supercoiled 100-bp (100t9; $\sigma = -0.05$) and 106-bp (106t9; $\sigma = -0.10$) minicircles formed distortions as a result of the superhelical stress, as previously determined experimentally in studies 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 as expected, the A form was found to be stable in 200-ns control simulations in a mixture of 85% ethanol and 15% water (vol/vol) (**Supplementary Fig. 21**). Parmbsc1 also reproduced the unfolding of DNA d(GGCGGC)$_2$ in a 4 M 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.

On the basis of the wide series of tests reported here, 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. Parameters (**Supplementary Software**) and trajectories are available at http://mmb.irbbarcelona.org/ParmBSC1/.

## METHODS

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

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*


## ACKNOWLEDGMENTS

M.O. thanks the Spanish Ministry of Science (BIO2012-32868), the Catalan SGR, the Instituto Nacional de Bioinformática and the European Research Council (ERC SimDNA) for support. M.O. is an academia researcher in the Catalan Institution for Research and Advanced Studies (ICREA). M.O. thanks the Barcelona Supercomputing Center for CPU and GPU time on MareNostrum and MinoTauro. C.A.L., S.A.H. and A.N. thank the UK HECBioSim Consortium for time on the ARCHER high-performance computing system (grant EP-L000253-1). A.N. was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; grant BB-I019294-1) and thanks ARC Leeds for computational resources. P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores; ANII, Uruguay) researcher. D.A.C. thanks C. Liu for assistance with the crystal simulation analysis.


## AUTHOR CONTRIBUTIONS

I.I. derived the parmbsc1 force-field parameter set. I.I., P.D.D., A.N., A.P., I.F., A.H., J.W., A.B., G.P., F.B., C.A.L. and S.A.H. performed validation simulations. C.G., M.V. and G.P. validated results from NMR-based experiments. C.G. obtained *de novo* NMR spectroscopy measurements. D.A.C. performed crystal MD simulations. R.G., P.A., A.H. and J.L.G. created the database infrastructure and web application. All authors contributed to data analysis. M.O. had the idea for the study, directed the project and wrote the manuscript, which was improved by the rest of the authors.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.




1. Pérez, A., Luque, F.J. & Orozco, M. *Acc. Chem. Res.* **45**, 196–205 (2012).
2. Pérez, A., Luque, F.J. & Orozco, M. *J. Am. Chem. Soc.* **129**, 14739–14745 (2007).
3. Várnai, 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. *Triple-helical Nucleic Acids* 1st edn. (Springer-Verlag, 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. USA* **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. USA* **94**, 14418–14422 (1997).
20. Du, Q., Kotlyar, A. & Vologodskii, A. *Nucleic Acids Res.* **36**, 1120–1128 (2008).

## ONLINE METHODS

**General parameterization strategy.** AMBER charges and van der Waals parameters for DNA can be used to reproduce high-level QM data[21–23] and hydration free energies[24–26], as well as to produce reasonable hydrogen-bond stabilities[2,21–23,27] and complex properties such as sequence-dependent stability 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 on parameterization of the backbone degrees of freedom: sugar puckering, glycosidic torsion, and ε and ζ rotations (taking the recently reparameterized α and γ torsions from parmbsc0 (ref. 4)). Parameterization of the different torsion angles (described 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 as reference high-level Self-Consistent Reaction Field (SCRF) QM data. In cases where fitting of one force-field parameter required knowledge of another parameter for optimization, we used an iterative procedure with parmbsc0 parameters in the first iteration.

**QM calculations.** Model compounds (**Supplementary Fig. 23**) were first geometrically optimized at the B3LYP/6-31++G(d,p) level[30], and from these single-point energies were calculated at the MP2/aug-cc-pVDZ level[31]. To minimize errors in the fitting, we performed optimizations 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 at 10° increments in the region of interest (ε = (175°, 275°), ζ = (220°, 330°)) and at 40° increments in the rest of the profile. Profiles of χ were sampled at 15° increments and profiles of sugar pucker were sampled at 10° increments in the range of phase angles from 0° to 180°, 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 points along the potential energy surface (for ε and ζ, these points were the $B_I$, $B_{Trans}$ and $B_{II}$ states; χ minima of *anti* and *syn* regions, and maxima between them; and minima of North, East and South conformations for the sugar pucker). These calculations entailed optimization at the MP2/aug-cc-pVDZ level followed by single-point calculations at the MP2/aug-cc-pVXZ (X = triplex and quadruplex) levels. We obtained CBS energies by extrapolating to an 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 (described below). All QM calculations were performed with Gaussian09 (http://www.gaussian.com).

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

**Molecular mechanics and potential of mean force.** Molecular mechanics (MM) reference calculations for the QM-optimized structures *in vacuo* were obtained from MM single-point energy calculations carried out with the AMBER 11 package (http://www.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 (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 on the order of 1–2 ns. Simulation details in PMF-US calculations were the same as those outlined below for the validation of MD simulations.

**Force-field fitting.** The procedure for force-field fitting was similar to the parmbsc0 parameterization process[4]. 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 the 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 via standard RESP methods mimicking the original AMBER parameterization. We used the standard torsion definitions 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 $\nu_1$ = O4′-C1′-C2′-C3′, the δ backbone and the $\nu_2$ = C1′-C2′-C3′-C4′ dihedrals, as they connect the two corrections: ε/ζ and χ (refs. 43–45).

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

$$E_{\mathrm{dih},x} = E_{\mathrm{QM}} - E_{\mathrm{ffbsc0}(x=0)} \quad (1)$$

where $x$ stands for a specific torsion or 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 (used in reference MM or US calculations noted above). The dihedral term was defined as

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

where $n$ stands for the periodicity of the torsion, $V_n$ is the rotational barrier, $\varphi$ is the torsion angle and $\alpha$ is the phase angle.

Our flexible Metropolis Monte Carlo algorithm allows for 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 of information from different

profiles obtained in different conditions or with different levels of QM data. Fittings were done taking all the data into 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 such as sugar puckering, detailed attention was needed to properly reproduce the transition region, which we did by increasing the importance of the energy maximum and introducing weights to the slopes in the calculations (**Supplementary Figs. 24–26**). As described before[4], around five to ten acceptable solutions of the Monte Carlo refinement were tested on short MD simulations (~50–100 ns) for one small duplex $d(CGATCG)_2$, rejecting those leading to distorted structures. The optimum parameter set (**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 other, already validated RNA force fields is recommended[45].

**Validation of MD simulations.** We performed MD simulations with the PMEMD code from AMBER 11-12 (http://www.ambermd.org) or with GROMACS[46], depending on the simulation. As shown in **Supplementary Figure 27**, results were insensitive to the simulation engine and to the use of CPU- or GPU-adapted codes[47]. Unless otherwise noted, normal temperature and pressure conditions with default temperature and pressure settings at 300 K and 1 atm, respectively, were used. Calculations used 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. We used the TIP3P[50] or SPCE[51] water model with a minimum 10-Å buffer solvation layer beyond the solute, and we neutralized negatively charged DNA with $Na^+$ or $K^+$ ions[52]. Test simulations with added salt (NaCl) showed that DNA helical conformations were not strongly dependent on the surrounding ionic strength in the range of 0–0.5 M (**Supplementary Discussion** and **Supplementary Fig. 28**). Long-range electrostatic interactions were calculated using the particle mesh Ewald method[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 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 ID 1D23) using 2-μs simulations with 12 unit cells (or 32 duplexes) in the simulation periodic box (**Supplementary Fig. 14**) for a total of 64 μs of duplex simulation.

For annotation of conformational regions at the nucleotide level, we used standard criteria for sugar puckering (C3′-endo for P between 0° and 36° (canonical North), C4′-exo for P between 36° and 72°, O4′-endo for P between 72° and 108° (canonical East), C1′-exo for P between 108° and 144°, C2′-endo for P between 144° and 180° (canonical South), C3′-exo for P between 180° and 216°, C4′-endo for P between 216° and 252°, O4′-exo for P between 252° and 288° (canonical West), C1′-endo for P between 288° and 324°, and C2′-exo for P between 324° and 360°), glycosidic torsion (*anti* for 90° to 180° or −60° to −180° and *syn* for −60° to 90°), BI (ε *trans*, ζ gauche-) and BII (ε gauche-, ζ *trans*). An H bond was annotated using standard GROMACS rules and was considered broken when the 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.

We performed a variety of analyses to characterize the mechanical properties of DNA on the basis of MD simulations. For flexibility analysis we used essential dynamics algorithms[56–58], base-step stiffness analysis[17,59,60] and quasi-harmonic entropies computed with either Andricioaei-Karplus[61] or Schlitter[62] procedures. We determined similarities between essential deformation movements using standard Hess metrics[63] as well as energy-corrected Hess metrics[59]. We calculated polymer deformation parameters (persistence length, stretch and twist torsion modules) by means of different approaches in order to minimize errors associated with 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) implementation of Olson's hybrid approach, which required additional Monte Carlo simulations using MD-derived stiffness matrices[65]. We computed dielectric constants of DNA using Pettit's procedure[66,67]. We used the DDD sequence to compare parmbsc1 to other modern force fields (**Supplementary Discussion** and **Supplementary Fig. 29**). We paid special attention to fraying of the terminal base pairs when analyzing MD trajectories (**Supplementary Fig. 30**) and *de novo* NMR experiments (below and **Supplementary Fig. 31**).

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

**NMR analysis.** We analyzed the ability of MD trajectories to reproduce NMR observables (NOE-derived interatomic distances and residual dipolar couplings) using the last 950 ns of microsecond trajectories. We used the single-value decomposition method implemented in the program PALES[71] to obtain the orientation tensor that best fit the calculated and observed residual dipolar coupling 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 $D_2O$ or $H_2O$-$D_2O$ 9:1 in 25 mM sodium phosphate buffer, 125 mM NaCl, pH 7. NMR spectra were acquired in Bruker spectrometers operating at 800 MHz and processed with Topspin software. Double quantum filter correlation spectroscopy, total correlation spectroscopy and NOE spectroscopy (NOESY) experiments were recorded in $D_2O$ and $H_2O$-$D_2O$ 9:1. The NOESY spectra were acquired with mixing times of 75, 100, 200 and 300 ms, and the total correlation spectra were recorded with a standard MLEV-17 spin-lock sequence and 80-ms mixing time. NOESY spectra were recorded at 5 °C and 25 °C.





doi:10.1038/nmeth.3658

We used the spectral-analysis program Sparky (https://www.cgl.ucsf.edu/home/sparky) for semi-automatic assignment of the NOESY cross-peaks and quantitative evaluation of the NOE intensities. We obtained quantitative distance constraints from NOE intensities by using a complete relaxation matrix analysis with the program MARDIGRAS[73]. We estimated error bounds in the interprotonic distances by carrying out several MARDIGRAS calculations with different initial models, mixing times and correlation times (2.0, 4.0 and 6.0 ns). We obtained final constraints 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 were 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) was not required because direct reading of AMBER topology files was possible.

**Data management.** We placed trajectories and the analysis performed 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 nucleic acid database will be presented elsewhere.




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. Chemistry* **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* Vol. 4 (ed. Leszczynski, J.) 191–225 (World Scientific Publishing, 1999).
27. Pérez, A. *et al. Chemistry* **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, T.H. Jr. *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. Miertuš, S. & Tomasi, J. *Chem. Phys.* **65**, 239–245 (1982).
36. Cancès, 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 (2012).
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, T.E. III. *Nat. Commun.* **5**, 5152 (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. Comput. 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* **1850**, 1059–1071 (2015).
55. Arnott, S. & Hukins, D.W.L. *Biochem. Biophys. Res. Commun.* **47**, 1504–1509 (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* **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 Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **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. USA* **111**, E3624–E3630 (2014).
67. Yang, L., Weerasinghe, S., Smith, P.E. & Pettitt, P.M. *Biophys. 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. *J. Magn. Reson.* **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).


doi:10.1038/nmeth.3658

## SUPPLEMENTARY DISCUSSION

**QM data fitting.**

As shown in the **Supplementary Table 12** refined parmbsc1 parameters fit very well high-level QM data. The *syn-anti* equilibrium, which was non-optimal in parmbsc0, is now well reproduced (**Supplementary Fig. 24**). The fitting to sugar puckering profile was improved by increasing the East barrier, and by displacing the North and South minima to more realistic regions (**Supplementary Table 12** and **Supplementary Fig. 25**). Additionally, parmbsc1 provides ε and ζ conformational map almost indistinguishable from the CCSD(T)/CBS results in solution (**Supplementary Fig. 26**), with errors in the estimates of relative BI/BII stability and transition barrier equal to 0.2 and 0.0 kcal $mol^{-1}$ respectively.

**Force-field benchmark simulations**.

It is not our purpose here to perform a comprehensive comparison of parmbsc1 with previous force-fields. This would require the analysis of >100 structures with up to six other force-fields, clearly out of the scope of this work. We performed, however, a first critical evaluation of the most used force-fields using the well-known Drew Dickerson dodecamer as reference. We tested parmbsc0[1–3], parmbsc0-OL1[4] (ε and ζ corrections from Šponer's group), parmbsc0-OL4[5] (χ corrections), parmbsc0-OL1+OL4[4,5], CHARMM36[6], and a modified parmbsc0 developed by mixing corrected χ values and scaled-down van der Waals interactions (parmbsc0-CG, Cheng-Garcia)[7]. In all cases simulations were extended for at least 1 μs under identical simulation conditions. The value of this benchmark must not be overestimated, since different behavior may be found for other DNA sequences or conformations, but it can be useful to obtain an approximate idea of the range of error expected in parmbsc1 with respect to other modern force-fields. Results are summarized in **Supplementary Table 2** and **Supplementary Figs. 29–31**. All the force-fields are able to maintain the general B-like

conformation in the central part of the duplex. However, significant distortions are found in the terminal pairs for parmbsc0, parmbsc0-OL1 (ε and ζ corrections), and CHARMM36, which show large openings (**Supplementary Fig. 29**) and very frequent fraying, with the formation of non-canonical interactions. The distortion induced by the opening of the terminal C-G pairs is especially dramatic in CHARMM36 simulations (**Supplementary Fig. 29**), but it is not negligible for parmbsc0[8] and parmbsc0-OL1, where aberrant *trans* Watson-Crick contacts involving a cytosine in *syn*, are dominant (**Supplementary Fig. 30**). It is clear that duplexes are flexible and reversible opening and closing of terminal base pair should exist, as found for example in parmbsc1 simulations (**Supplementary Fig. 30**). However, detailed analysis of new NMR spectra (**Supplementary Fig. 31**) shows that there are just minor differences between terminal and interior base pairs, which mean that open states should be short-lived, and not prevalent as in CHARMM36 simulations. Furthermore, no NMR evidence exists (**Supplementary Fig. 31**) supporting the existence of stable unusual contacts involving terminal pairs, or the prevalence of non-*anti* conformations, which are observed in parmbsc0, parmbsc0-OL1 or CHARMM36 simulations.

The introduction of χ corrections removes the excessive fraying of terminal pairs, preserving better the integrity of the entire helix in parmbsc1, parmbsc0-OL4[8], parmbsc0-CG (Cheng-Garcia, and parmbsc0-OL1+OL4 (ε, ζ, and χ corrections together) trajectories (**Supplementary Figs. 29** and **30**). The duplex sampled from parmbsc0-CG calculations is however far from the experimental structures: RMSd around 4 Å (compared to values clearly below 2.0 Å for parmbsc1 simulations), strong under-twisting, poor groove geometry and incorrect description of the BI/BII equilibrium (**Supplementary Table 2**). The sequence dependence of the helical properties, which is clear for the rest of bsc0-based force-fields, is also lost here (**Supplementary Fig. 29**).

Parmbsc0-OL4 and parmbsc0-OL1+OL4 provide reasonable representations of the DDD geometry. However, the use of parmbsc1 leads to clear improvements in all structural

descriptors. Thus, parmbsc1 balances better the sugar puckering (see **Supplementary Fig. 29**), leads to a better balance of BI/BII states (**Supplementary Table 2**), improves very significantly the average roll which is now very close to the NMR estimates, avoiding the excess of roll found in other calculations (**Supplementary Table 2** and **Supplementary Fig. 29**). Parmbsc1 improves very clearly the average twist and its sequence-dependence (RMSd difference between NMR and parmbsc1 twist profiles is 1.9 º, compared with 3.7 º for parmbsc1-OL1+OL4, or 5.6 º for CHARMM36. Not surprisingly, the improvement in twist, roll and puckering is reflected in much more realistic groove dimensions. For example the average difference in groove widths is only 0.3 Å between parmbsc1 and NMR values, while for the parmbsc0-OL1+OL4 force-field error is above 1 Å. In summary, at least for DDD, parmbsc1 provide results of better quality than those obtained with the most recent force-fields for DNA available.

**The effect of ionic strength and the nature of counterion**.

To evaluate potential differences in simulations arising from the ionic strength we performed additionally 2 μs simulations of DDD with extra salt: $Na^+Cl^-$ 150 mM, and 500 mM. These additional calculations were performed using the same conditions outlined previously, showing results that are quite independent on the exact choice (in the 0–500 mM range) of the added extra salt (**Supplementary Fig. 28**).

## SUPPLEMENTARY REFERENCES


1. Pérez, A. *et al. Biophys. J.***92,** 3817–3829 (2007).
2. Cornell, W.D. *et al. J. Am. Chem. Soc.***117,** 5179–5197 (1995).
3. Cheatham III, T.E., Cieplak, P. & Kollman, P.A. *J. Biomol. Struct. Dyn.***16,** 845–862 (1999).
4. Zgarbová, M. *et al. J. Chem. Theory Comput.***9,** 2339–2354 (2013).
5. Krepl, M. *et al. J. Chem. Theory Comput.***8,** 2506–2520 (2012).

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

7. Cheng, A.A., Garcia, A.E. *Proc. Natl. Acad. Sci. USA***110**, 16820–25 (2013).

8. Zgarbová, M., Otyepka, M., Šponer, J., Lankaš, F. & Jurečka, P. *J. Chem. Theory Comput.***10,** 3177–3189 (2014).

## SUPPLEMENTARY TABLES

**Supplementary Table 1.** DNA sequences used for validation of the parmbsc1 force-field. The nature of the structure, the origin of the starting conformation and the length of the production trajectories are also reported. The validation set is divided in several blocks separated in the table by double lines (from top to bottom): i) Normal B-DNA structures (including mismatches, epigenetic modifications and polymeric sequences); ii) very large oligomers; iii) Complexes of DNA with proteins or drugs; iv) Unusual DNA structures; v) dynamic transitions.; parmbsc1 validation; and vi) parmbsc1 benchmarking.

| Sequence | Family | Origine / PDB id | Length (ns) |
|---|---|---|---|
| $d(CGCGAATTCGCG)_2$ | B-DNA | 1BNA, 1NAJ | 1x 800<br>2x 1000<br>1x 12001x 10000 |
| $d(CCATACaATACGG)_2$ | B-DNA mismatch AA | Fiber | 500 |
| $d(CCATACgATACGG)_2$ | B-DNA mismatch GG | Fiber | 500 |
| $d(CGCGA_{5mC}GTCGCG)_2$ | B-DNA 5methylC | Fiber | 250 |
| $d(CGCGA_{5hmC}GTCGCG)_2$ | B-DNA 5hydroxy-methylC | Fiber | 250 |
| $d(CGCGT_{5mC}GACGCG)_2$ | B-DNA 5methylC | Fiber | 500 |
| $d(CGCGACGTCGCG)_2$ | B-DNA | Fiber | 500 |
| $d(CGCGTCGACGCG)_2$ | B-DNA, | Fiber | 500 |
| $d(GCCTATAAACGCCTATAA)_2$ | B-DNA | Fiber | 1000 |
| $d(CTAGGTGGATGACTCATT)_2$ | B-DNA | Fiber | 1000 |
| $d(CACGGAACCGGTTCCGTG)_2$ | B-DNA | Fiber | 1000 |
| $d(GGCGCGCACCACGCGCGG)_2$ | B-DNA | Fiber | 1000 |
| $d(GCCGAGCGAGCGAGCGGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCCTAGCTAGCTAGCTGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCTGCGTGCGTGCGTGGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCGATCGATCGATCGAGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCGAGGGAGGGAGGGAGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCGCGGGCGGGCGGGCGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCGGGGGGGGGGGGGGGC)_2$ | B-DNA | Fiber | 1000 |
| $d(GCGTGGGTGGGTGGGTGC)_2$ | B-DNA | Fiber | 1000 |
| $d(CTCGGCGCCATC)_2$ | B-DNA | 2HKB | 590 |
| $d(CCTCTGGTCTCC)_2$ | B-DNA | 2K0V | 590 |

| | | | |
|---|---|---|---|
| $d(CGCATGCTACGC)_2$ | B-DNA | 2L8Q | 590 |
| $d(GGATATATCC)_2$ | B-DNA | 2LWG | 590 |
| $d(GCGCATGCTACGCG)_2$ | B-DNA | 2M2C | 590 |
| $d(CCTCAGGCCTCC)_2$ | B-DNA | 2NQ1 | 590 |
| $d(CGCGAAAAAACG)_2$ | B-DNA (A-track) | 1D89 | 200 |
| $d(GGCAAAAAACGG)_2$ | B-DNA (A-track) | 1FZX | 200 |
| $d(GCAAAATTTTGC)_2$ | B-DNA (A-track) | 1RVH | 200 |
| $d(CTTTTAAAAG)_2$ | B-DNA (A-track) | 1SK5 | 200 |
| $d(AGGGGCCCCT)_2$ | B-DNA (A-track) | 440D | 200 |
| $d(GGCAAGAAACGG)_2$ | B-DNA (A-track) | 1G14 | 1000 |
| $d(CGATCGATCG)_2$ | B-DNA crystal | 1D23 | 32x 2000 |
| $d(ATGGATCCATAGACCAGAACATGATGTTCTCA)_2$ | B-DNA 32mer | Fiber | 1000 |
| $d(CGCGATTGCCTAACGAGTACTCGTTAGGCAATCGCG)_2$ | B-DNA 36mer | Fiber | 2x 300 |
| $d(CGCGATTGCCTAACGGACAGGCATAGACGTCTATGCCTGTC CGTTAGGCAATCGCG)_2$ | B-DNA 56mer | Fiber | 1x 290<br>1x 500 |
| $d(CGTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCC GAACTCAGAAGTGCG)_2$ | B-DNA 56mer | Fiber | 300 |
| $d(CGCCGGCAGTAGCCGAAAAAATAGGCGCGCGCTCAAAAAAA TGCCCCATGCCGCGC)_2$ | B-DNA 56mer | Fiber | 1x 360<br>1x 440<br>1x 500 |
| $d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)_2$ | 106-mer circle 10 turns (reg A) | Fiber | 100 |
| $d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)_2$ | 106-mer circle 10 turns (reg B) | Fiber | 100 |
| $d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)_2$ | 106-mer circle 10 turns (reg C) | Fiber | 100 |
| $d(ATCTTTGCGGCAGTTAATCGAACAAGACCCGTGCAATGCTA TCGACATCAAGGCCTATCGCTATTACGGGGTTGGGAGTCAATG GGTTCAGGATGCAGGTGAGGAT)_2$ | 106-mer circle 9 turns | Fiber | 50 |
| $d(ATCTTGGCAGTTAATCGAACAAGACCCGTGCAATGCTATCG ACATCAAGGCCTATCGTTACGGGGTTGGGAGTCAATGGGTTCA GGATGCAGGTGAGGAT)_2$ | 100-mer circle 9 turns | Fiber | 100 |
| 147mer nucleosome | DNA-histones | 1KX5 | 500 |
| DNA:HU complex | DNA-HU protein | 1P71 | 1000 |
| DNA:HU complex | DNA-HU protein | 1P71 (without mismatches and flipped bases) | 1000 |

| | | | |
|---|---|---|---|
| DNA:TRP repressor | DNA-repressor | 1TRO | 1000 |
| DNA:leucine zipper | DNA-transc factor | 2DGC | 1000 |
| DNA:P22 c2 | DNA-represor | 3JXC | 1000 |
| $d(CGCAAATTTGCG)_2$-distamycin | DNA-mG binder | 2DND | 700 |
| $d(CTTTTCGAAAAG)_2$-Hoescht | Drug cooperativity | 1QSX | 10x 10 |
| $d(CGTACG)_2$-daunomycin | DNA-intercalator | 1D11 | 600 |
| $d(GGGG)_4$ | PS quadruplex | 352D (without Thymine loops) | 440 |
| $d(GGGG)_4$ | APS quadruplex | 156D (without Thymine loops) | 440 |
| $d(T•A•T)_{10}$ | PS triplex | Fiber | 440 |
| $d(G•G•C)_{10}$ | PS triplex | Fiber | 440 |
| $d(G•G•C)_{10}$ | APS triplex | Fiber | 440 |
| $d(ATATATATATAT)_2$ | H-duplex | 1GQU | 720 |
| $d(CGATATATATAT)_2$ | H-duplex | 2AF1 | 400 |
| d(AAGGGTGGGTGTAAGTGTGGGTGGGT) | G_quadruplex | 2LPW | 5000 |
| d(AGGGTTAGGGTTAGGGTTAGGG) | G-loop quadruplex(HTQ) | 1KF1 | 1000 |
| $d(GGGGTTTTGGGG)_2$ | G quadruplex (OxyQ) | 1JRN | 1000 |
| $d(CCGGTACCGG)_4$ | Holliday Junction | 1DCW | 1000 |
| $d(CGCGCGCGCG)_2$ | Z-DNA, duplex | 1I0T | 2x 385 |
| d(GCGAAGC) | Hairpinfold (REXMD) | 1PQT | 1000 |
| $d(CGCGAATTCGCG)_2$ | A-form in ethanol | 1BNA | 200 |
| $d(CGCGAATTCGCG)_2$ | A to B transition ($H_2O$) | 1BNA | 5x40 |
| $d(GGCGCC)_2$ | DNA unfolding (Pyridine) | 1P25 | 400 |
| $d(CGCGAATTCGCG)_2$ | DDD, 0.15M NaCl | 1BNA | 2000 |
| $d(CGCGAATTCGCG)_2$ | DDD, 0.5M NaCl | 1BNA | 3000 |
| $d(CGCGAATTCGCG)_2$ | parmBSC0 | 1BNA | 1500 |
| $d(CGCGAATTCGCG)_2$ | parmBSC0-OL1 | 1BNA | 1500 |
| $d(CGCGAATTCGCG)_2$ | parmBSC0-OL4 | 1BNA | 1500 |
| $d(CGCGAATTCGCG)_2$ | parmBSC0-OL1-OL4 | 1BNA | 1500 |
| $d(CGCGAATTCGCG)_2$ | parmBSC0-Cheng-Garcia | 1BNA | 1500 |
| $d(CGCGAATTCGCG)_2$ | CHARMM36 | 1BNA | 1500 |
| $d(CGCGAATTCGCG)_2$ | DDD, Amber GPU | 1BNA | 100 |
| $d(CGCGAATTCGCG)_2$ | DDD, Amber CPU | 1BNA | 100 |
| $d(CGCGAATTCGCG)_2$ | DDD, Gromacs GPU | 1BNA | 100 |
| $d(CGCGAATTCGCG)_2$ | DDD, Gromacs CPU | 1BNA | 100 |

**Supplementary Table 2.** MD-averaged helical parameters (on 1.2 μs simulation time) of Drew-Dickerson dodecamer in parmbsc1 simulations (and, as a control, other modern force-fields) compared with the NMR and X-ray estimates. [a]

| | Twist | Roll | Slide | Rise | Shift | Tilt | BI(%) | Major groove width | Minor groove width |
|---|---|---|---|---|---|---|---|---|---|
| **Parmbsc1** | **34.3±5.4** | **1.5±5.4** | **-0.3±0.5** | **3.3±0.3** | **0.0±0.8** | **0.0±4.5** | **77** | **11.9±1.7** | **5.4±1.2** |
| Parmbsc0 | 32.8±5.8 | 2.7±5.8 | -0.4±0.6 | 3.3±0.3 | 0.0±0.7 | 0.0±4.3 | 84 | 12.9±1.8 | 3.9±1.2 |
| OL1 | 33.3±5.7 | 2.7±5.9 | -0.2±0.6 | 3.3±0.3 | 0.0±0.7 | 0.0±4.4 | 83 | 12.2±1.4 | 6.1±1.3 |
| OL4 | 33.3±6.4 | 2.6±5.9 | -0.1±0.6 | 3.3±0.3 | 0.0±0.7 | 0.0±4.5 | 85 | 12.1±1.4 | 6.5±1.3 |
| OL1+OL4 | 33.0±6.1 | 2.8±5.7 | -0.3±0.6 | 3.3±0.3 | 0.0±0.7 | 0.0±4.3 | 86 | 12.4±1.5 | 6.0±1.2 |
| C36 [d] | 34.5±11 | 5.1±8.8 | 0.8±1.0 | 3.6±0.8 | -0.1±1.1 | 0.9±8.0 | 66 | 10.5±1.5 | 8.3±1.7 |
| Cheng-Garcia(CG) | 32.5±3.4 | 1.5±5.2 | -1.7±0.5 | 3.4±0.3 | 0.0±0.4 | 0.0±4.3 | 100 | 15.3±1.6 | 5.5±0.9 |
| **X-ray [b]** | **35.2±0.6** | **-0.7±1.1** | **0.1±0.1** | **3.3±0.1** | **-0.1±0.1** | **-0.4±0.9** | | **11.2±0.1** | **4.6±0.3** |
| **NMR [c]** | **35.6±0.8** | **1.6±1.0** | **-0.3±0.1** | **3.2±0.1** | **0.0±0.1** | **0.0±0.7** | **73[e]** | **11.9±0.3** | **4.7±0.3** |

[a] Translational parameters and groove widths are in Å, while rotational parameters are in degrees. Note that for MD trajectories the standard deviations are computed from sequence-averages and time-averages. [b] X-ray mean values and standard deviations were obtained averaging the following structures (PDB id): 1BNA[1], 2BNA[2], 7BNA[3] and 9BNA[4]. [c] NMR mean values and standard deviations were obtained by averaging over the ensemble of structures contained in the PDB id 1NAJ[5].[d] These average values are contaminated by the opening of terminal base pairs (note large standard deviations in roll and twist). [e] Average value of BI population taken by averaging direct NMR estimates[6,7]. See also Supplementary Discussion and **Supplementary Figs. 29-31** for a discussion on the relative performance of parmbsc1 with respect to other force-fields.


1. Drew, H.R. *et al. Proc. Natl. Acad. Sci. U. S. A.***78,** 2179–2183 (1981).
2. Drew, H.R., Samson, S. & Dickerson, R.E. *Proc. Natl. Acad. Sci. U. S. A.***79,** 4040–4044 (1982).
3. Holbrook, S.R. *et al. Acta Crystallogr.,Sect.B***41,** 255–262 (1985).
4. Westhof, E. *J. Biomol. Struct. Dyn.***5,** 581–600 (1987).
5. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B. & Bax, A.*J. Biomol. NMR***26,** 297–315 (2003).
6. Tian, Y., Kayatta, M., Shultis, K., Gonzalez, A., Mueller, L.J. & Hatcher, M.E., *J. Phys. Chem. B***113**, 2596–2603 (2008).
7. C. D. Schwieters, C.D. & Clore, G.M. *Biochemistry***46**, 1152–1166 (2007).

**Supplementary Table 3.** Ability of MD-ensembles obtained from parmbsc0 and parmbsc1 force fields to reproduce NMR observables for Drew-Dickerson dodecamer. The first block correspond to residual dipolar couplings Q-factor, $q = \sqrt{\sum\left(RDC_{calc} - RDC_{exp}\right)^2} \Big/ \sqrt{\sum RDC_{exp}^2}$, where $RDC_{exp}$ has been determined using PALES[1], and the second block to NOEs (146 restraints).

| | NMR | X-ray | Fiber model B-DNA | Fiber model A-DNA | BSC1 | BSC0 |
|---|---|---|---|---|---|---|
| Bicelles, 1NAJ [a], 129 RDCs | 0.17 | 0.49 | 0.51 | 0.87 | 0.32 | 0.36 |
| Bicelles, 1DUF [b], 204 RDCs | 0.23 | 0.53 | 0.66 | 0.92 | 0.34 | 0.38 |
| Sum of violations (A) | 0.01 | 10.0 | 7.6 | 42.01 | 0.4 | 2.6 |
| Largest violation (A) | 0.01 | 1.0 | 0.4 | 1.3 | 0.2 | 1.3 |
| Num. of violated restraints | 1 | 35 | 36 | 84 | 2 | 5 |

[a] Data taken from ref. 2. [b] Data taken from ref. 3.


1. Zweckstetter, M. *Nat. Protoc.*, ***3***, 679-690 (2008).
2. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B. & Bax, A.*J. Biomol. NMR***26,** 297–315 (2003).
3. Tjandra, N., Tate, S. I., Ono, A., Kainosho, M. & Bax, A. *J. Am. Chem. Soc.***122,** 6190–6200 (2000).

**Supplementary Table 4.** Different metrics showing the quality of parmbsc1 simulations for B-DNA duplexes.[a]

| DNA seq or PDB id | Ref | RMSd | RMSd/bp | % H-bond | Avg. twist | Avg. roll |
|---|---|---|---|---|---|---|
| 1BNA (12mer) | C | 2.1 / 1.7 | 0.18 / 0.17 | 96 / 98 | 35.6 / 34.3 | 2.8 / 1.5 |
| 1NAJ (12mer) | N | 1.7 / 1.4 | 0.15 / 0.15 | 96 / 98 | 35.6 / 34.3 | 2.8 / 1.5 |
| CCATACgATACGG[b] | N | 2.9 / 2.3 | 0.22 / 0.21 | 91 / 91 | 33.5 / 34.2 | 8.8 / 1.6 |
| CCATACaATACGG[c] | N | 3.3 / 3.1 | 0.26 / 0.28 | 93 / 94 | 33.7 / 34.1 | 2.7 / 2.5 |
| CGCGACGTCGCG | F | 2.0 / 1.5 | 0.17 / 0.15 | 98 / 99 | 34.8 / 34.6 | 3.1 / 2.0 |
| CGCGTCGACGCG | F | 2.6 / 1.5 | 0.22 / 0.16 | 97 / 99 | 34.1 / 34.5 | 3.4 / 2.3 |
| GCGAGGGAGGGAGGGAGC | F | 2.7 / 2.3 | 0.15 / 0.15 | 97 / 99 | 33.5 / 33.3 | 2.5 / 2.9 |
| GCGCGGGCGGGCGGGCGC | F | 2.3 / 2.0 | 0.13 / 0.13 | 97 / 99 | 33.7 / 33.7 | 2.8 / 3.3 |
| GCGGGGGGGGGGGGGGGC | F | 3.0 / 2.7 | 0.17 / 0.17 | 98 / 99 | 32.8 / 32.6 | 3.0 / 3.5 |
| GCGTGGGTGGGTGGGTGC | F | 2.2 / 1.9 | 0.12 / 0.12 | 97 / 99 | 33.1 / 33.0 | 2.7 / 3.2 |
| GCCGAGCGAGCGAGCGGC | F | 2.9 / 2.4 | 0.17 / 0.15 | 98 / 99 | 34.7 / 34.5 | 2.1 / 2.6 |
| GCCTAGCTAGCTAGCTGC | F | 2.2 / 1.9 | 0.13 / 0.12 | 97 / 98 | 34.3 / 34.2 | 1.6 / 2.1 |
| GCTGCGTGCGTGCGTGGC | F | 2.2 / 2.0 | 0.13 / 0.13 | 97 / 98 | 32.6 / 34.5 | 2.3 / 2.8 |
| GCGATCGATCGATCGAGC | F | 2.0 / 1.8 | 0.11 / 0.12 | 97 / 98 | 34.8 / 34.7 | 1.9 / 2.3 |
| GCCTATAAACGCCTATAA | F | 2.9 / 2.8 | 0.17 / 0.18 | 94 / 97 | 34.7 / 34.4 | 1.6 / 2.0 |
| CTAGGTGGATGACTCATT | F | 3.3 / 2.9 | 0.18 / 0.18 | 94 / 97 | 30.9 / 31.8 | 1.2 / 4.6 |
| CACGGAACCGGTTCCGTG | F | 3.0 / 2.9 | 0.17 / 0.18 | 95 / 97 | 34.6 / 33.8 | 2.7 / 2.0 |
| GGCGCGCACCACGCGCGG | F | 3.4 / 2.7 | 0.19 / 0.17 | 96 / 98 | 33.2 / 34.4 | 3.5 / 2.4 |
| 1D89 (12mer) | C | 2.3 / 1.9 | 0.19 / 0.19 | 93 / 98 | 35.6 / 33.9 | 3.0 / 1.7 |
| 1FZX (12mer) | N | 1.8 / 1.7 | 0.16 / 0.18 | 95 / 96 | 33.9 / 33.8 | 2.4 / 2.3 |
| 1RVH (12mer) | N | 1.9 / 1.7 | 0.16 / 0.17 | 98 / 98 | 33.9 / 34.0 | 2.2 / 2.6 |
| 1SK5 (10mer) | C | 2.1 / 1.8 | 0.21 / 0.23 | 93 / 97 | 34.2 / 34.3 | 1.7 / 1.7 |
| CGATATATATATCG | F | 1.9 / 1.6 | 0.16 / 0.17 | 96 / 97 | 34.4 / 34.4 | 2.9 / 1.7 |
| 2HKB (12mer) | N | 1.8 / 1.7 | 0.15 / 0.17 | 96 / 97 | 34.1 / 33.8 | 2.3 / 2.6 |
| 2K0V (12mer) | N | 2.4 / 2.1 | 0.20 / 0.22 | 95 / 96 | 33.9 / 33.5 | 2.2 / 1.9 |
| 2L8Q (12mer) | N | 1.9 / 1.5 | 0.16 / 0.16 | 95 / 97 | 34.4 / 34.1 | 2.7 / 2.5 |
| 2LWG (10mer) | N | 1.8 / 1.5 | 0.18 / 0.19 | 98 / 99 | 34.5 / 34.6 | 2.4 / 1.5 |
| 2M2C (14mer) | N | 2.5 / 2.3 | 0.18 / 0.20 | 96 / 97 | 34.4 / 34.0 | 2.7 / 2.5 |

[a] The reference structures used for comparison were taken from X-ray crystallography (C), NMR (N) or fiber (F) data, as available. Except otherwise mentioned, all the duplexes were self-complementary and only one strand is noted. For structures available in the Protein Data Bank we display only the PDB code. RMSd are in Å and average rotational parameters are in degrees. Note that the first value in each cell corresponds to a sequence average considering the complete oligomer, while the second value in each cell was computed excluding the terminal residues. [b] Structure containing a G:G mismatch. The NMR structure used as reference was solved after parmbsc1 was derived[1]. [c] Same than [b] but containing an A:A mismatch.

1. Rossetti, G., Dans, P.D. *et al. Nucleic Acids Res.***43**, 4309-4321 (2015).

**Supplementary Table 5.** Long oligomers RMSd, helical parameters, and bending (reported herein as % of shortening) values, for all the residues or excluding the terminal ones, with respect to the ideal helix built using average dinucleotide X-ray helical parameters.

| | Seq1[c] | Seq2a | Seq2b | Seq3 | Seq4a | Seq4b |
|---|---|---|---|---|---|---|
| RMSd | 4.4±1.3 | 4.2±1.5 | 4.3±1.3 | 6.7±2.8 | 7.2±2.7 | 7.4±2.7 |
| RMSd (no ends) | 4.2±1.2 | 4.0±1.4 | 4.1±1.2 | 6.4±2.6 | 6.9±2.6 | 7.0±2.5 |
| RMSd / bp[a] | 0.14 | 0.12 | 0.12 | 0.12 | 0.14 | 0.13 |
| RMSd / bp (no ends) | 0.14 | 0.12 | 0.12 | 0.12 | 0.13 | 0.13 |
| Avg. twist (º) | 34.9±7.3 | 35.0±5.3 | 34.5±5.4 | 34.2±5.6 | 34.8±5.3 | 34.3±5.8 |
| Avg. roll (º) | 2.1±8.4 | 1.5±5.8 | 1.7±5.8 | 2.2±5.7 | 1.7±5.8 | 2.0±6.0 |
| Avg. slide (Å) | -0.4±0.7 | -0.2±0.5 | -0.3±0.6 | -0.4±0.6 | -0.2±0.5 | -0.3±0.5 |
| Shortening[b] | 4±2 (16) | 5±2 (20) | 5±2 (17) | 6±3 (18) | 6±3 (23) | 6±3(21) |

[a] Values per base pair are indicated to avoid size-inconsistency. [b] Note that for helix shortening the maximum shortening percentages are reported in bracket.
[c] Seq1: ATGGATCCATAGACCAGAACATGATGTTCTCA in TIP3P water;
Seq2a: CGCGATTGCCTAACGAGTACTCGTTAGGCAATCGCG in SPCE water;
Seq2b: idem Seq2a in TIP3P water;
Seq3: CGCCGGCAGTAGCCGAAAAAATAGGCGCGCGCTCAAAAAAATGCCCCATGCCGCGC in TIP3P water;
Seq4a: CGCGATTGCCTAACGGACAGGCATAGACGTCTATGCCTGTCCGTTAGGCAATCGCG in SPCE water;
Seq4b: idem Seq4a in TIP3P water.

**Supplementary Table 6.** Statistic of NOE restraints violations for different nucleic acids (include: normal duplexes, hairpins, quadruplexes, and A-tracks).[a]

| Structure (PDB id) | Number Restraints | Average Violation | Largest Violation | Number violations |
|---|---|---|---|---|
| 1NAJ | 146 | *0.0001* | *0.01* | *1* |
| | | 0.003 | 2 | 1 |
| 2LPW | 938 | *0.0006* | *0.1* | *12* |
| | | 0.07[b] | 7.0 | 45 |
| 1PQT | 94 | *0.01* | *0.1* | *3* |
| | | 0.01 | 0.1 | 2 |
| 1G14 | 218 | *0.01* | *0.2* | *33* |
| | | 0.05 | 0.9 | 44 |
| 1RVH | 446 | *0.02* | *0.3* | *50* |
| | | 0.03 | 0.8 | 56 |
| 2LWG | 415 | *0.01* | *0.5* | *28* |
| | | 0.03 | 1.4 | 38 |
| 2K0V | 634 | *0.05* | *1.9* | *83* |
| | | 0.12 | 2.5 | 129 |
| 2L8Q | 172 | *0.0005* | *0.09* | *1* |
| | | 0.001 | 0.26 | 1 |
| 2M2C | 296 | *0.15* | *3.3* | *54* |
| | | 0.13 | 3.1 | 50 |
| 2NQ1 | 870 | *0.02* | *1.3* | *111* |
| | | 0.09 | 3.9 | 162 |

[a] For each PDB entry we show the number of experimental restraints, the average deviation (A), the maximum deviation (A), and the number of restraint violations. In each cell NMR results are reported in italic, *i.e.*, the values obtained when experimental restraints were enforced to solve the structure; while the MD results obtained using parmbsc1 simulations are reported with normal characters. [b] Since the NOE deviations were larger than usual for this hairpin, calculations were repeated using parmbsc0 and CHARMM36 force-fields, finding 73 and 64 violations respectively.

**Supplementary Table 7.** Quality factor (Q-factor), $q = \sqrt{\sum(RDC_{calc} - RDC_{exp})^2} \Big/ \sqrt{\sum RDC_{exp}^2}$, for the agreement between observed and predicted residual dipolar couplings (RDCs), using both experimental NMR structures and parmbsc1 MD simulations. [a]

| Structure | Alignment Method | Number RDCs | Q-factor (NMR) | Q-factor (MD) |
|---|---|---|---|---|
| 1NAJ | Bicelles | 204 | 0.23 | 0.34 |
| 2LPW | Bicelles | 57 | 0.25 | 0.54 |
| 1PQT | Pf1 | 29 | 0.11 | 0.41 |
| 1RVH | Pf1 | 72 | 0.13 | 0.27 |
| 2LWG | Pf1 | 46 | 0.18 | 0.29 |

[a] Note that lower Q-factor indicates better agreement. Typically data sets include both C-H and N-H dipolar couplings. The alignment media used to record NMR RDCs is indicated in all the cases. RDCs were back-calculated from the MD simulations using PALES.

**Supplementary Table 8.** Statistic of NOE violations for different nucleic acids, for oligomers solved after parmbsc1 development. NOE restraints here are determined using the full matrix relaxation and are more accurate than those typically found in the literature (rough data available upon request). [a]

| Duplex | Number restraints | Average violation | Largest violation | Number violations[b] | Rfactor$_{2\alpha}$[c] |
|---|---|---|---|---|---|
| GG mismatch | 246 | *0.004* | *0.090* | *73\|15\|0* | *0.204* |
| | | 0.012 | 0.302 | 64\|36\|7 | 0.172 |
| AA mismatch | 230 | *0.003* | *0.160* | *64\|6\|1* | *0.290* |
| | | 0.006 | 0.083 | 51\|27\|0 | 0.292 |
| A***C***GT control | 208 | *0.006* | *0.046* | *85\| 29\|0* | *0.261* |
| | | 0.022 | 0.123 | 106\|79\|12 | 0.250 |
| A***5mC***GT[d] | 102 | *0.034* | *0.205* | *57\|49\|14* | *0.197* |
| | | 0.035 | 0.189 | 60\|45\|18 | 0.243 |
| A***5hmC***GT[e] | 216 | *0.004* | *0.045* | *63\|18\|0* | *0.232* |
| | | 0.014 | 0.218 | 86\|57\|2 | 0.236 |

[a] Note that the comparisons are made between metrics obtained for the NMR ensemble (the set of structures refined by imposing NMR restraints) in italics, and those coming from the unbiased MD trajectory in roman. [b] To define "number of violations" we used three criteria: i) the distances given by the flat well limits (left value in the cell), ii) the boundaries of the "contact" are extended by ±0.2 Å (middle value), and finally iii) the upper-limit is multiplied by 1.25 (right value in the cell). [c] The global quality factor Rfactor$_{2\alpha}$ [1, 2] take values around 0.6 and 0.7 for B and A-DNA respectively. The sequences considered here are reported in **Supplementary Table 1**.[d] 5mC stands for 5-methyl-cytosine. [e] 5hmC stands for 5-hydroxymethyl-cytosine.

1. Gonzalez, C., Rullmann, J.A.C., Bonvin, A., Boelens, R. & Kaptein, R. *J. Magn. Reson.***91,** 659–664 (1991).
2. Gronwald, W. *et al. J. Biomol. NMR***17,** 137–151 (2000).

**Supplementary Table 9.** Different metrics of DNA flexibility in the Cartesian space for the Drew-Dickerson dodecamer simulation using parmbsc0 and parmbsc1 force-fields.

| Metrics | Parmbsc1 | Parmbsc0 |
|---|---|---|
| Entropy all heavy [a] | 2.14 | 2.14 |
| | *2.00* | *2.00* |
| Entropy backbone | 1.16 | 1.15 |
| | *1.11* | *1.10* |
| First three eigenvalues [b] | 176,127,102 | 204,135,104 |
| Eigenvalues 10, 20 and 30 | 20,8,4 | 23,9,4 |
| Self-similarity (10 eigenvalues)[c] | 0.89 | 0.94 |
| Similarity parmbsc1/parmbsc0[d] | 0.81 | |
| Relative similarity[e] | 0.89 | |
| Energy weighted similarity | 0.88 | |
| Relative weighted similarity | 0.93 | |

[a] Entropies in kcal $mol^{-1}$ $K^{-1}$ are determined using Schlitter (roman) and Andrioacei-Karplus (italics) for the entire 1.2 μs simulations. [b] Eigenvalues (in $Å^2$) are computed by diagonalization of the covariance matrix and ordered according to their contribution to the total variance. [c] Self-similarity is computed by comparing the first and second halves of the same trajectory. [d] Similarity and weighted similarity indexes are computed using the Hess matrix[1], or following reference[2]. [e] Relative similarities are computed from absolute similarities and self-similarities as described elsewhere[3].

1. Hess, B. *Phys. Rev. E***62,** 8438 (2000).
2. Pérez, A. *et al. J. Chem. Theory Comput.***1,** 790–800 (2005).
3. Orozco, M., Pérez, A., Noy, A. & Luque, F.J. *Chem. Soc. Rev.***32,** 350–364 (2003).

**Supplementary Table 10.** Sequence-dependent dinucleotide force constants associated with the deformation of a single helical degree of freedom.[a]

| bps | Twist | Tilt | Roll | Shift | Slide | Rise |
|---|---|---|---|---|---|---|
| AA | 0.028 | 0.037 | 0.020 | 1.72 | 2.13 | 7.64 |
| | *0.036* | *0.045* | *0.023* | *1.68* | *2.91* | *9.33* |
| | **0.043** | **0.044** | **0.022** | **2.45** | **3.56** | **9.47** |
| | (0.092) | (0.100) | (0.049) | (3.98) | (6.16) | (21.75) |
| AC | 0.036 | 0.038 | 0.023 | 1.28 | 2.98 | 8.83 |
| | *0.047* | *0.045* | *0.027* | *1.54* | *3.67* | *10.44* |
| | **0.034** | **0.034** | **0.025** | **1.55** | **3.33** | **8.31** |
| | (0.073) | (0.111) | (0.080) | (2.94) | (6.37) | (23.86) |
| AG | 0.028 | 0.037 | 0.019 | 1.40 | 1.78 | 7.04 |
| | *0.031* | *0.049* | *0.025* | *1.54* | *2.78* | *9.73* |
| | **0.036** | **0.045** | **0.022** | **2.00** | **2.82** | **9.35** |
| | (0.064) | (0.149) | (0.096) | (3.21) | (7.19) | (29.50) |
| AT | 0.031 | 0.035 | 0.022 | 1.05 | 3.77 | 9.34 |
| | *0.031* | *0.033* | *0.024* | *1.24* | *4.10* | *9.23* |
| | **0.032** | **0.032** | **0.023** | **1.21** | **3.49** | **7.32** |
| | (0.070) | (0.166) | (0.055) | (3.17) | (10.69) | (25.55) |
| CA | 0.015 | 0.025 | 0.016 | 1.05 | 1.80 | 6.30 |
| | *0.028* | *0.028* | *0.016* | *0.77* | *2.69* | *7.66* |
| | **0.032** | **0.027** | **0.018** | **1.60** | **2.19** | **6.71** |
| | (0.043) | (0.082) | (0.048) | (3.73) | (2.40) | (18.24) |
| CC | 0.026 | 0.042 | 0.020 | 1.43 | 1.57 | 7.86 |
| | *0.032* | *0.049* | *0.021* | *1.50* | 1.78 | *9.59* |
| | **0.030** | **0.043** | **0.021** | **1.53** | **1.74** | **8.96** |
| | (0.041) | (0.119) | (0.064) | (2.43) | (3.54) | (30.31) |
| CG | 0.014 | 0.026 | 0.016 | 1.05 | 1.91 | 6.11 |
| | *0.024* | *0.032* | *0.016* | *1.10* | *2.47* | *7.61* |
| | **0.032** | **0.024** | **0.017** | **1.82** | **2.48** | **6.64** |
| | (0.047) | (0.068) | (0.050) | (1.59) | (3.30) | (14.16) |
| GA | 0.024 | 0.038 | 0.020 | 1.32 | 1.88 | 8.48 |
| | *0.034* | *0.045* | *0.023* | *1.40* | *2.66* | *10.08* |
| | **0.040** | **0.041** | **0.024** | **2.27** | **3.40** | **10.12** |
| | (0.071) | (0.087) | (0.046) | (6.54) | (2.78) | (22.82) |
| GC | 0.022 | 0.036 | 0.026 | 1.18 | 2.59 | 9.47 |
| | *0.031* | *0.043* | *0.025* | *1.32* | *3.19* | *11.16* |
| | **0.027** | **0.031** | **0.028** | **1.70** | **4.79** | **9.43** |
| | (0.055) | (0.082) | (0.082) | (3.35) | (6.24) | (25.86) |
| TA | 0.018 | 0.019 | 0.015 | 0.64 | 1.25 | 6.08 |
| | *0.028* | *0.025* | *0.015* | *0.50* | *2.16* | *7.47* |
| | **0.036** | **0.021** | **0.015** | **0.93** | **1.52** | **6.61** |
| | (0.052) | (0.148) | (0.029) | (3.86) | (2.35) | (21.91) |

[a] Parmbsc0 (roman)[1], parmbsc1 (italics), and CHARMM27 (bold) force-fields are compared with stiffness values derived from inspection of the X-Ray structural variability of the different base pair steps (in brackets)[2]. Note that values for a particular base pair step are diagonal entries of its stiffness matrix. Values reported in the table are averages over all the equivalent steps. The rotational values are in kcal $mol^{-1}$ $deg^{-2}$ and translational ones are in kcal $mol^{-1}$ $Å^{-2}$.

1. Perez, A., Lankas, F., Luque, F. J. & Orozco, M. *Nucleic Acids Res.***36,** 2379–2394 (2008).
2. Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M. & Zhurkin, V.B. *Proc. Natl. Acad. Sci.***95,** 11163–11168 (1998).

**Supplementary Table 11.** Elastic properties derived from atomistic MD simulations of three sequences of DNA.[a]

| | Persistence length | | | | | | Other stiffness descriptors | |
|---|---|---|---|---|---|---|---|---|
| *DNA* | *Roll* | *Tilt* | *Isotropic* | *Dynamics* | *Static* | *Total* | *Torsion module* | *Stretch Module* |
| Seq3[b] | 41±10 | 63±16 | 49±11 | *63±1* | *566±150* | *57±2*<br>41±20<br>49±20 | 48±19<br>*101±9* | 1,373±195<br>*1,857±22* |
| Seq4a | 41±8 | 64±14 | 50±9 | *71±1* | *608±150* | *64±2*<br>42±23<br>50±23 | 49±13<br>*102±10* | 1,430±210<br>*1,567±42* |
| Seq4b | 41±7 | 65±15 | 50±9 | *71±1* | *310±44* | *57±2*<br>39±20<br>48±21 | 46±13<br>*107±12* | 1,476±185<br>*1,832±45* |
| Avg. | 41±14 | 64±26 | 50±17 | *68±2* | *495±211* | *59±4*<br>41±30<br>49±30 | 47±26<br>*104±18* | 1,426±341<br>*1,752±65* |

[a] Persistence lengths and torsion modules are in nm, and stretch module are in pN. Values in roman correspond to 2 bp windows, while values in italic correspond approximately to one DNA turn windows[1]: (i) persistence lengths are calculated by linearly fitting the directional decay from 2 bp until 11 bp sub-fragments, and the static contributions come from the distribution of sequence-dependent static bends obtained through the MD average structure; (ii) stretch modulus are obtained by linearly fitting end-to-end variances of all central sub-fragments containing from 8 bp up to 16 bp to avoid the very long end-effect; (iii) torsional modulus is evaluated by averaging the 38 central sub-fragments containing 11 bp. Only the central 48-mer of the 56-mers was considered to minimize end-effects. Underlined values were obtained using a local implementation of Olson's Monte Carlo procedure[2], without additional corrections, or including (underlined with a curved line) partial variance corrections as discussed in Noy and Golestanian 2012[1].[b] See **Supplementary Table 5** for the definition of the sequences. As reference experimental estimates for persistence lengths are around 50 nm[3], for static persistence lengths are in the range of 200-1,500 nm[4, 5], for stretch modulus are around 1,100-1,500 pN[6, 7] and for torsion (twist) constants are in the range 80-120 nm[8, 9].

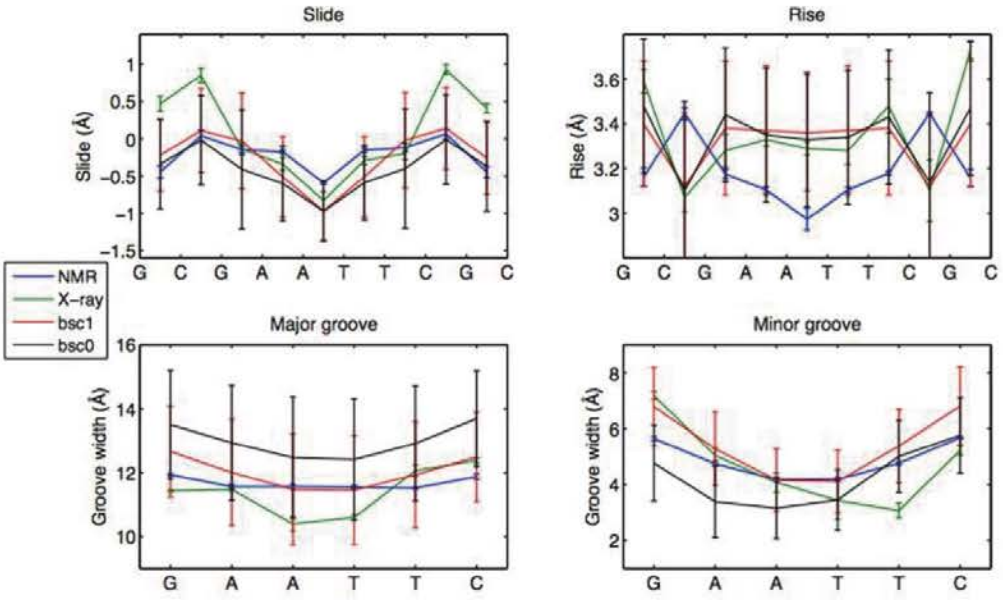

1. Noy, A. & Golestanian, R. *Phys. Rev. Lett.***109,** 228101 (2012).
2. Zheng, G., Czapla, L., Srinivasan, A.R. & Olson, W.K. *Phys. Chem. Chem. Phys.***12,** 1399–1406 (2010).
3. Mazur, A.K. & Maaloum, M. *Nucleic Acids Res.***42,** 14006-14012 (2014).
4. Smith, S.B., Finzi, L. & Bustamante, C. *Science***258,** 1122–1126 (1992).
5. Moukhtar, J. *et al. J. Phys. Chem. B***114,** 5125–5143 (2010).
6. Smith, S.B., Cui, Y. & Bustamante, C. *Science***271,** 795–799 (1996).
7. Gross, P. *et al. Nat. Phys.***7,** 731–736 (2011).
8. Strick, T.R., Allemand, J.-F., Bensimon, D., Bensimon, A. & Croquette, V. *Science***271,** 1835–1837 (1996).
9. Moroz, J.D. & Nelson, P. *Proc. Natl. Acad. Sci.***94,** 14418–14422 (1997).

**Supplementary Table 12**. Differences between QM and force-field estimates for the parameterized systems. Values refer to calculations performed in water.

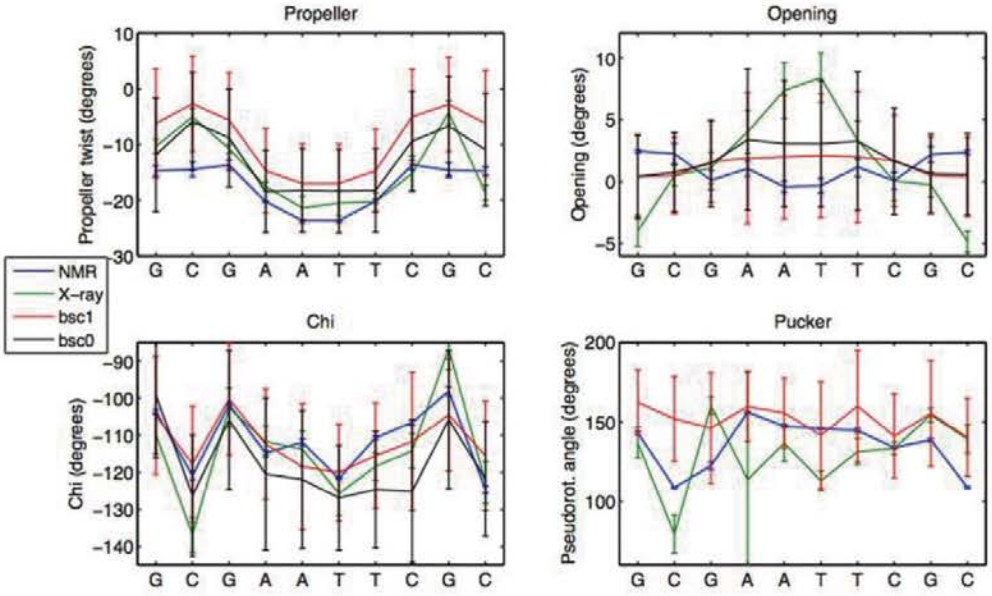
| Torsion | Adenosine | Guanosine | Cytosine | Thymidine |
|---|---|---|---|---|
| | | Glycosidic torsion (χ) | | |
| | | *Geometries (°)* [a] | | |
| Anti | 14 / 40 | 9 / 40 | 2.5 / 1 | 2.5 / 1 |
| Barrier | 1.5 / 11 | 2.5 / 15 | 13 / 10 | 11 / 11 |
| Syn | 7 / 32 | 2.5 / 30 | 12 / 30 | –12 / 30 |
| | | *Energies (kcal mol$^{-1}$)* [b] | | |
| Anti/Syn | 0.0 / –0.3 | –0.4 / –0.6 | –1.1 / 1.3 | –0.8 / 1.7 |
| Barrier [c] | 0.3 / –2.0 | 0.0 / –2.1 | –0.6 / –0.7 | –0.9 / –1.2 |
| Profile | 0.3 / 2.5 | 1.2 / 2.8 | 0.9 / 4.0 | 0.9 / 3.9 |
| | | Phase angle (P) | | |
| | | *Geometries (°)* [a] | | |
| North | 10 / 30 | 10 / 10 | 10 / 40 | 0 / 10 |
| East | 0 / 10 | 0 / 0 | 10 / 10 | 0 / 10 |
| South | 0 / 0 | 10 / 10 | 0 / 0 | 0 / 0 |
| | | *Energies (kcal mol$^{-1}$)* [b] | | |
| North/South | –0.1 / –1.5 | 0.0 / –1.0 | –0.6 / –1.6 | 0.5 / –0.5 |
| East Barrier | –0.2 / 0.4 | –0.5 /0.7 | –0.1 / 1.2 | –0.8 / 0.0 |
| Profile | 0.4 / 0.6 | 0.5 / 0.4 | 0.4/ 0.7 | 0.2 / 0.5 |

[a] Errors in the position of the minima and transition state when parmbsc1 (first number in the cell) or parmbsc0 (second number in the cell) values are compared with MP2 geometries. [b] Errors in the estimates of the relative stability and transition barrier when parmbs1 (first number in the cell) or parmbsc0 (second number in the cell) values are compared with single-point CCSD(T)/CBS results. [C] Energy values refer to barrier at χ around 120 degrees, note that the large barrier located at χ around 0 is very well reproduced at the parmbsc1 level, but very poorly at the parmbsc0 one (**Supplementary Fig. 24**).

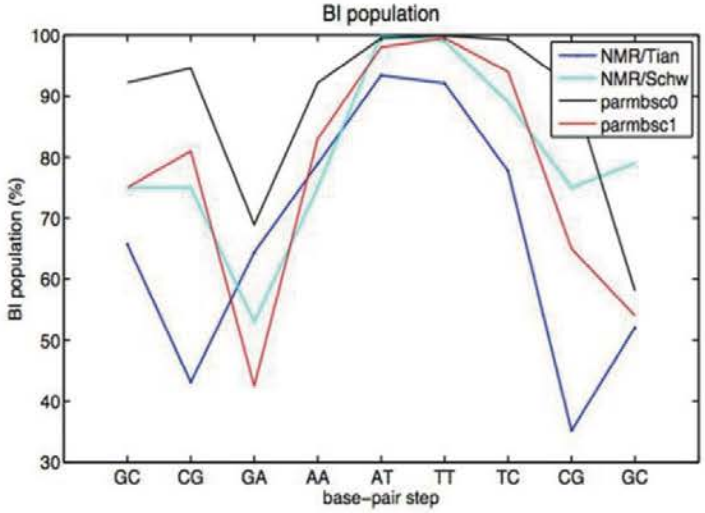
## SUPPLEMENTARY FIGURES




**Supplementary Figure 1| Helical parameters of DDD: Slide, Rise and grooves' width.** Comparison of slide, rise, major and minor groove width average values per base-pair step coming from NMR structure pdb: 1NAJ (blue), X-ray structure pdb: 1BNA (green), 1 μs run using parmbsc0 force-field (black) and 1.2 μs run using parmbsc1 force-field.

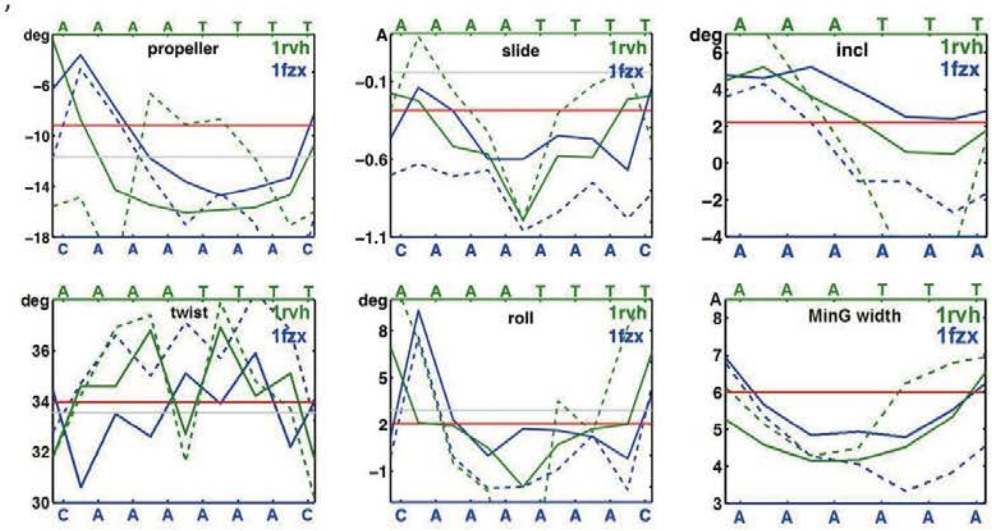



**Supplementary Figure 2| Helical parameters per base-pair of DDD.** Comparison of propeller twist, base opening, χ (chi) and pseudo-rotational angle (pucker) average values per base-pair step coming from NMR structure pdb:1NAJ (blue), X-ray structure pdb:1BNA (green), 1 μs run using parmbsc0 force-field (black), and 1.2 μs run using parmbsc1 force-field.

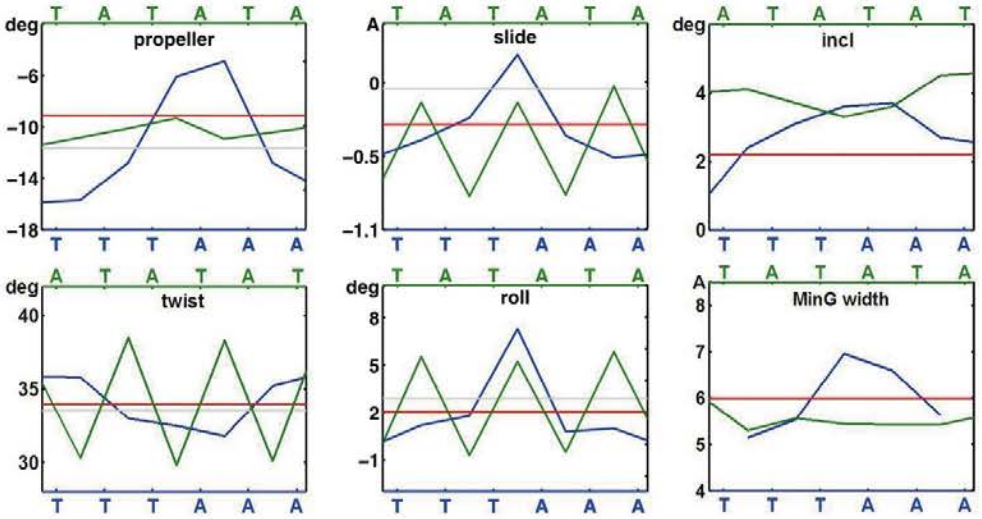



**Supplementary Figure 3| BI/BII populations of DDD.** Comparison of BI population percentage per base-pair step for DDD. Values coming from NMR/Tian *et al.*[1] (blue), NMR/ Schwieters *et al.*[2] (light blue), 1 μs run using parmbsc0 force-field (black) and 1.2 μs run using parmbsc1 force-field (red).

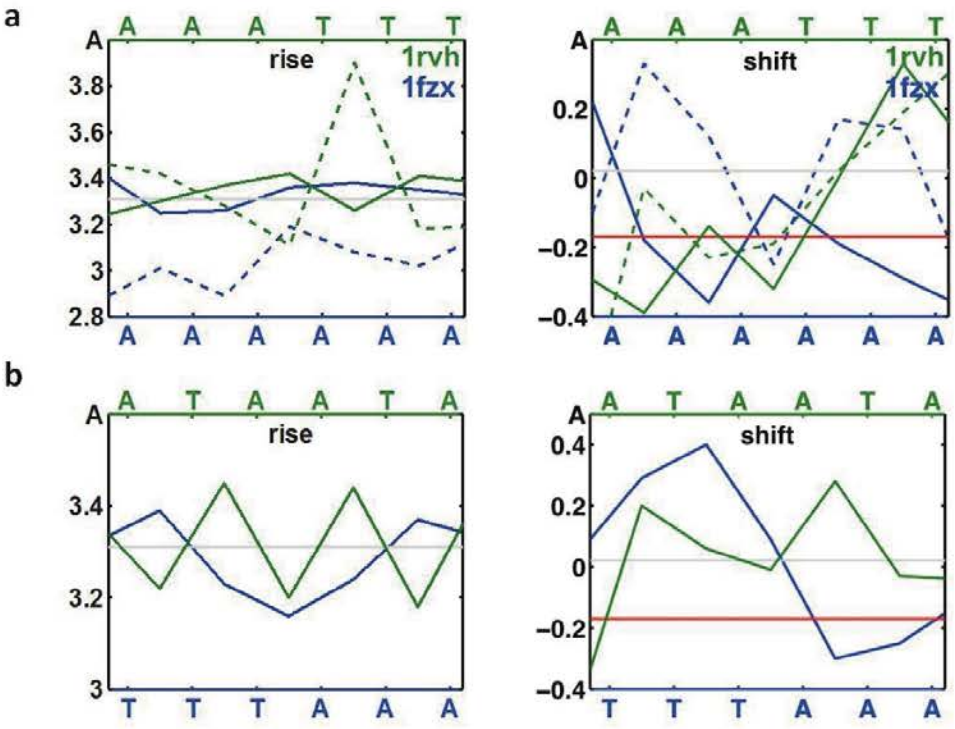
1. Tian, Y., Kayatta, M., Shultis, K., Gonzalez, A., Mueller, L.J., & Hatcher, M.E. *J. Phys. Chem. B***113**, 2596–2603 (2008).
2. Schwieters, C.D. & Clore, G.M., *Biochemistry***46**, 1152–1166 (2007).



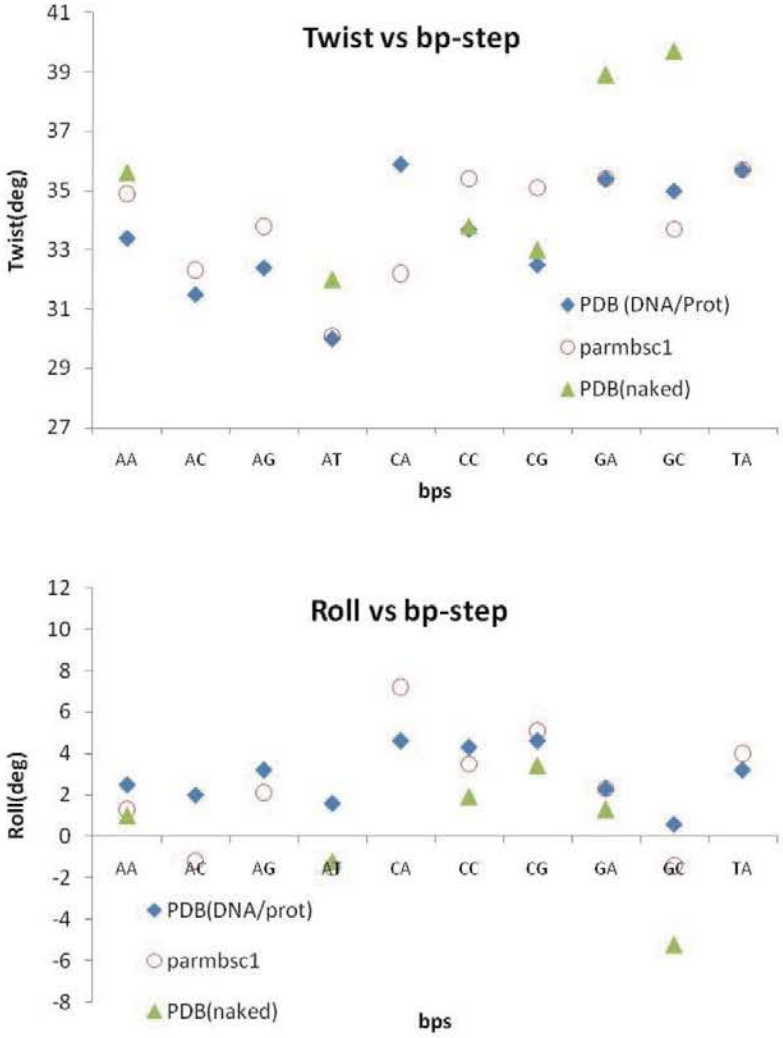

**Supplementary Figure 4| Helical parameters of A-tract sequences: AATT and AAAA.** Comparison in structural characteristics such as propeller twist, slide, inclination, twist, roll and minor groove width of values obtained using parmbsc1 force-field (full line) and experimental values (dashed lines) for AATT (pdb code:1RVH) (green) and AAAA (pdb code: 1FZX) (blue) sequences. Experimental average is represented with a grey line, while parmbsc1 average is represented with a red line.




**Supplementary Figure 5| Helical parameters of A-tract sequences: ATAT and TTAA.** Comparison in structural characteristics such as propeller twist, slide, inclination, twist, roll and minor groove width of values obtained using parmbsc1 force-field (full line) and experimental values (dashed lines) for ATAT (green) and TTAA (blue) sequences. Experimental average is represented with a grey line, while parmbsc1 average is represented with a red line.

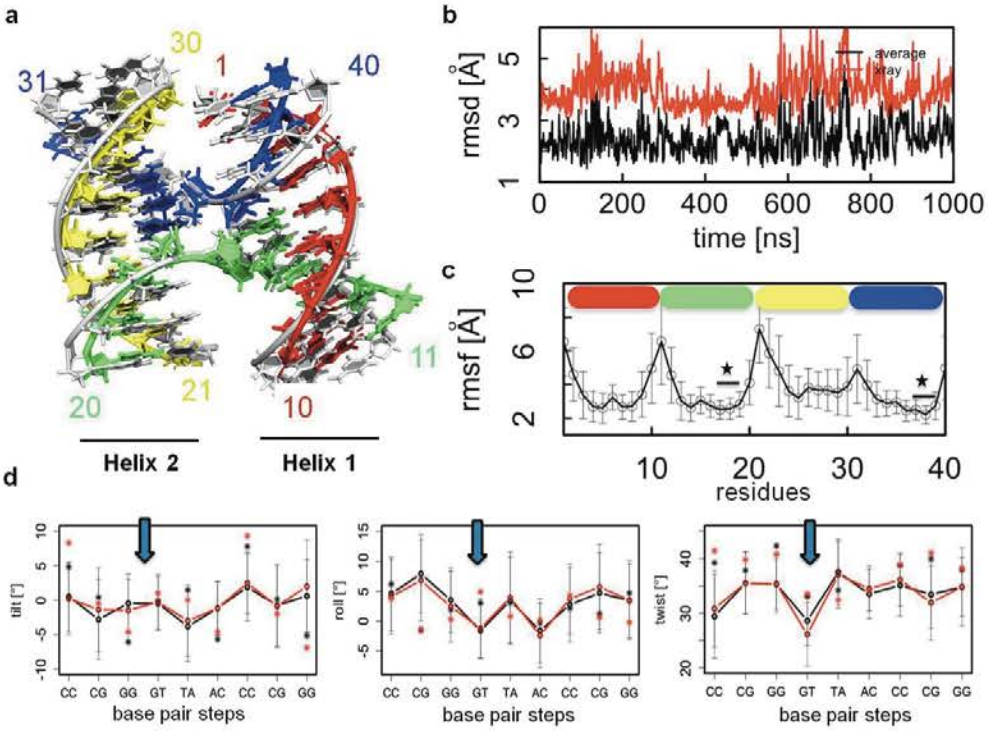



**Supplementary Figure 6| Base-pair step helical parameters of A-tract sequences.** Comparison in rise and shift of values obtained using parmbsc1 force-field (full line) and experimental values (dashed lines) for **(a)** AATT (pdb code:1RVH) (green) and AAAA (pdb code: 1FZX) (blue) and **(b)** ATAT (green) and TTAA (blue) sequences. Experimental average is represented with a grey line, while parmbsc1 average is represented with a red line.



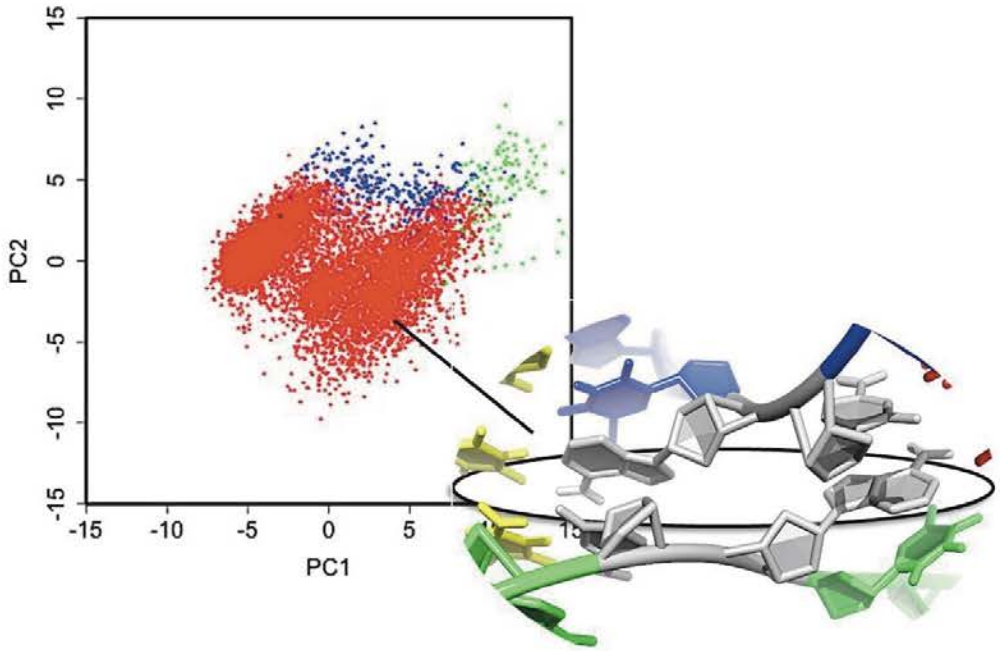

**Supplementary Figure 7| Sequence-dependent variability of twist and roll.** Comparison of DNA-protein complexes (blue), naked DNA (green) and parmbsc1 (red) values for twist (top) and roll (bottom) values per base-pair step. Values of DNA-protein complex come from analysis of 636 structures from PDB, while values of naked DNA come from analysis of 103 structures from PDB[1].

1. Dans, P.D., Pérez, A., Faustino, I., Lavery, R. & Orozco, M. *Nucleic Acids Res.***40**, 10668–10678 (2012).

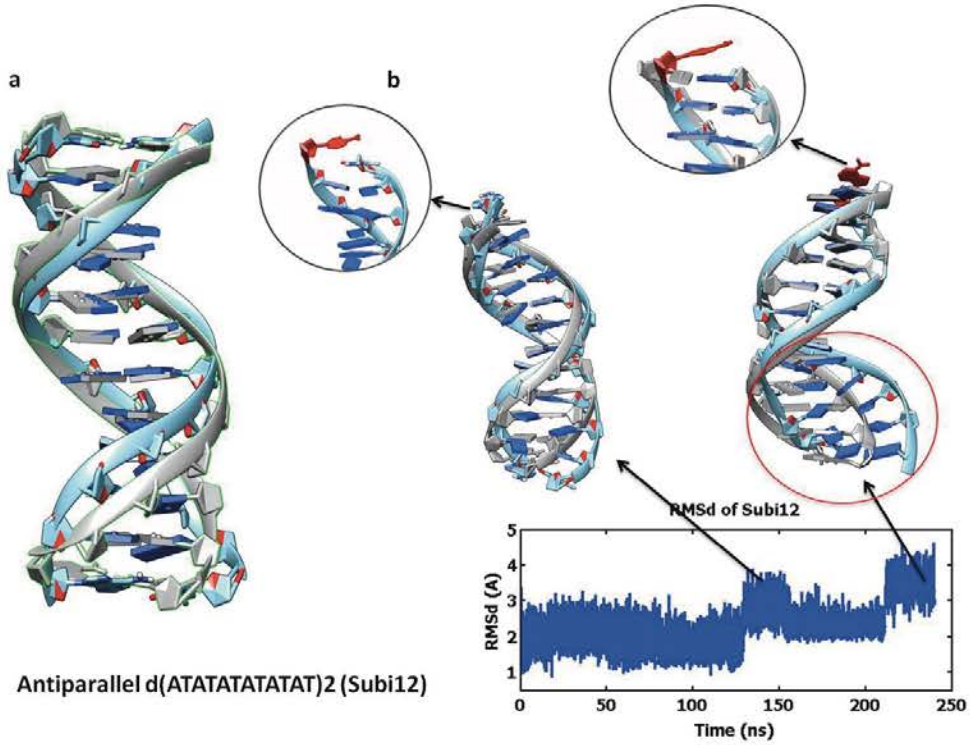



**Supplementary Figure 8| Holliday junction structural features are close to x-ray (1DCW) structure.** (**a**) Structural comparison of the time-averaged structure (in colors) with the x-ray reference structure (grey). (**b**) All heavy atoms RMSD and (**c**) per-residue RMSD from 1 μs MD simulation. X-ray structure was also taken as reference in the per-residue RMSD calculation. Note the higher RMSD values correspond to end strand bases. Starred residues are placed in the junction between helices. (**d**) Selected time-averaged helical parameters for the symmetric helices I and II. For experimental reference structures see ref. 1.

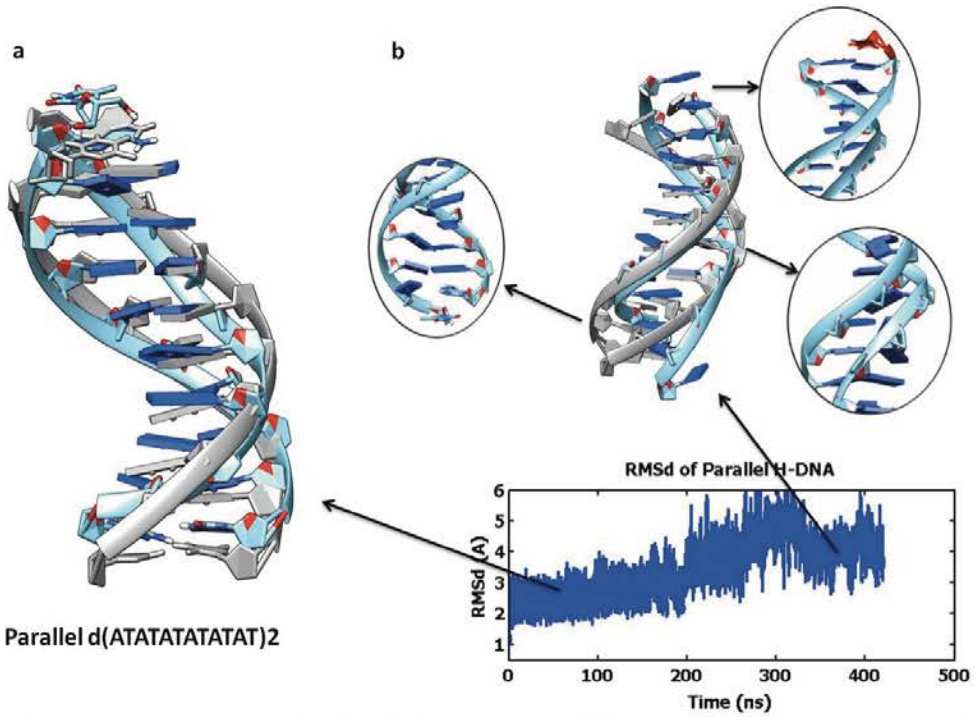
1. McKinney, S.A., Déclais, A.-C., Lilley, D.M.J. & Ha, T. *Nat. Struct. Mol. Biol.***10,** 93–97 (2003).



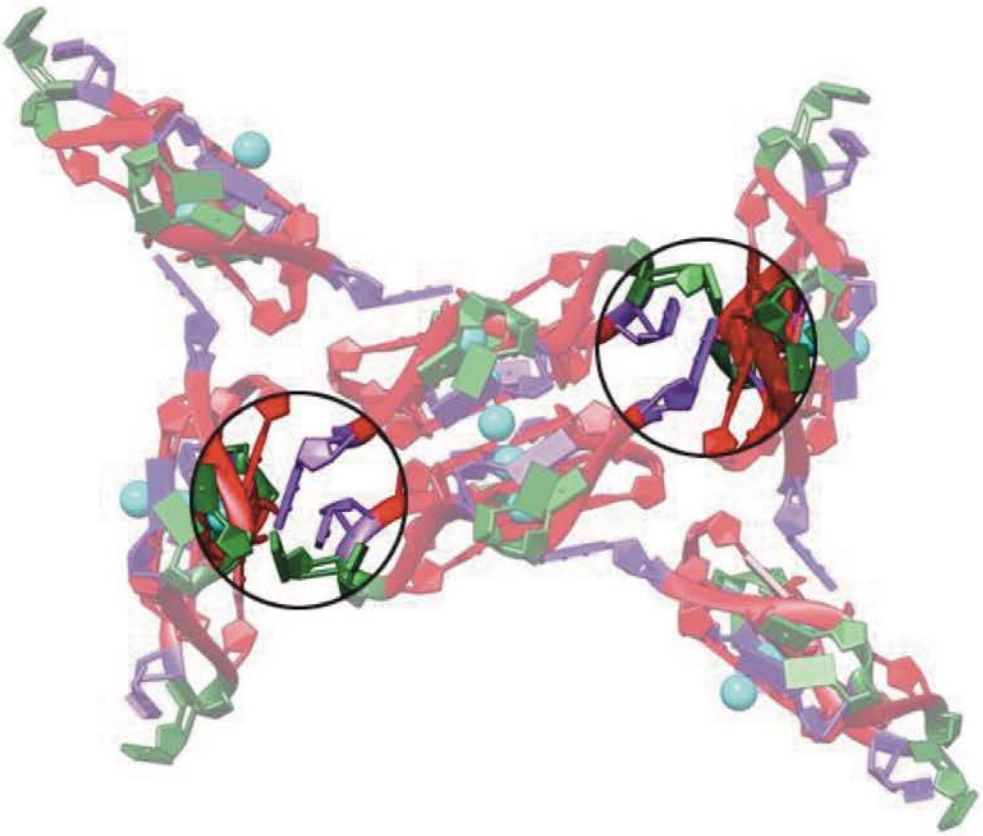

**Supplementary Figure 9| Holliday junction PCA results.** Projection to the first two PCA-eigenvectors based on the heavy atoms of junction bases (residues 16, 17, 36, and 37). The major conformation (in red) is present over ~95% of the simulation.



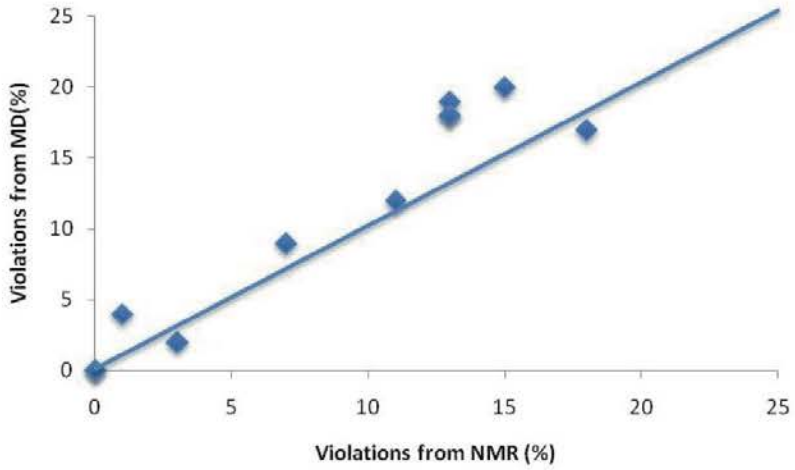

**Supplementary Figure 10| Simulation antiparallel of H-DNA. (a)** Comparison of experimental structure (made from pdb code: 1GQU) (grey) with the last snapshot of a 250 ns run using parmbsc1 (light blue). Bellow is an illustration of the duplex sequence. **(b)** RMSd of the 250 ns run with several snapshots plotted along the trajectory (light blue) compared with the experimental structure (grey) with highlighted distortions in the duplex.



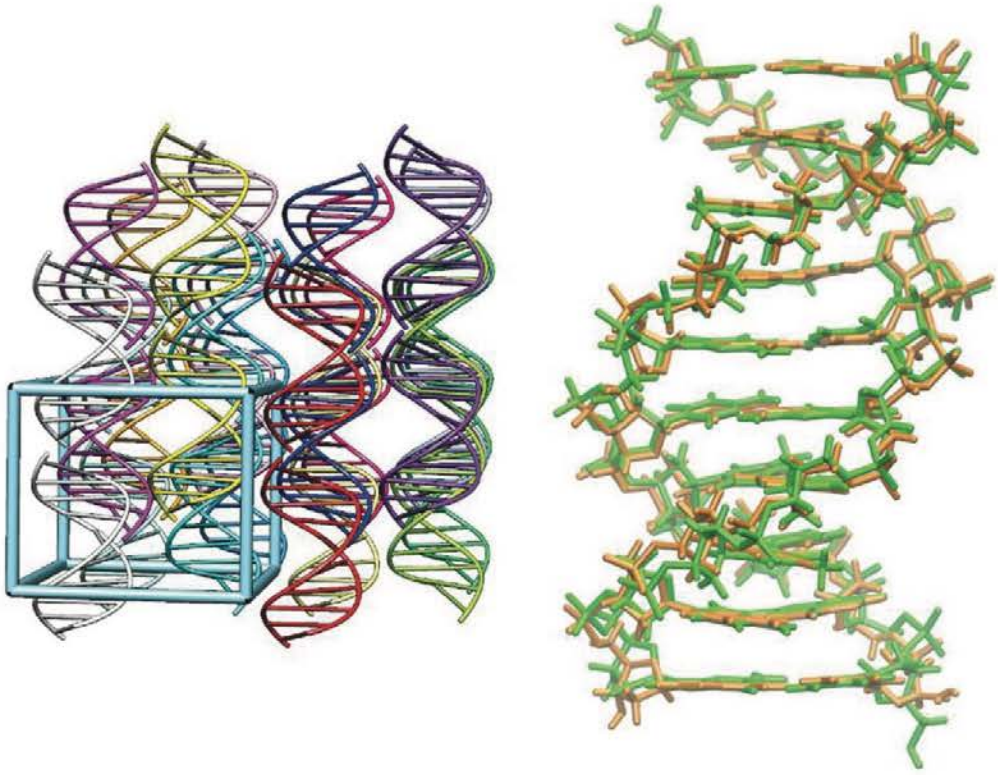

**Supplementary Figure 11| Simulation of parallel H-DNA. (a)** Comparison of experimental structure (grey) with a snapshot from a 400 ns run using parmbsc1 (light blue). **(b)** RMSd of the 400 ns run with several snapshots plotted along the trajectory (light blue) compared with the experimental structure (grey) with highlighted sever distortions in the duplex.



**Supplementary Figure 12| Crystal packing of Human Talomeric Quadruplex (HTQ).** Crystal packing of HTQ quadruplex (pdb code: 1KF1) showing interactions between loops' bases and other crystal units. Loop residues stacked to the neighboring units are highlighted in the circles.



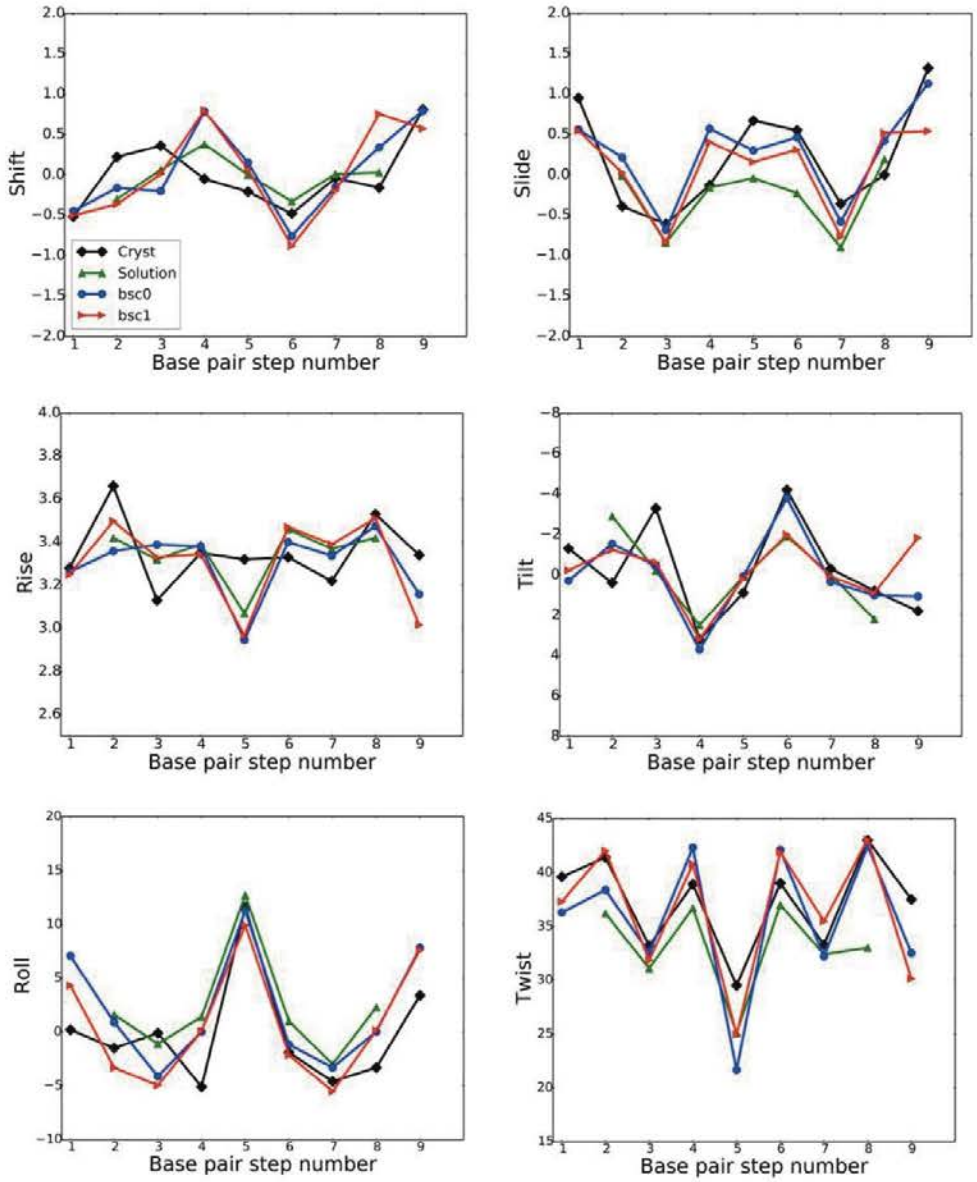

**Supplementary Figure 13| Correlation between the number of violations in NOE restraints found in MD-parmbsc1 trajectories and corresponding NMR models.** See **Supplementary Table 7** for details on structures.



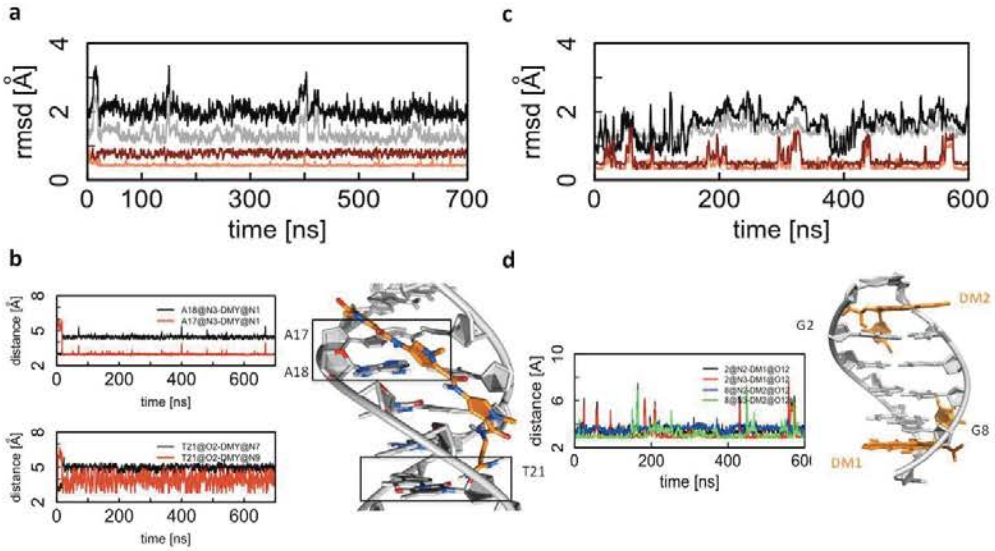
**Supplementary Figure 14| Representation of the crystal structure simulation of a B-DNA duplex (PDB: 1D23).** The simulation box used in the crystal simulations is shown on the left, while comparison between the best-fit average structure from parmbsc1 simulations (orange) and the crystal structure (green) are shown on the right. Note that the RMS deviation for all DNA heavy atoms of the simulation average structure (compared to the PDB structure) is 0.70 Å. This can be compared to 0.77 Å for a crystal simulation using parmbsc0, and 1.83 Å for a solution simulation also using parmbsc0[1].

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

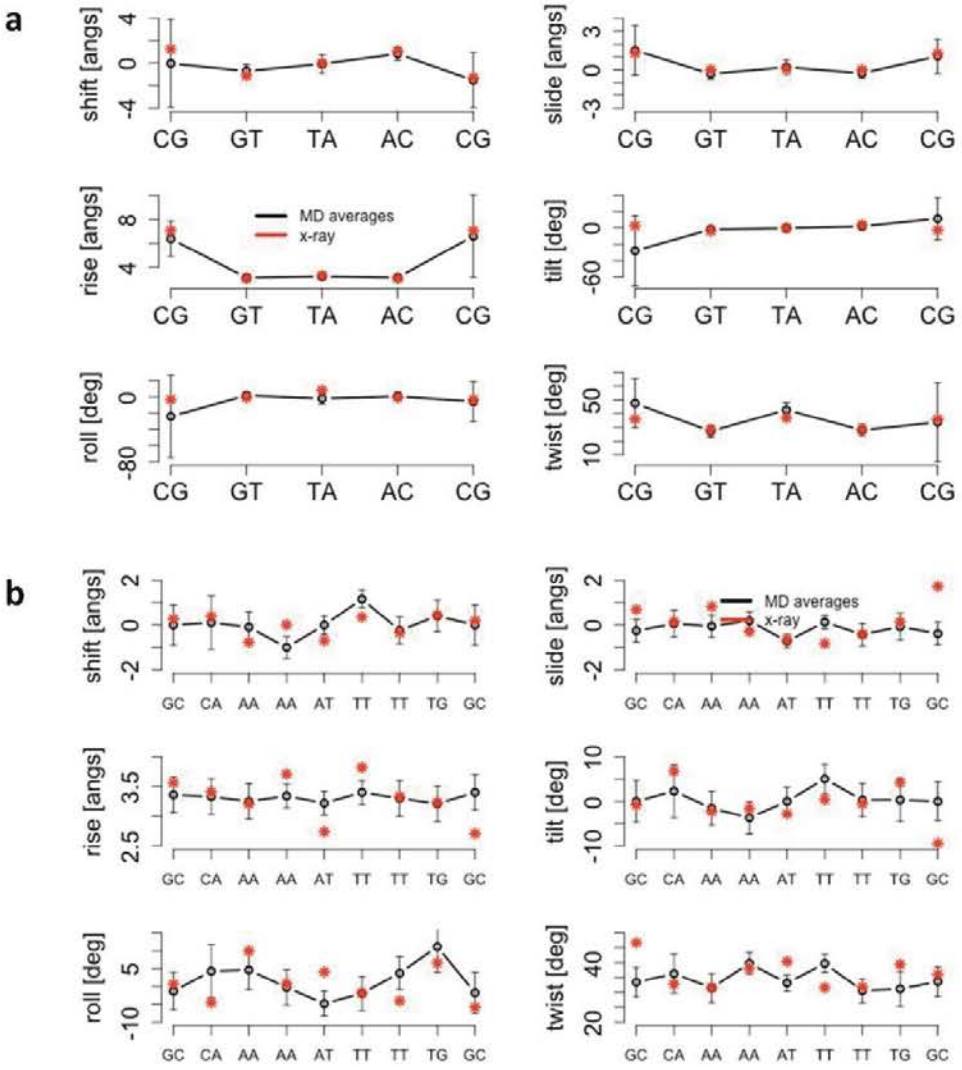
.



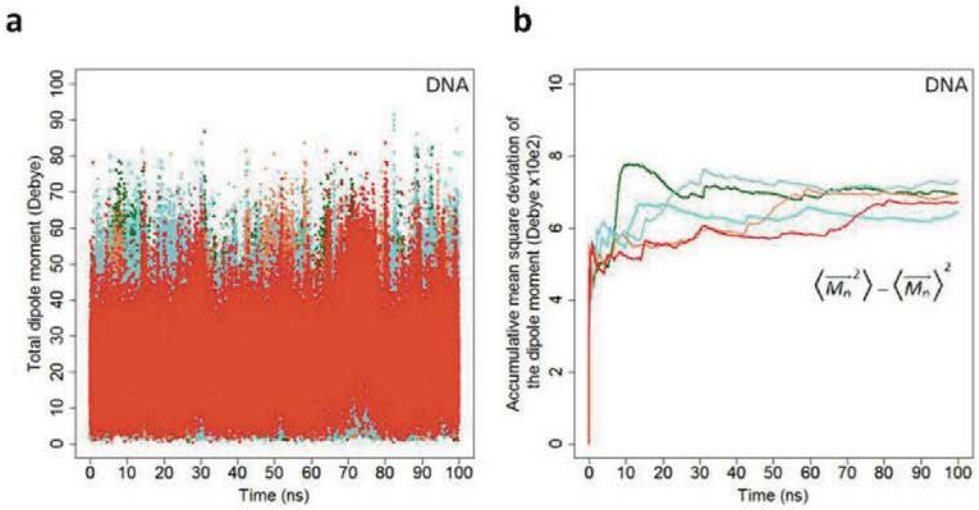

**Supplementary Figure 15| Helicoidal analysis of a simulation of a B-DNA duplex (PDB: 1D23) within crystal environment.** Helical parameters comparing results from simulation using parmbsc0 (blue) and parmbsc1 (red) force-fields, a simulation in solution (green) and the crystal structure (black).




**Supplementary Figure 16| Representative stability properties in drug-DNA complexes with parmbsc1.** RMSD (**a**) and representative distance between the distamycin A and the closest residues. (**b**) RMSD plots relative to x-ray (PDB id: 2DND), and MD-average structures for DNA (black and grey respectively) and distamycin A (red and orange respectively). Original contacts with the DNA are rapidly replaced by neighboring atoms keeping distamycin A within the minor groove. RMSD (**c**) and representative distances between the first daunomycin (PDB id: 1D11) and the closest guanine. (**d**) Second daunomycin's RMSd values are similar. Stabilizing interactions (h-bonds) between the N3 of guanine (residues 2 and 8 respectively) and a hydroxyl group in the daunomycin were stable along time.



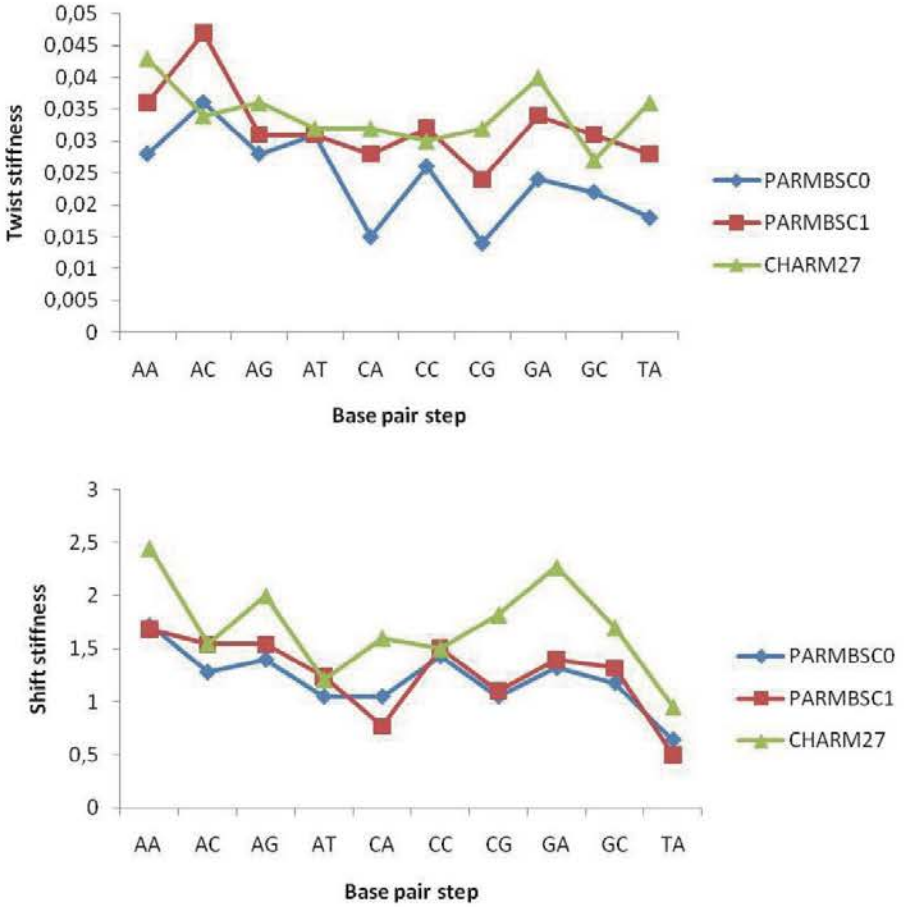

**Supplementary Figure 17| Representative helical base pair step parameters in drug-DNA complexes.** Time-averaged values associated to the DNA in complex with daunomycin (**a**) and distamycin A (**b**) in black compared with the original values from the X-ray structures (red, PDB id: 1D11 and 2DND for daunomycin and distamycin respectively).




| Component | Volume ($nm^3$) | Dielectric constant parmbsc1 ($\varepsilon_r$) | Dielectric constant AFM ($\varepsilon_r$) |
|---|---|---|---|
| DNA | 10.1 ± 0.2 | 8.0 ± 0.3 | 8.5 ± 1.4 |
| $DNA_{sugar-phosphate}$ | 3.4 ± 0.2 | 19.9 ± 1.1 | |
| $DNA_{sugar-base}$ | 7.6 ± 0.2 | 2.1 ± 0.2 | |

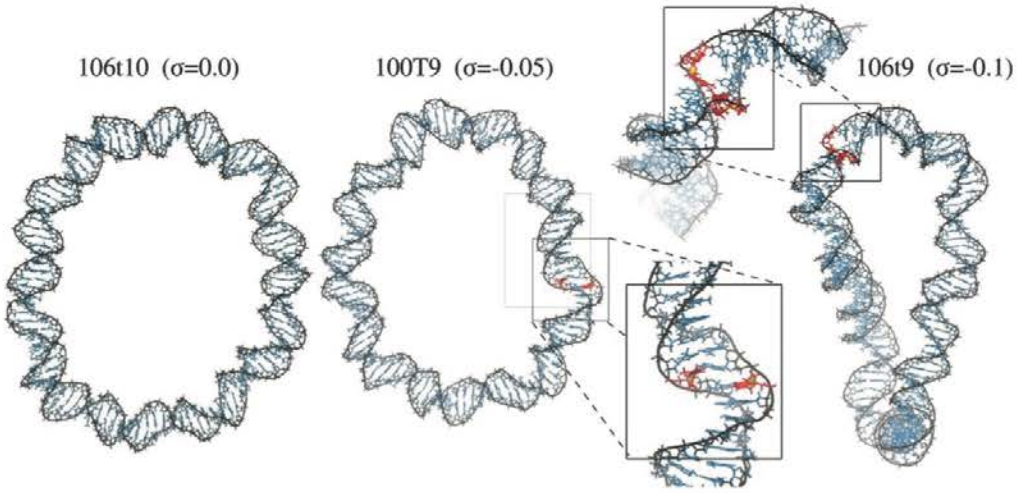
**Figure 18| DNA dielectric constant.** (**a**) Total dipole moment over time for 5 different replicas (100 ns each) taken from the microsecond long DDD simulation. (**b**) Accumulative mean square deviation of the dipole moment for the five replicas showing fairly good convergence after 30–40 ns. Values of whole DNA, sugar and phosphate groups, and sugar and base contributions are shown in the table below. See ref. 1 for the detailed procedure followed herein.

1. Cuervo, A., Dans, P. D.*et al. Proc. Natl. Acad. Sci.***111,** E3624–E3630 (2014).

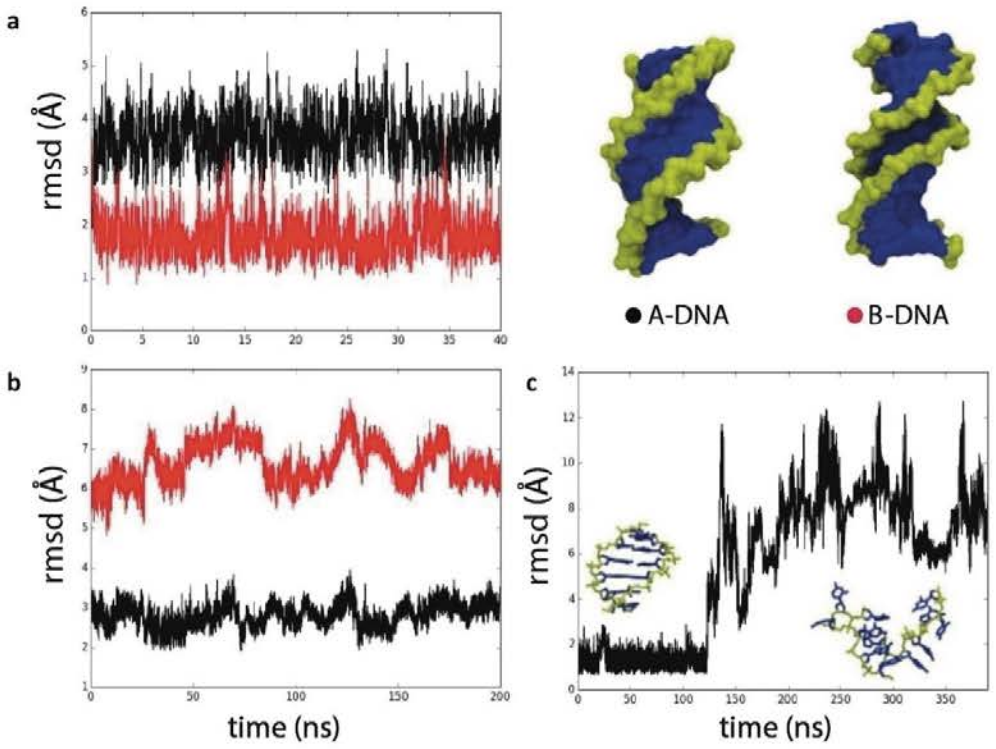



**Supplementary Figure 19| Sequence dependent helical deformability.** Variability of Twist (top) and Shift (bottom) stiffness constants for 10 unique base-steps. Parmbsc0 and CHARMM27 values are taken from ref 1.

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

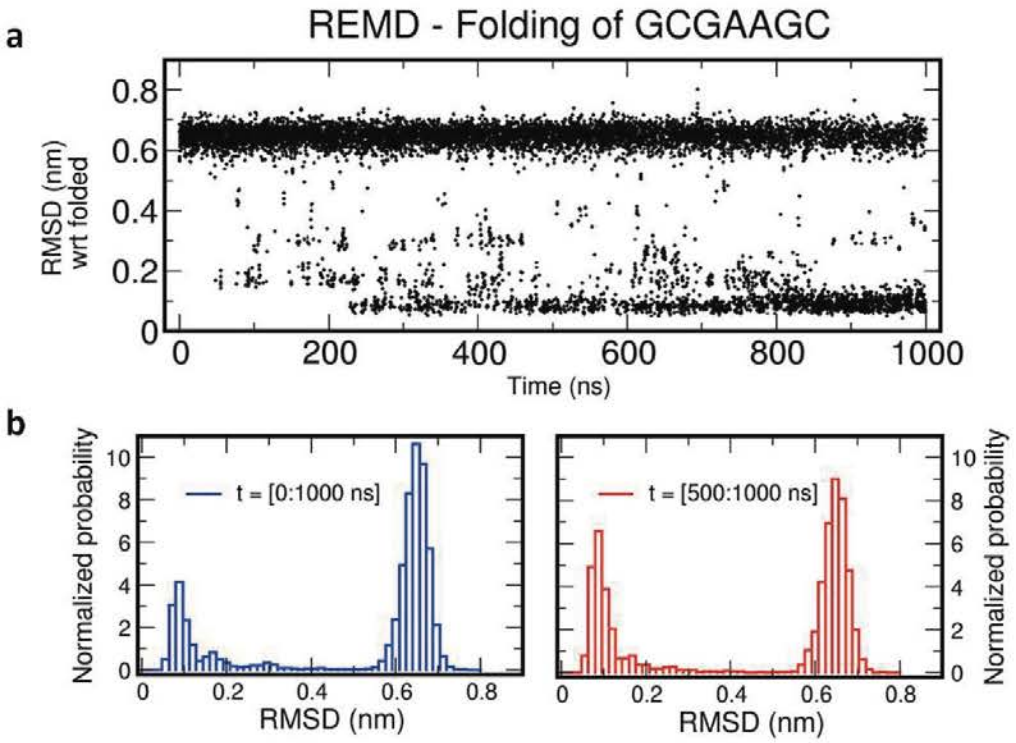



**Supplementary Figure 20| Analysis of DNA minicircles.** Final frames of the minicircles MD simulations. The secondary structure of the relaxed loop with 106 bp and 10 helical turns (106t10) remains intact, while the 2 negatively supercoiled circles show significant denaturalization. The 100 bp circle with 9 turns (100t9) presents 2 adjacent pyrimidine base-flipping towards the major groove, and the 106 bp circle with 9 turns (106t9), denature over multiple consecutive base pairs.



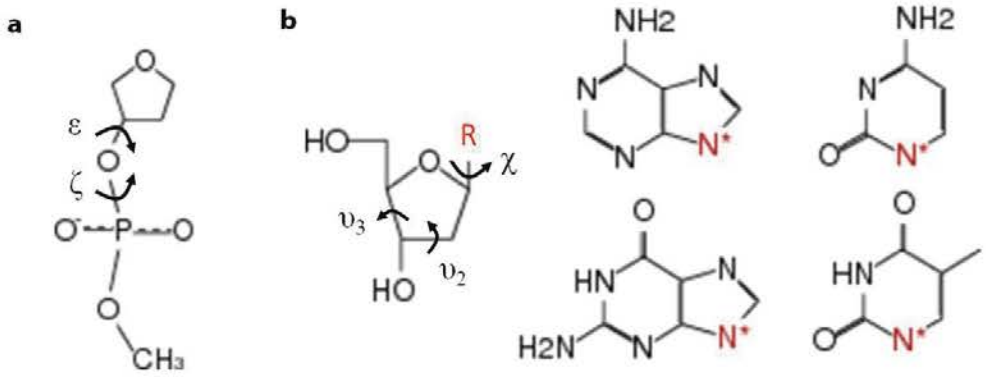

**Supplementary Figure 21| MD simulations of conformational changes.** (**a**) A to B transition simulation of DDD, where A-DNA form is presented in black with B-DNA in red. (**b**) Simulation of DDD in mixture of water and ethanol (see refs. 1 y 2 for additional discussion). (**c**) Unfolding of $d(GGCGGC)_2$ in 4 M pyridine water solution[3].

1. Soliva, R., Luque, F.J., Alhambra, C. & Orozco, M. *J. Biomol. Struct. Dyn.***17,** 89–99 (1999).
2. Ivanov, V.I., Minchenkova, L.E., Minyat, E.E., Frank-Kamenetskii, M.D. & Schyolkina, A.K. *J. Mol. Biol.***87,** 817–833 (1974).
3. Perez, A. & Orozco, M. *Angew. Chemie Int. Ed.***49,** 4805–4808 (2010).

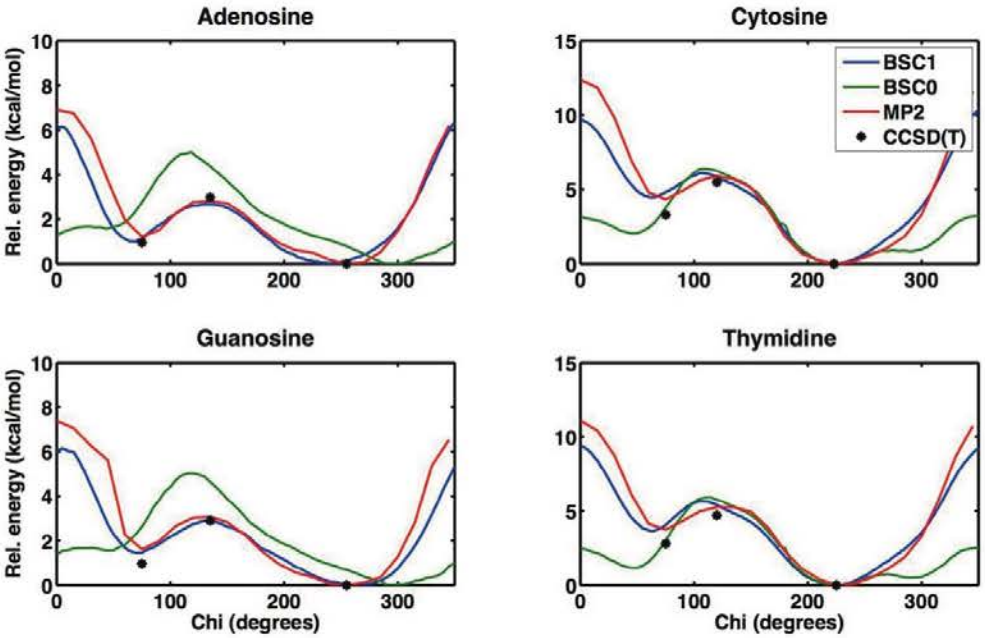



**Supplementary Figure 22| Hairpin folding.** Replica exchange MD (REMD) simulations of the folding of the small hairpin d(GCGAAGC) in water using parmbsc1 force-field. (**a**) RMSD with the respect to the folded state. (**b**) Probabilities of RMSDs in whole (blue) and second part (red) of microsecond runs of REMD. Structures are clearly recognizing the folded conformation and keeping it. For technical details see reference 1.

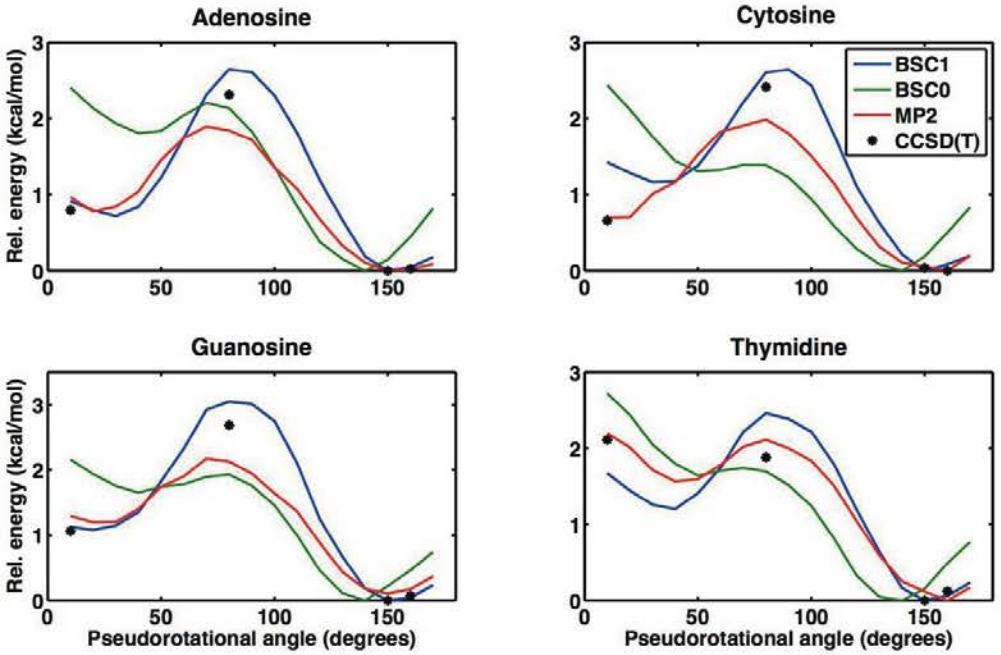
1. Portella, G., Orozco, M. *Angewandte chemie Int. Ed.***49**, 7673–7676 (2010).



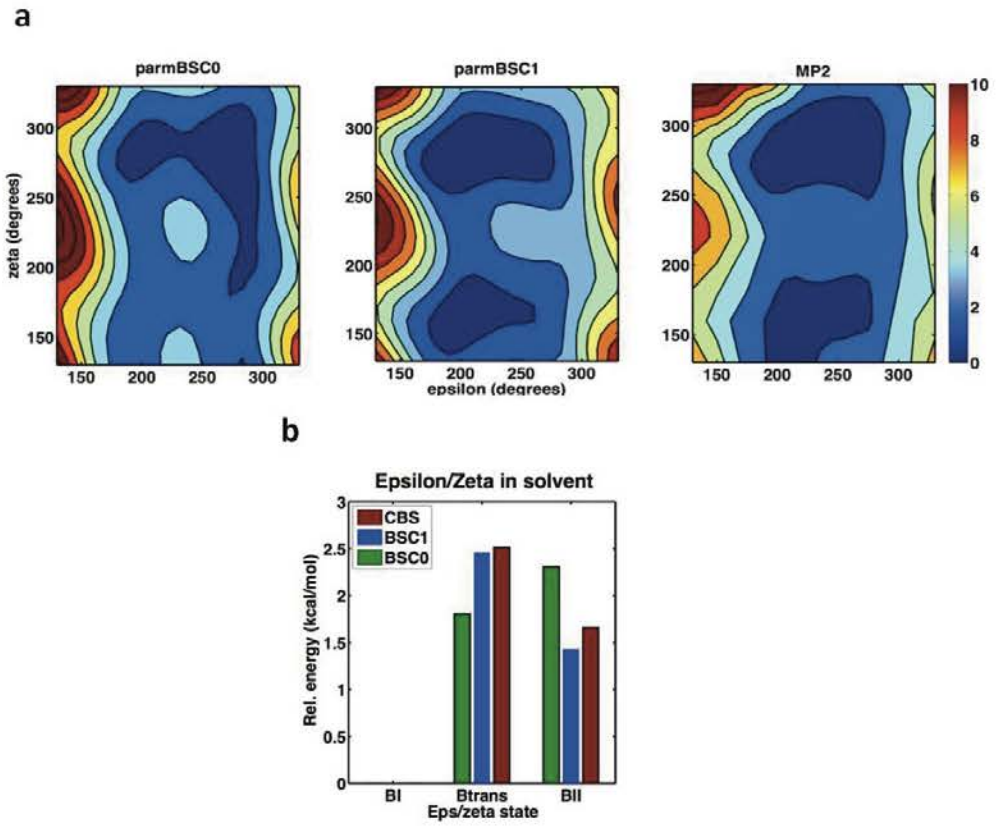

**Supplementary Figure 23| Model compounds used in QM optimization.** (**a**) Compound used for ε/ζ parameterization. (**b**) Compounds used for χ and sugar puckering parameterizations, where R represents the base, shown on the right.



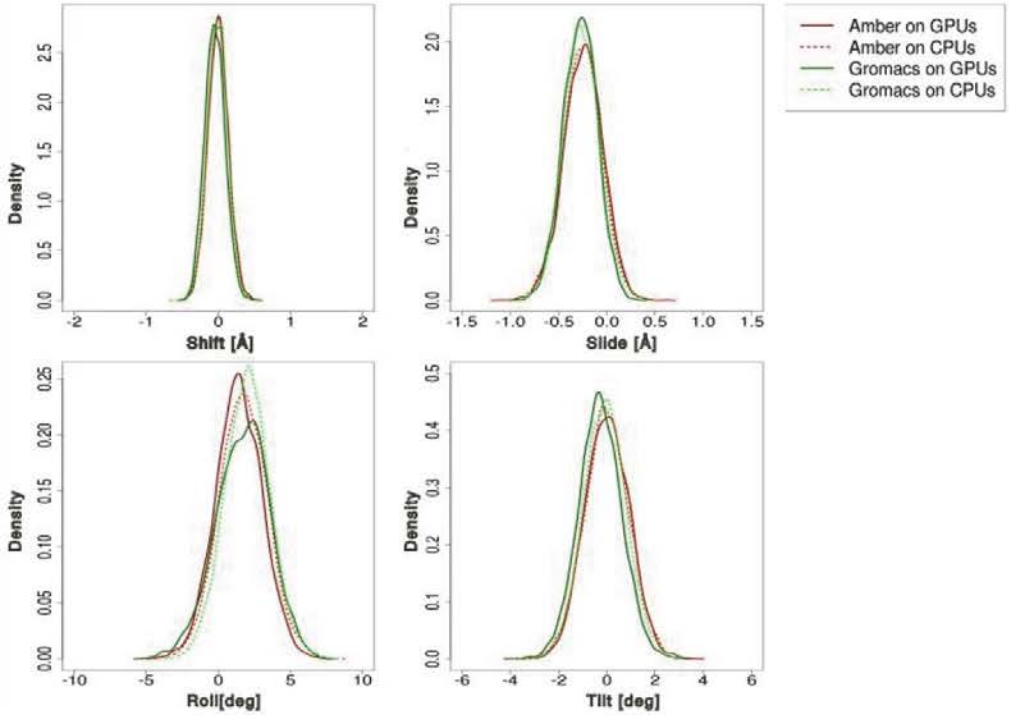

**Supplementary Figure 24| Profiles of χ (chi) dihedral for 4 DNA bases in solution.** Comparison of profiles obtained from QM using MP2/aug-cc-pVDZ (red) method with solvent corrections (Supplementary Notes), and PMF profiles using parmbsc0 (green) and parmbsc1 (blue) force-fields. Complete basis set (CBS) values for specific points are represented with a black dot.



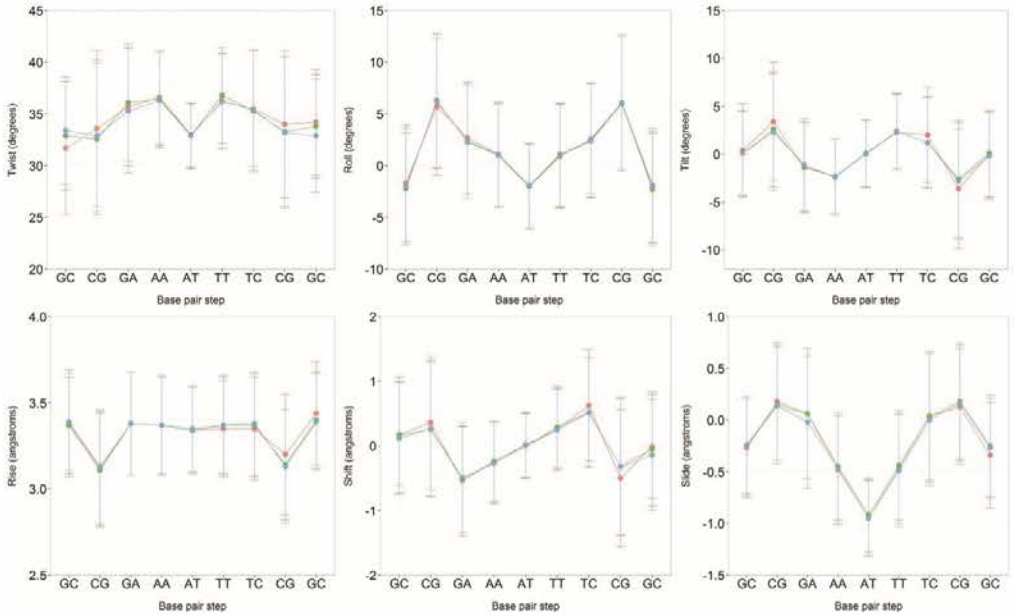

**Supplementary Figure 25| Profiles of pseudorotational angle for 4 DNA bases in solution.** Comparison of profiles obtained from QM using MP2/aug-cc-pVDZ (red) method with solvent corrections (Supplementary Notes), and PMF profiles using parmbsc0 (green) and parmbsc1 (blue) force-fields. Complete basis set (CBS) values for specific points are represented with a black dot.



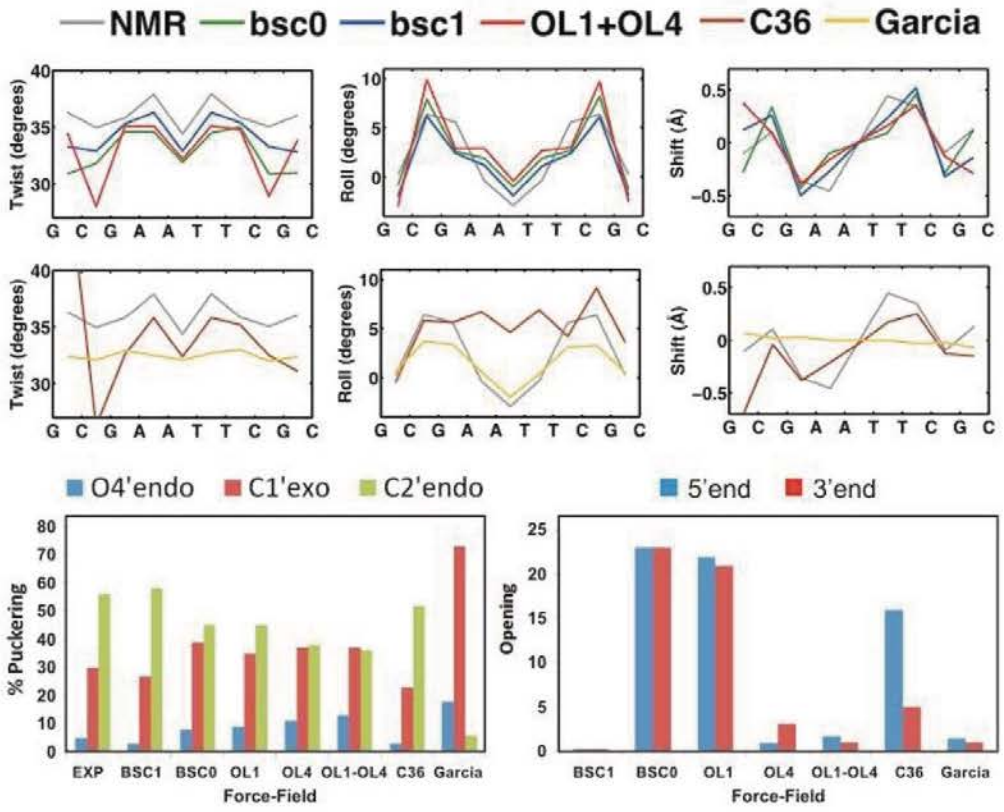

**Supplementary Figure 26| ε/ζ (epsilon/zeta) profiles in solution.** (**a**) Contour profiles of epsilon/zeta from QM calculations using MP2/aug-cc-pVDZ method (right), and PMF profiles using parmbsc0 (left) and parmbsc1 (middle) force-fields. Energies are given in kcal $mol^{-1}$ and the color bar goes from blue (0 kcal $mol^{-1}$) to red (10 kcal $mol^{-1}$). (**b**) Values at key points of the profile comparing parmbsc0 (green), parmbsc1 (blue) and complete basis set (CBS) (dark red) values.



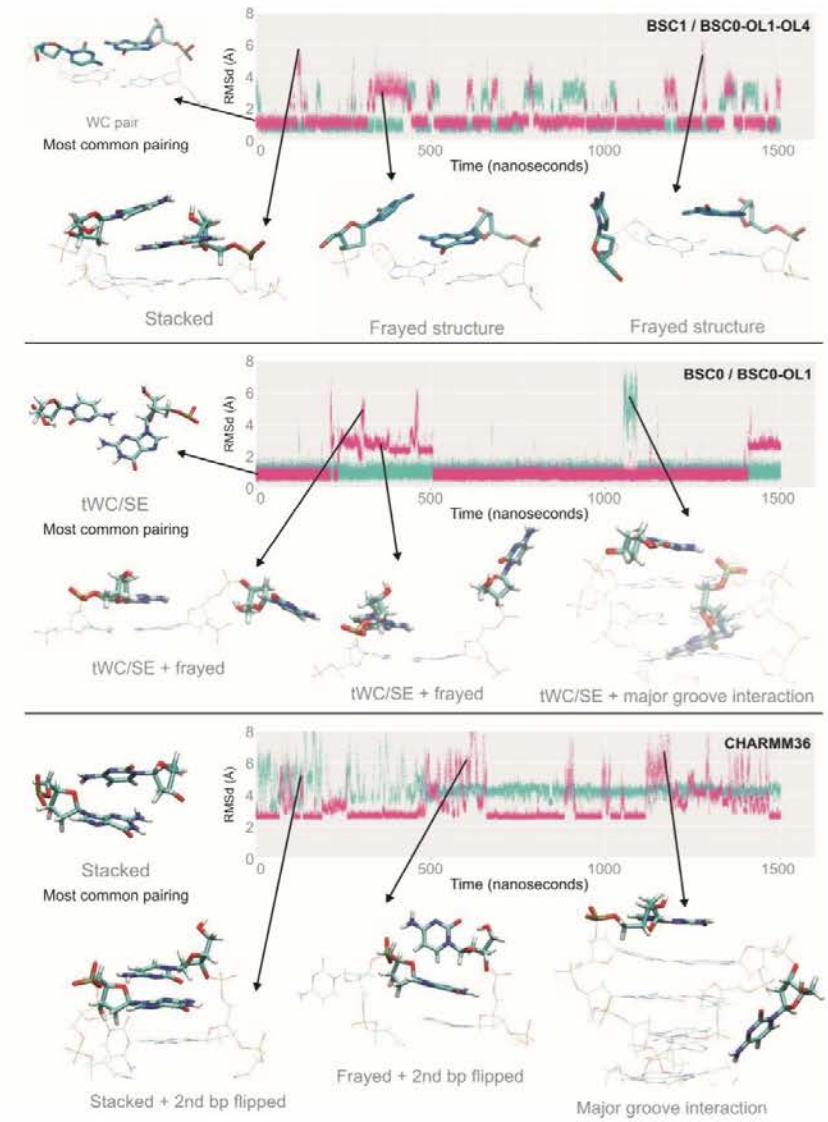

**Supplementary Figure 27| Using DDD to compare different simulation engines.** Normalized distributions of the helical parameters shift, slide, roll and tilt are shown for the four MD simulations (AMBER vs GROMACS, and GPU vs CPU codes). Due to the shortness of the simulation runs (100 ns), slight differences in roll angle can be detected using different MD engines.



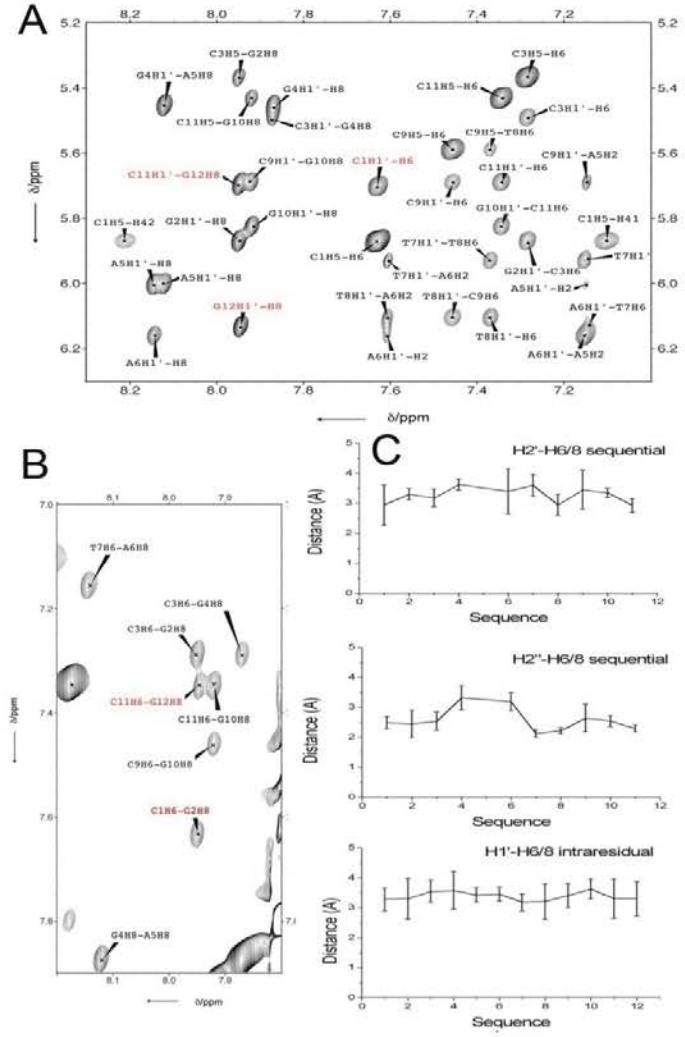

**Supplementary Figure 28| Variation of helical parameters along the sequence for 2 µs of MD simulation of DDD with added salt (NaCl) concentrations**: minimum $Na^+$ for neutrality (green), 150 mM (red) and 500 mM (blue). PME was used in all the cases.




**Supplementary Figure 29| Structural characteristics of DDD in MD simulations with different force-fields**. First row variation of key helical coordinates along sequence in parmbsc0, parmbsc1 and parmbsc0-OL1+OL4 (those force-fields providing the best average parameters in **Supplementary Table 2**). Second raw correspond to force-fields providing less accurate average values in **Supplementary Table 2** (CHARMM36 and parmbsc0-Cheng-Garcia). In these two rows only the 10 mer segment is shown (to avoid dramatic scale bias in case of fraying of terminal bases), and only NMR results are used as reference (to make more clear the plots; note that nearly identical profiles are obtained from X-Ray (see **Fig. 1**)). The third row corresponds to the distribution of sugar puckering (taking as experimental reference the average of NMR and X-Ray structures) and the average opening at the terminal basis. The superior behavior of parmbsc1 is evident in all plots, as well the prevalence of fraying artifacts for some of the force-field, and the presence of non-negligible distortions in CHARMM36 and parmbsc0-CG trajectories, even for the central portion of the helix.




**Supplementary Figure 30| Details of the evolution of the terminal base pairs**. RMSd of the terminal base pairs (C1:G24 in pink and G12:C13 in cyan) along 1.5 μs of MD trajectories. Firs row: profiles for a force-field showing no fraying artifacts (but indeed frequent short-living openings) such a parmbsc1 (parmbsc0-OL1+OL4 and parmbsc0-OL4 provide similar profiles, while parmbsc0-CG (Cheng-Garcia) shows completely frozen terminal base pairs). Second row: profile for a force-field like parmbsc0 which suggest fraying and the formation of unusual contacts (parmbsc0-OL1 provides identical profiles) with tWC pairing and *syn* nucleotides. Third row: profiles obtained for CHARMM36, where despite the center of the duplex is well conserved terminal Watson-Crick pairings are mostly lost and substituted by a myriad of alternative contacts. In all cases structures sampled along specific time frames are shown.




**Supplementary Figure 31| NOE data on the terminal base steps of DDD.** A) H1´-aromatic region of the NOESY spectra of DDD (mixing time 200 ms, buffer conditions 125 mM NaCl, 25 mM sodium phosphate, pH 7, T = 25 º C). Some relevant cross-peaks involving terminal residues are labeled in red color. B) Aromatic-aromatic region of the NOESY spectra (same experimental conditions). Note that NOE intensities involving terminal residues (i.e. C1H6-G2H8, C11H6-G12H8 in red) are not significantly lower than those involving central residues, indicating that the terminal bases remain stacked on top of their neighbors. C) Some experimental distances obtained from a full relaxation matrix analysis of the NOE data *vs* sequence. Sequential H2´-H6/8 and H2"-H6/8 do not exhibit dramatic changes for the terminal base steps, indicating that the fraying effect in these residues is not significant under these experimental conditions. All intra-residual H1´-H6/8 distances, including the terminal base residues, are around 3-4 Å, characteristic of glycosidic angle conformation in *anti*.

## 4.2 Drew-Dickerson dodecamer dynamics (Publication 2)

After developing and thoroughly testing our new force field, we decided to take a more extensive look into the Dickerson–Drew dodecamer (DDD), one of the most studied DNA sequences with over 60 entries in PDB database. DDD is a prototypic B-DNA molecule of a palindromic sequence $d[CGCGAATTCGCG]_2$, which is indisputably the most studied (theoretically and experimentally) oligo in the history (Pérez, Luque, et al. 2007; Drsata et al. 2012).

Previous MD studies of DDD pointed to some of the problems of parmbsc0 force field that were addressed with parmbsc1. As noted above, notable imperfections in the case of canonical B-DNA were the underestimation of some of the helical parameters, improper description of the bimodality of the distribution of certain helical parameters (for example twist for CG step) (Heddi et al. 2008), and excessive distortions at the ends of the duplexes, namely tWC/SE conformation (where cytosine is turned around the glycosidic bond into *syn* conformation to form a non-WC pair resembling the *trans* WC/Sugar edge C•G pair, which occurs extremely rarely in experiments; see (Drsata et al. 2012) for more details). This artifact generated severe end-effects, especially changes in the profiles of twist in several base pairs away.

The idea behind this study was to check if the good performance of parmbsc1 would be preserved in a longer timescale and possibly capture some slow conformational changes that are not visible on 1μs timescale. For that reason, we performed an extended (10 μs) simulation of DDD under physiological conditions, and a variety of control, shorter simulations varying environmental conditions. Helical parameters analysis was able to capture some correct details of DDD like sequence variability of twist, $B_I/B_{II}$ populations, and higher χ values (*high-anti*) for guanosine (see Supplementary Figure S17 in the following publication). Twist distribution for CG base pair step shows a clear bimodality (see Supplementary Figure S20 in the following publication), as it has a high propensity for $B_I \leftrightarrow B_{II}$ transitions (see Supplementary Figure S19 in the following publication). The analysis of terminal residues from this long trajectory shows a dramatic decrease in terminal opening with most of the fraying being transient sampling of large opening angles (see Figure 8 in the following publication), with very few event where χ values explored *syn* conformation and almost none (just one short instance) of the tWC/SE conformation (described above) that caused distortions in helical description when parmbsc0 was used. Lastly, the 10 μs trajectory allowed us to explore convergence issues on a different level than before. We analyzed the trajectory in 1, 2 and 5 μs segments doing principle component analysis, where we observed small divergence between 1 μs segments, but no significant differences (see Supplementary Figure S21 in the following publication). Similar approach was done in entropy calculations, where we observed difference less than 2% difference between 2 μs segments,

regardless of the method used to calculate the entropy (see Supplementary Figure S22 in the following publication).

Another concern regarding parmbsc1 proficiency was that its performance would be biased toward the selection of the solvent and ion models. We considered two most popular water models: TIP3P and SPC/E, while for salt ($Na^+Cl^-$ or $K^+Cl^-$) we considered models by Smith-Dang, Joung-Cheatham, Jense-Jorgensen and Beglov-Roux (see the Methods section in the following publication), ranging from 0.15 M to higher salt concentrations (up to 2 M). Averaged base pair step profiles show no significant differences between the ion and solvent model used, suggesting a robustness that parmbsc1 has with respect to the selection of ion and solvent force fields. To check the similarity in DNA flexibility between the different simulations of DDD, we performed essential dynamics analysis and computed the similarities for the central 10 bp of DDD (see Supplementary Figure S3 in the following publication). We concluded that the dynamics obtained with parmbsc1 has above 75% similarity with respect to the reference simulation done with parmbsc0 force field. This similarity was even higher (>90%) if we considered energy-averaged indices. Moreover, helical stiffness analysis confirmed the similarity between parmbsc0 and parmbsc1.

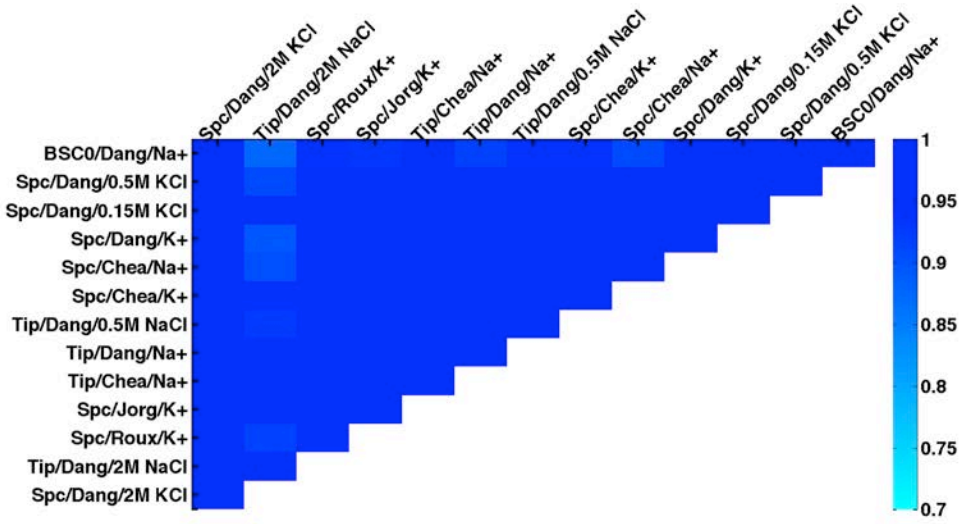



**Figure 4.2. Relative energy-weighted similarity index matrix between trajectories of DDD in different environments.** For more details on similarity indexes calculation see (Pérez et al. 2005).

In summary, this study provided additional confirmation of the good performance of parmbsc1 force field, even on 10 μs level. Careful examination was used to characterize slow and infrequent conformational changes in DDD, leading to the identification of previously uncharacterized conformational states of this duplex, which can explain biologically relevant conformational transitions. With a total of

more than 43 μs of unrestrained molecular dynamics simulation, this study is the most extensive investigation of the dynamics of DDD published at that time.

Nucleic Acids Research Advance Access published April 15, 2016

*Nucleic Acids Research, 2016* **1**
*doi: 10.1093/nar/gkw264*

# Long-timescale dynamics of the Drew–Dickerson dodecamer

**Pablo D. Dans[1,2], Linda Danilāne[1,2,3], Ivan Ivani[1,2], Tomáš Dršata[4], Filip Lankaš[4,5], Adam Hospital[1,2], Jürgen Walther[1,2], Ricard Illa Pujagut[1,2], Federica Battistini[1,2], Josep Lluis Gelpí[6], Richard Lavery[7] and Modesto Orozco[1,2,6,*]**

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona, Spain, [2]Joint BSC-IRB Research Program in Computational Biology, Baldiri Reixac 10-12, 08028 Barcelona, Spain, [3]School of Chemistry, University of East Anglia (UEA), Norwich Research Park, Norwich NR4 7TJ, UK, [4]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo nám 2, 166 10 Prague, Czech Republic, [5]Laboratory of Informatics and Chemistry, University of Chemistry and Technology Prague, Technická 5, 166 28 Prague, Czech Republic, [6]Department of Biochemistry and Molecular Biology, University of Barcelona, 08028 Barcelona, Spain and [7]Bases Moléculaires et Structurales des Systèmes Infectieux, Université Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, Lyon 69367, France

Received December 29, 2015; Revised March 25, 2016; Accepted March 31, 2016

## ABSTRACT

**We present a systematic study of the long-timescale dynamics of the Drew–Dickerson dodecamer (DDD: d(CGCGAATTGCGC)$_2$) a prototypical B-DNA duplex. Using our newly parameterized PARMBSC1 force field, we describe the conformational landscape of DDD in a variety of ionic environments from minimal salt to 2 M $Na^+Cl^-$ or $K^+Cl^-$. The sensitivity of the simulations to the use of different solvent and ion models is analyzed in detail using multi-microsecond simulations. Finally, an extended (10 μs) simulation is used to characterize slow and infrequent conformational changes in DDD, leading to the identification of previously uncharacterized conformational states of this duplex which can explain biologically relevant conformational transitions. With a total of more than 43 μs of unrestrained molecular dynamics simulation, this study is the most extensive investigation of the dynamics of the most prototypical DNA duplex.**

## INTRODUCTION

The static picture of DNA derived from the early X-Ray studies is now challenged by a myriad of experimental and theoretical studies which show DNA to be a highly flexible entity, undergoing many conformational alterations and even modifications of its covalent structure. Simple inspection of the Protein Data Bank (PDB) illustrates how different sequences adopt different conformations, but also how identical sequences can be found in different conformations to due to the presence of ligands or of changes in the environment (1). Clearly, DNA structure should be explained in terms of conformational ensembles rather than in terms of individual structures.

Recently experimental techniques (2–5) are providing invaluable information on DNA dynamics, however most of what we know about the sequence-dependent flexibility of DNA comes from atomistic molecular dynamics (MD) simulations (6–10). As computer power increases and the reliability of force fields improve, more reliable information is derived from atomistic MD simulations (11). Such simulations have revealed the extent, and the complexity, of DNA movements and their tight coupling to the nature and dynamics of the environment (8,12–14). Unfortunately, MD simulations are extremely dependent on the quality of the force field (11,15–17) and, as simulations become longer, errors induced by force fields accumulate, generating erroneous patterns of flexibility (18–20). Continuous refinement of the force field is therefore required in order to profit from computational improvements and to gain better insight into the structure and dynamics of DNA. With this aim in mind, we have recently developed the PARMBSC1 force field, a new functional with an excellent ability to describe a variety of DNA structures on the microsecond timescale (20).

Here we use PARMBSC1 (20) to make a detailed exploration the dynamics of the best known fragment of DNA: the Drew–Dickerson dodecamer (DDD, (21)). DDD is an ideal model system: (i) it contains a biologically relevant sequence that fits well into the canonical B-form of DNA, (ii) it has been extensively studied experimentally (135 structures with the DDD sequence are available in the PDB, some of them solved at very high resolution) and (iii) it

*To whom correspondence should be addressed. Tel: +34 93 4037156; Fax: +34 93 4037156; Email: modesto.orozco@irbbarcelona.org

has also been widely studied by means of nanosecond-to-microsecond MD simulations (7,22,23). In summary, DDD is the best-known model system of DNA, and its analysis is likely to produce results that can be extrapolated to any canonical B-DNA.

In a first step, we evaluate the impact on DNA of a wide variety of solvent and ion models. In a second step, we analyze the impact that changes in ionic strength can have on the collected conformational samples. Finally, we explore in detail the long-timescale dynamics of DNA by using many multi-microsecond trajectories and one extended (10 μs) single trajectory. Our study reveals that the main conformational characteristics of DNA are quite insensitive to the nature of the models used to describe the solvent and ion environment. Changes due to the nature of the salt ($Na^+Cl^-$ or $K^+Cl^-$), or to the ionic strength, are also quite modest. PARMBSC1 provides very stable conformational samplings, which agree well with experimental information on this duplex, but also highlight unusual anharmonic deformations of DNA that can explain some biologically relevant transitions. Overall, in the framework of fixed-charge all-atom force fields, we present here the broadest and conceivably the most accurate study of the multi-microsecond timescale dynamics of duplex B-DNA to date.

## MATERIALS AND METHODS

### System set-up

Starting geometries for all systems used either Arnott-B DNA canonical values (24) or the high resolution X-Ray structure of DDD with PDB ID: 1JGR (25). The systems were then solvated by adding TIP3P or SPC/E waters to a truncated octahedral box and neutralized by adding 22 $Na^+$ or $K^+$. Both Smith-Dang (S&D; (26)) and Joung-Cheatham (J&C; (27)) ion models were considered and extra salt ($Na^+Cl^-$ or $K^+Cl^-$) was added up to a chosen ionic strength (0.15, 0.5 and 2.0 M added salt). Counterions were initially placed randomly, at a minimum distance of 5 Å from the solute and 3.5 Å from one another. For simulations involving potassium, two extra ion parameterizations were tested: Jensen-Jorgensen (J&J; (28)) and Beglov-Roux (B&R; (29)). All the systems were energy minimized, thermalized and pre-equilibrated using our standard multi-step protocol (22,30) followed by 50 ns of equilibration. All the systems were then simulated (production runs) on the microsecond timescale (see Table 1).

### Simulation details

All systems were simulated in the isothermal-isobaric ensemble ($P$ = 1 atm; $T$ = 298 K) using the Berendsen algorithm (31) to control the temperature and the pressure, with a coupling constant of 5 ps. Although this has been the standard protocol adopted by the ABC consortium (8), and many others, for the simulation of short B-DNA sequences, readers should be aware that the Berendsen thermostat may produce a non-uniform temperature distribution. While this was demonstrated for proteins (32), the compact structure of the DDD dodecamer and the weak coupling of the thermostat (5 ps) are likely to minimize such effects in our simulations. In a previous work, the Nosé–Hoover (33) thermostat was also used in combination with PARMBSC1, giving results without perceptible differences (20) (data not shown). Center of mass motion was removed every 10 ps to limit build up of the translational kinetic energy of the solute. SHAKE (34) was used to keep all bonds involving hydrogen at their equilibrium values, which allowed us to use a 2 fs step for the integration of Newton equations of motion. Long-range electrostatic interactions were accounted for by using the Particle Mesh Ewald method (35) with standard defaults and a real-space cutoff of 10 Å. The PARMBSC1 force field (20) was used to represent DNA interactions. All simulations were carried out using the PMEMD CUDA code module (36) of AMBER 14 (37).

### Analysis

During production runs, data was typically collected every 1 ps, which allowed us to study infrequent, but fast movements. All the trajectories were pre-processed with the CPPTRAJ (38) module of the AMBERTOOLS 15 package (37), the NAFlex server (39) and tools developed in the group (http://mmb.irbbarcelona.org/www/tools). DNA helical parameters and backbone torsion angles associated with the each base pair (bp) and base pair step (bps) were measured with the CURVES+ and CANAL programs (40). The substates of the torsion angles of the backbone (α, γ, ε and ζ) were categorized following the standard definition: *gauche positive* (g+) = 60 ± 40 degrees; *trans* (t) = 180 ± 40 degrees; and *gauche negative* (g−) = 300 ± 40 degrees. For the analysis of the vast majority of helical parameters we took advantage of the palindromic nature of the DDD sequence considering both strands independently, or as an average between the Watson and Crick strands. For comparison with the data available in the experimental databases, which were obtained in different environments, we built a single theoretical conformational space containing almost 40 million structures taken from all the independent trajectories and constituting an aggregated simulation time of 43 μs.

*Experimental structures of the Drew–Dickerson dodecamer.* The experimental conformational space of DDD was defined as a set of experimental structures in the PDB with the sequence: $d(CpGpCpGpApApTpTpCpGpCpG)_2$; see Supplementary Table S1 for a detailed list. The final ensemble contained structures of DDD either isolated or in complex with small organic compounds (Supplementary Table S1). Both ligands and those sequences containing non-canonical covalent modifications were removed. After this selection procedure, the remaining 93 structures were analyzed with CURVES+ (40) and used as a reference conformational ensemble.

*Solution X-ray scattering profiles from MD simulations.* We computed SAXS/WAXS spectra from MD, with PARMBSC1, by taking 1000 structures from the last microsecond of the simulation with 0.15 M $Na^+Cl^-$ in TIP3P water, and generating 100 spectra, each being the average of 10 snapshots. The conditions in that simulation are the closest to the experimental ones, obtained in 0.10 M of added $Na^+Cl^-$ plus 0.05 M of Tris·HCl (41). With an estimated

**Table 1.** Overall simulation information for the systems studied

| System name | Initial structure | Solvent model | Number of waters | Ion type | Ion model | Number of ions | Total time | Sample |
|---|---|---|---|---|---|---|---|---|
| **PARMBSC1** | | | | | | | | |
| J&C Na neutral | fiber | SPC/E | 4968 | Na+ | J&C | 22 | 2 μs | 1 ps |
| S&D Na neutral | fiber | SPC/E | 4968 | Na+ | S&D | 22 | 2 μs | 1 ps |
| S&D Na neutral | fiber | TIP3P | 4998 | Na+ | S&D | 22 | 2/10 μs | 1/20 ps |
| J&C Na neutral | fiber | TIP3P | 4970 | Na+ | J&C | 22 | 2 μs | 1 ps |
| S&D NaCl 0.15M | fiber | TIP3P | 5324 | Na+/Cl− | S&D | 36/14 | 2 μs | 1 ps |
| S&D NaCl 0.5M | fiber | TIP3P | 5118 | Na+/Cl− | S&D | 64/42 | 3 μs | 1 ps |
| S&D NaCl 2.0M | fiber | TIP3P | 5095 | Na+/Cl− | S&D | 162/140 | 2 μs | 1 ps |
| J&C neutral | 1JGR | SPC/E | 5037 | K+ | J&C | 22 | 3 μs | 1 ps |
| S&D neutral | 1JGR | SPC/E | 5187 | K+ | S&D | 22 | 3 μs | 1 ps |
| S&D neutral TIP | 1JGR | TIP3P | 5187 | K+ | S&D | 22 | 2 μs | 1 ps |
| S&D 0.15M | 1JGR | SPC/E | 5159 | K+/Cl− | S&D | 36/14 | 5 μs | 1 ps |
| S&D 0.5M | 1JGR | SPC/E | 5118 | K+/Cl− | S&D | 64/42 | 3 μs | 1 ps |
| S&D 2.0M | 1JGR | SPC/E | 5095 | K+/Cl− | S&D | 162/140 | 2 μs | 1 ps |
| J&J neutral | 1JGR | SPC/E | 8609 | K+ | J&J | 22 | 1 μs | 1 ps |
| B&R neutral | 1JGR | SPC/E | 4993 | K+ | B&R | 22 | 1 μs | 1 ps |
| **PARMBSC0** | | | | | | | | |
| S&D neutral TIP | 1BNA | TIP3P | 4998 | Na+ | S&D | 22 | 4 μs | 1 ps |
| S&D 0.15M | fiber | SPCE | 5044 | K+/Cl− | S&D | 36/14 | 2.4 μs | 10 ps |
| J&C 0.15M | fiber | SPCE | 5046 | K+/Cl− | J&C | 36/14 | 0.6 μs | 10 ps |
| S&D NaCl 0.15M | fiber | SPCE | 5044 | Na+/Cl− | S&D | 36/14 | 0.6 μs | 10 ps |
| J&C NaCl 0.15M | fiber | SPCE | 5049 | Na+/Cl− | J&C | 36/14 | 0.6 μs | 10 ps |

resolution of 2 Å, this experimental WAXS spectrum is, as far as we know, the most accurate available for the DDD sequence (41). To measure the intensities we used the method developed by Park *et al.* (42), which was implemented in AmberTools by Case group. Conceptually, X-ray scattering compares the scattering intensity from the sample of interest, in this case the full solvated DNA, to a 'blank' with just solvent present, and reports the difference, or 'excess' intensity. Consequently, we simulated a water box with 0.15 M of added $Na^+Cl^-$ (50 ns of production run), with the same settings mentioned above, and used it as the 'blank' sample. Only waters and ions within 10 Å distance from the nearest DNA atom were considered to build the spectra, and hydrogen atoms from the DNA were explicitly considered. In addition, we used the recent RISM model (43) to compute the WAXS spectra of the experimental structures 1BNA (X-ray), 1GIP (nuclear magnetic resonance; NMR) and the average structure from the MD. 1GIP is known to be the experimental structure that best matches the experimental solution scattering profile (44). The distribution function of waters and ions computed with RISM also considered a TIP3P solution with 0.15 of added $Na^+Cl^-$.

*Analysis of the cations.* The new CANION module from CURVES+ (45–47) was used to determine the position of each cation in curvilinear helicoidal coordinates for each snapshot of the simulations with respect to the instantaneous helical axis. Given a distance D along the helical axis, ion distributions were computed for each bps (defined here as N-0.2 ≤ D ≤ N+1.2 for a generic bps $N_ipN_{i+1}$) inside the grooves (distance from the axis $R \leq 10.25$ Å), dividing the contribution between the minor groove ($A = 33$–$147°$) and the major groove (polar angle $A = 33°$ to $0°$ to $147°$), as shown in Supplementary Figure S1. We analyzed the ion distribution in one-dimensional (R, D or A) and two-dimensional (RA, DA, DR) curvilinear helicoidal coordinates. Three-dimensional distributions were also reconstructed in Cartesian coordinates using an average structure for the DNA oligomers obtained with CPPTRAJ (38) from the full-length simulations. Ion densities were obtained in units of molarity as detailed elsewhere (45). Special attention was paid to the convergence of the ion population both inside and outside the DNA major and minor grooves for each bps as previously described (47).

*Similarities, global and local flexibility in Cartesian and Helical spaces.* Deformation modes were determined from a principal component analysis of the collected simulations using PCASUITE (http://mmb.pcb.ub.es/software/pcasuite/pcasuite.html). DNA entropy values were obtained from trajectories using the Schlitter (48) and the Andricioaei–Karplus (49) methods for all heavy atoms (excluding terminal base-pairs). Similarity indices were calculated using Hess metrics (50), and energy-weighted similarities (9). Eigenvalues (in $Å^2$) were computed by diagonalizing the covariance matrix and were ordered according to their contribution to the total variance. Self-similarities of the first 10 eigenvalues were computed by comparing the first and second halves of a given trajectory. Relative similarities were computed as described in our previous work (6,51,52). Stiffness constants were determined using base pair step helical stiffness matrices, and base-resolution stiffness matrices, always obtained from the inversion of covariance matrices derived from the atomistic simulations (10). Persistence lengths (in nm) were obtained according to (53), considering all possible DNA sub-fragments. The sequence used to calculate the persistence lengths was artificially extended by taking the inner 8 bp of the DDD sequence and multiplying this segment by 20 to create a 160 bp oligomer: $(CGAATTCG)_{20}$. The calculations were executed on ensembles of $10^4$ structures generated by an in-house implementation of Olson's Monte Carlo procedure (54). As discussed elsewhere (53,54), persistence length is a macroscopic descriptor of the polymeric flexibility of duplex DNA that can be compared with experimental measures.

*Sampling of extreme cases and anharmonic distortions.* Certain fluctuations in the global structure of the DNA cannot be reasonably explained in the harmonic regime. Oth-

ers, while harmonic, represent extreme cases found at the margins of the distribution of sampled conformations. The latter were detected by looking at the distribution of deformation energies, calculated for a reduced set of DNA conformations taken from each MD simulation (one structure every 2 ns), with respect to a reference state defined by the MD-derived basepair step stiffness matrix (6) and the average DNA conformation. We approximate the distribution of deformation energies to a normal distribution, and consider extreme deformations to be those structures with energies above either two or three standard deviations from the average. Anharmonicity was evaluated by applying the Shapiro–Wilk test (55), since none of the reduced sets of deformation energies obtained for each trajectory had more than 5000 values. Furthermore, the complete ensemble of deformation energies (combining all the trajectories) was analyzed graphically by means of Q-Q and box plots, characterizing the structures sampled by the force field beyond the harmonic approximation. In these analyses the terminal base pairs were not considered.

### Statistics, graphics and molecular plots

The statistical analysis, including the Bayesian Information Criterion (BIC) and linear correlations, as well as associated graphics, were obtained with the R 3.0.1 statistical package (56) and the ggplot2 library (57). The molecular plots were generated using either VMD 1.9 (58), or the UCSF Chimera package version 1.8.1 (59).

## RESULTS AND DISCUSSION

### Environmental impacts on DNA

MD trajectories are dependent on the solvent and ion environment in two different ways: one legitimate, since changes in solvent, ionic strength or the nature of the ions should impact simulations, sometimes in a dramatic way (13,60), and one illegitimate, linked to the uncertainties in the force fields used to represent water or ions. Before analyzing the detailed dynamics of DNA, it is therefore necessary to evaluate the uncertainties introduced into simulations by the ion and/or solvent models used. For water, we considered the two most popular three-point models: TIP3P (61) and SPC/E (62), while for salt ($Na^+Cl^-$ or $K^+Cl^-$) we considered models by Joung-Cheatham (J&C), Smith-Dang (S&D), Jensen-Jorgensen (J&J) and Beglov-Roux (B&R) (the last two only for $K^+Cl^-$, see 'Materials and Methods' section). All trajectories involved at least 2–3 μs of simulation and were extended up to 5 μs when ion convergence was not clear.

Consistent with the general article describing the parameterization and validation of the PARMBSC1 force field (20), all the simulations yielded average RMSDs in the range 1.9–2.3 Å with respect to Arnott B-DNA values (1.6–2.0 Å if terminal base pairs were excluded), ∼2.2 Å with respect to the X-Ray structures (PDB IDs: 1BNA, 2BNA, 7BNA and 9BNA) and ∼1.8 Å with respect to the ensemble of NMR structures in the PDB ID: 1NAJ. About 91–98% of the Watson–Crick hydrogen bonds were maintained during the multi-microsecond simulations explored in this work (see Supplementary Table S2). Helical parameters, groove dimensions and torsion angles sampled in the simulations in all cases matched the values expected from experimental structures. A more detailed comparison with a large number of high resolution X-ray and solution NMR structures is carried out below.

As already suggested by previous studies (7), the impact of using two different solvent models (SPC/E and TIP3P) is negligible in terms of the structural properties of DNA. We analyzed in detail the six helical base pair step parameters along the DDD sequence (Figure 1), and the complete set of 16 helical parameters and 8 torsion angles (see Supplementary Table S3 ($Na^+Cl^-$) and S4 ($K^+Cl^-$)). The impact of changing the ion force-field parameters also led to very small differences in terms of the global properties of DNA (see Figure 1 and supplementary Tables S3 and 4), in good agreement with previous PARMBSC0 (19) results (7,63). The use of $Na^+Cl^-$ or $K^+Cl^-$ has also little impact on the DNA structure, again in good agreement with previous simulations (46,63). Finally, increasing the ionic strength (here we tested ∼0.15, ∼0.5 and ∼2 M), also seemed to produce little effect (see Figure 1) on the average structure of DNA (for local effects linked to ionic strength see the discussion below).

As the MD conformational ensembles appeared to be quite robust to changes in the solvent or ionic atmosphere, we combined all the trajectories to create a 43 μs meta-trajectory from which a 'theoretical' conformational space of B-DNA can be defined and compared with that derived from experimental structures (see 'Materials and Methods' section and Supplementary Table S1), which were also solved in different environments (93 structures, which thanks to the palindromic nature of DDD sequence provide 186 experimental estimates of helical parameters for each type of base pair step in the sequence: CpG, GpC, GpA, ApA, ApT and TpT). As shown in Figure 2, no experimental conformation lies outside the 'theoretical' conformational space derived from our simulations. Furthermore, experimentally observed conformations lie in the regions of higher density in the MD-derived sampling (see Figure 2 for a comparison of the rotational space, and Supplementary Figure S2 for the translational space). DDD actually covers quite a wide conformational space, as reflected by the variety of experimental structures, well reproduced by the MD simulations.

Although X-ray and NMR derived structures have been considered for the last 20 years the gold-standard for force field comparison (64), they both suffer from some limitations when it comes to represent the structure of the DNA in solution (65). Crystal packing and crystallization artifacts in X-ray techniques, or user-biased integration of the peaks and refinements based on all-atom force fields in NMR experiments, are just some of the known limitations. To complement this data, Small-Angle X-ray Scattering (SAXS) and Wide-Angle X-ray Scattering (WAXS) are able to deliver information about the shape and size of the molecules in solution. At high resolution, structural polymorphism such as the B-DNA/B'-DNA can be detected (41), although the spatial averaging carried out to derive the profiles also leads to a loss of information in SAXS/WAXS compared to crystallography. To complement our findings, we compared our simulations to the high-resolution (2 Å) WAXS

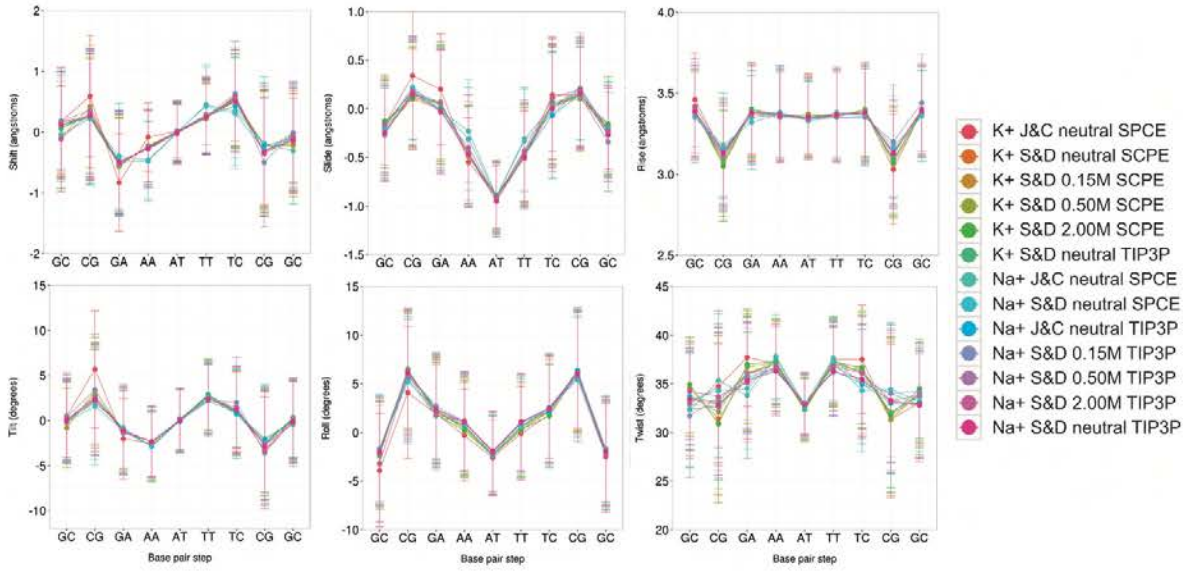



**Figure 1.** Averaged base pair step helical parameters along sequence for all the simulation performed with PARMBSC1. See Table 1 for a detailed description of the simulated systems. Translational parameters (shift, slide and rise) are reported in Å, and rotational ones (tilt, roll and twist) in degrees. The terminal base pairs were removed from the analysis.

spectrum obtained for DDD by Zuo and Tiede (41). Nevertheless, the reader should be aware that comparisons between theoretically-derived and experimental spectra have to be made with caution, due to the problems in generating profiles from structural models (specially related to the different way to treat the solvent; see 'Materials and Methods' section), the different conditions in the simulation and experiments, and the lack of definition in certain regions of the spectra. With these cautions in mind, it seems clear from Figure 3 that PARMBSC1 is able to overcome some of the deviations from experiment described previously using PARMBSC0 (60), and provide spectra that fit well the experimental ones. Quantitative comparison of peak location reveals that PARMBSC1 recovers the first peak (P1) near $q \approx 0.45$ Å$^{-1}$, which was reported to be absent in PARMBSC0 simulation (65). The major deviation from experiment was found at wide angles (P5), where PARMBSC1 is slightly shifted respect to the experimental value, but where the resolution of both theory (43) and experiment is also lower and peak location is not so clear. It is worth noting that in general PARMBSC1 fits the experimental profiles with a quality similar or better than the best experimentally derived structures (by NMR or X-Ray), even in the most complicated region (P1–P3) that reflects the structure of the sugar-phosphate backbone (see Supplementary Table S5).

In summary, PARMBSC1 trajectories reproduce experimental observables accurately and seem robust with respect to the (somewhat arbitrary) selection of ion and solvent force fields. In terms of general DNA structure, the trajectories are also quite insensitive to ionic strength (over a 'physiological' range) and to the nature of the salt ($Na^+Cl^-$ or $K^+Cl^-$). These results suggest that globally, despite the use of simple additive potentials (66,67) PARMBSC1 is performing very well. Further improvements are likely to require the inclusion of new factors such as polarization.

### Similarities, global and local flexibility

Processing of the covariance matrix obtained from atomistic MD simulations provides a direct measure of DNA flexibility in Cartesian space, which can be described in terms of essential deformation movements and quasi-harmonic entropies (see 'Materials and Methods' section and reference (51)). These estimates cannot be directly compared with experimental observables, but are very useful in determining the similarity between deformation patterns and stiffnesses derived from different force fields (6). As shown in Supplementary Figure S3, the dynamics of the central 10 bp of DDD obtained with PARMBSC1 have a very high (>75%) similarity with respect to a reference simulation using PARMBSC0. This similarity increases to more than 90% if Boltzmann indices are considered (Supplementary Figure S3D). We can conclude that the nature and the magnitude of the principal deformations of DNA are very similar with both force fields. This is confirmed by an analysis of the Cartesian entropies and the stiffnesses associated with the main deformation movements (Supplementary Table S6). Helical stiffness analysis provides an alternative picture of DNA dynamics by considering local perturbations of helical parameters (10). Results at the base pair level (Supplementary Figure S4) and at the base pair step level (Supplementary Figure S5) show the expected sequence dependence (6) and confirm the similarity between PARMBSC0 and PARMBSC1 results. Finally, polymeric MD-stiffnesses derived from the extension of DDD to a very long duplex (see 'Materials and Methods' section) demonstrate that PARMBSC1 also reproduces the persis-

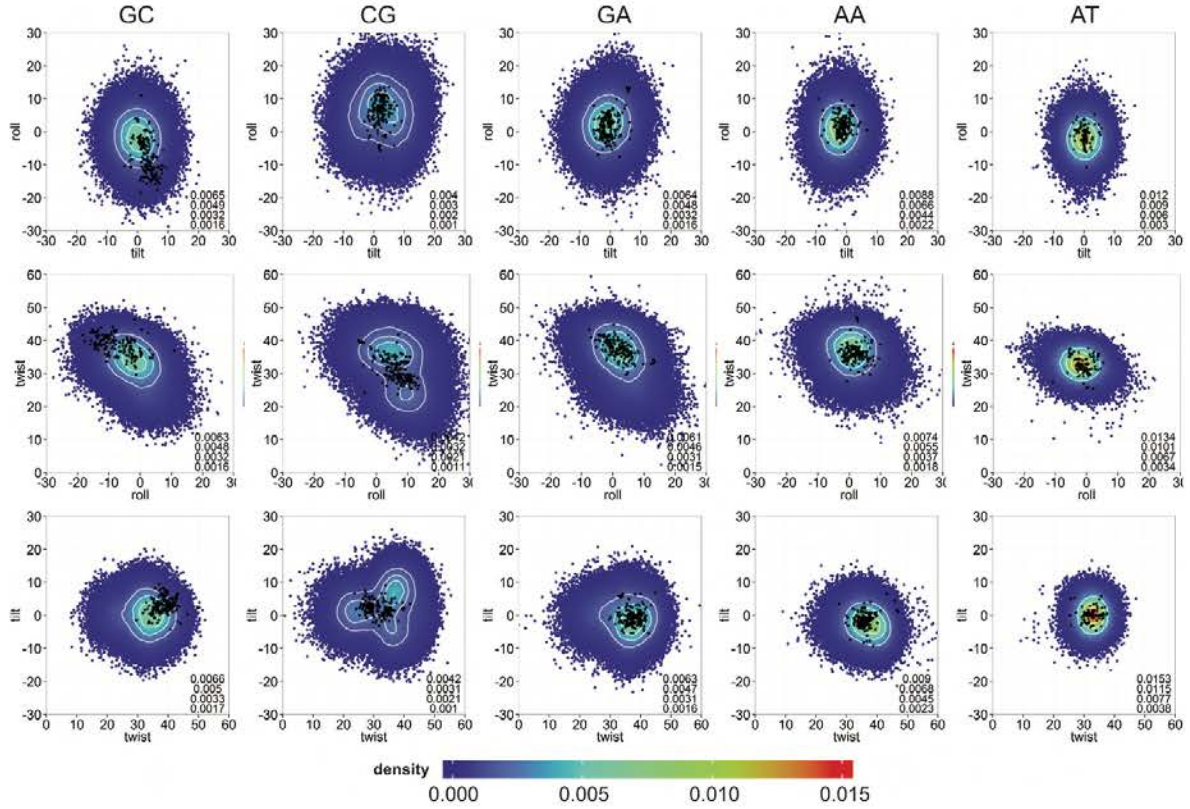



**Figure 2.** Comparison at the bps level between the theoretical and experimental rotational spaces. Rotational parameters (tilt, roll and twist) are reported in degrees. All distinct bps found in DDD are shown (removing the ends): GC (first column), CG (second column), GA (third column), AA (fourth column) and AT (fifth column). Smoothed 2D densities, estimated by fitting the observed distributions to a bivariated normal kernel (evaluated on a square grid of 90 × 90 bins), are depicted by coloring the points coming from the MD simulations with a color gradient from low (blue) to high (red) density. Four iso-density curves are shown in white, and are quantified on the bottom right side of each plot. Experimental conformations are shown as black dots (supplementary Table S1).

tence length of duplex DNA with values ranging from 48 to 57 nm depending on simulation conditions (Supplementary Table S6), compared to experimental estimates of around 50 nm (68).

To further investigate the capacity of MD to sample extreme conformations and also structural distortions beyond the harmonic regime, we fitted the deformation energies (see 'Materials and Methods' section) with a normal distribution obtaining an average ($\alpha$) of 1.8 kcal mol$^{-1}$ with a standard deviation ($\sigma$) of 0.4 for the meta-trajectory (note that we obtained the same $\alpha$ and $\sigma$ for the single 10 $\mu$s trajectory). If the movements of the DNA could indeed be described by the harmonic regime, one would expect that the tails of the distribution beyond $\alpha \pm 2\sigma$ would account for 4.56% of the total probability distribution. Following the same reasoning, the probability that a normal deviate would lie beyond $\alpha \pm 3\sigma$ is at most 0.27%. Counting the number of times these extreme regions are sampled in the 10 $\mu$s long trajectory led to 5.64 and 1.32% beyond the $\alpha \pm 2\sigma$ and $\alpha \pm 3\sigma$ limits respectively. Using the complete meta-trajectory that describes 43 $\mu$s of DDD in different environments we obtained 8.91% ($\alpha \pm 2\sigma$) and 2.41% ($\alpha \pm 3\sigma$), clearly showing that PARMBSC1 simulations significantly sample extreme conformations, more frequently than expected from the harmonic regime. Furthermore, we applied the Shapiro–Wilk test to check whether the sets of distortion energies could be drawn from a normally distributed population. For all the trajectories, we rejected the null hypothesis with $P < 0.05$, supporting the deviation from normality. We also analyzed the complete space of deformation energies (combining the results obtained separately for each trajectory), using a graphical approximation. The distribution, Q-Q plot, and boxplot presented in Supplementary Figure S6, clearly supports the presence of anharmonic distortions related with highly bent structures.

We saw above that solvent and ion force-field models, including both $Na^+Cl^-$ or $K^+Cl^-$, had little impact on the global structure or flexibility of DDD. This may seem reasonable given the 'physiological' conditions range used in this work (see Noy and Golestanian (69)). However, while simulations carried out with minimal (neutralizing) ionic strength seem to lead to shorter persistence lengths (Supplementary Table S6), there is no systematic trend relating flexibility and ionic strength. Note that the use of the AMBER

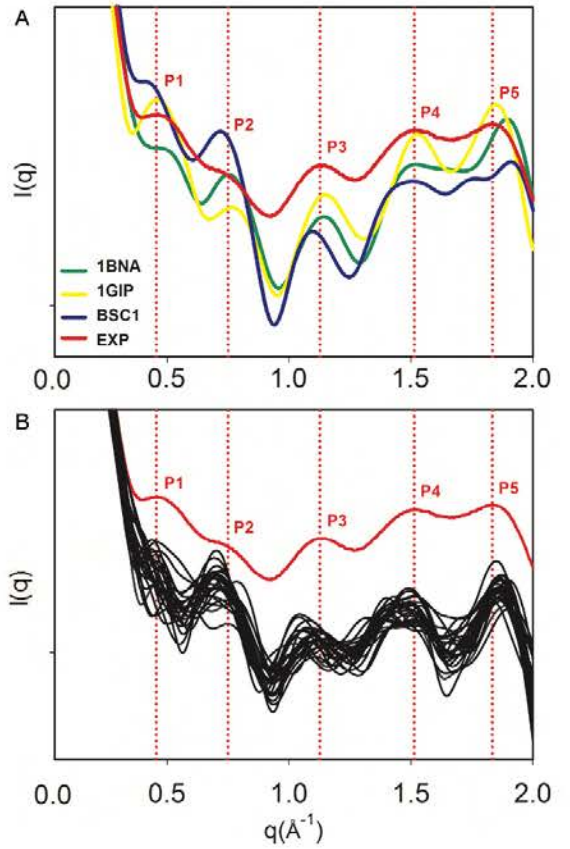



**Figure 3.** Solution scattering profiles. (**A**) Solution interference patterns computed with the RISM approach (43) for the DDD crystal (PDB ID: 1BNA, green), the NMR (PDB ID:1GIP, yellow) and the average structure from the MD simulation (PARMBSC1, blue), compared to the experimental profile (red). Vertical dotted lines in red represent the peaks determined experimentally. (**B**) Scattering profiles obtained from the MD simulation with PARMBSC1 using the method from Park *et al.* (42) (see 'Materials and Methods' section), compared to the experimental result. The positions of the peaks are reported in Supplementary Table S5. Note that the absolute intensities were accordingly shifted to a common origin to maximize the overlap. The data to produce the experimental curve was a courtesy of Prof David Tiede (41).

implementation of PME to treat long-range electrostatics precludes performing simulations at very low ionic strength (net-charged systems), where the connection between global flexibility and ionic strength could become significant. The reason is the implicit presence of a net-neutralizing plasma that appears due to the omission of the zeroeth-order term in the reciprocal Ewald sum (70,71).

### Ion atmosphere

Previous sections have shown that the global structure and dynamics of DNA is not dramatically dependent on the solvent or ion force field, the nature of the monovalent cations ($Na^+$ or $K^+$), or the ionic strength (within the range studied). However, this robustness of DNA to environmental conditions does not preclude local changes linked to the solvent or ionic atmosphere. We investigated this possibility in more detail by looking at the interactions of DNA with ions. The first point that becomes evident when looking at the trajectories is that while DNA structure is reasonably well converged in several hundred ns (46), the ionic environment may require significantly more time to converge, as suggested from earlier simulations (47). This is indeed what the analysis of ion population at the base pair step level (see 'Materials and Methods' section) shows (see Supplementary Figures S7–10 for the analysis of $K^+Cl^-$. Similar profiles were obtained for $Na^+Cl^-$). It is also clear that the convergence of the ion distribution depends on the ion force field (Supplementary Figures S7 and 8) and also on the region of DNA that is analyzed. As an example, ions represented with the J&J model converge quickly (200 ns) inside the grooves and around the DNA, whereas the J&C ion atmosphere is not fully converged in the 3 μs studied here (Supplementary Figures S7 and 8). It is also clear that convergence is in general faster in external regions (around the phosphates, Supplementary Figures S7 and 9), than within the grooves and notably within the narrow minor groove where saw tooth-like curves can be observed (Supplementary Figures S8 and 10). In these cases, convergence is not fully guaranteed even after 5 μs (see the 150 mM S&D simulation in Supplementary Figure S10). It is worth noting that convergence problems do not decrease when ionic strength is increased, despite the fact that more ions are available in the DNA environment, indicating that it is not a simple statistical problem. Indeed, the saw tooth-like population curves of S&D ions inside the minor groove (especially at the central AT step) in the minimal salt simulations are present, and sometimes even amplified, in simulations at higher ionic strength (see Supplementary Figure S10). This suggests that ions visiting some narrow regions in the grooves may be frustrated (47), trapped in an oscillatory regime between two different substates. Ions with long residence times inside the grooves, could also explain part of the oscillatory regime. Thus, using the S&D 0.15 M simulation we compared the volume of the groove, the time evolution of $K^+$ ions visiting the minor groove and the average residence at $A_6T_7$ and $C_3G_4$, which are the two most populated bps at physiological concentration (Supplementary Table S7). The average volume of the minor groove is significantly narrower for AT (193 $Å^3$) than for GC (239 $Å^3$) as previously reported (7,22). We also found that the average residence time of $K^+$ inside the minor groove was 108 ps for AT versus 50 ps for CG (if we consider an ion to be present when it stays at least 20 consecutive ps inside the groove (46)). Indeed, $K^+$ ions are able to remain within the AT step for several hundreds of ns (Supplementary Figure S11). During these long periods there is a higher probability of simultaneously finding two cations inside the narrow groove of AT, compared to CG. Based on the visual inspection of a single extended trajectory, this double occupancy seems to produce an imbalance that triggers the release of both ions from the groove within a few ps. This could explain the oscillatory ion population at the AT step, as it indeed occurs at each of the sawtooth-like peak we observed in the AT time series (Supplementary Figure S11). Nevertheless, a more systematic approach with statistical support should be undertaken

to confirm these findings. Similar events are not seen in the minor groove of CG steps (Supplementary Figure S12).

While remaining cautious with respect to convergence problems, we can reach some general conclusions on the impact of the ion force field on ion populations around DNA. Of the four $K^+$ models tested, the Lorentz-Berthelot (LB) implementation of J&J is the one showing the weakest affinity for DNA (Figure 4 and Supplementary Figure S8), leading to very low ion populations inside the grooves and failing to explain the regions of high cation density found experimentally (25,72,73) in the minor groove of the AATT segment (note that this different behavior could be due to the conversion from geometric to LB combination rules used to build the Lennard-Jones potential in AMBER (37,63), since these parameters were created to work with another van der Waals functional (28)). J&C is the one with the strongest DNA affinity, possibly explaining its severe convergence problems in regions with narrow grooves. Finally, the S&D and B&R models (the two most used in DNA simulations), at a first glance, give similar ion distributions (see Figure 4 and Supplementary Figure S8). A more detailed picture of ion environment can be obtained by looking at 3D density plots (see 'Materials and Methods' section) such as those shown for B&R, J&C and S&D in Figure 5.

Looking at the minimal salt trajectories, differences in ionic atmosphere are especially visible at the edges of the groove, around the phosphate groups (where only the J&C model generates ion density when analyzing the 1.5 M isomolarity surface, Supplementary Figure S13) and in the central minor groove region, where the J&C model tends to concentrate all ion density in the AT and CG steps. In contrast the J&J model predicts a low ion concentration of ions anywhere in the groove, while S&D and B&R models show a more homogeneous distribution of ions along the groove (Figure 4, see Supplementary Tables S7–10). Looking at high isomolarity surfaces (5 M) shows that not all the parameterizations are able to reproduce the location of the $K^+$ ions that co-crystallized with the DNA in the high-resolution X-ray experiment (25). It could be argued that those co-crystallized cations reached their final location in the crystal cell due to packing or crystallization effects. Nevertheless, the cations are found buried inside the grooves of the DNA in close interaction with the bases. The analysis of their precise location as been the subject of several studies, where the position of cations is correlated with changes in the groove widths, being part of a complex structured network involving cations, water and DNA interactions (22,73). While J&C and S&D models correctly predict the position of $K^+$ ions in both grooves (S&D being the more accurate), B&R only reproduces the cations found in the major groove, while a systematic shift of the density clouds is observed in the minor groove (Figure 5). When the ionic strength is increased (only studied for the S&D model), the general 3D distribution of $K^+$ does not change dramatically (see Supplementary Figure S14), except for the overall increase in ion density that is particularly visible in the major groove and at the groove edges close to the phosphate groups where new sites are populated. The general good agreement between the densities coming from the free dynamics in solution and the co-crystallized cations 'fixed' in the crystal cell, and the high concentration at which the cations were found inside the grooves of some specific bps (up to two orders of magnitude higher respect to the physiological background), make us think that these cations reached their final position in the crystal following a clear sequence-dependent pattern.

As discussed below, differences in cation population or density do not lead to significant structural or dynamic differences in DNA. However, a detailed analysis does show local changes linked to ion populations in the grooves. As an example, the J&J model which showed the weaker affinity for DNA leads to wider grooves (very visible in the AATT segment, see Supplementary Figure S14), while the J&C model, with the strongest DNA affinity, leads to narrower minor grooves in the central AATT segment. A clear correlation is also observed between increasing ion concentration and the width of both DNA grooves (Supplementary Figure S16). With the S&D model, the average minor groove width changes from 4.06 Å (neutral system) to 3.88 Å (2.0 M system; standard error of $2 \times 10^{-5}$ Å). The absolute difference between these two extreme conditions is small in absolute terms, but is enough to add extra structural frustration to K+ ions entering the minor groove. In general, an increased population of ions attached to DNA leads to an increase in the local stiffness associated to the central AATT tract, but the differences are evident only when the 'extreme' J&J and J&C models are compared (Supplementary Figure S15). In contrast to groove geometry, no noticeable changes were found in BI/BII populations. Overall, it is clear that B-DNA is more affected by the choice of ion force fields than by the bulk ion concentration (Supplementary Figures S15 and 16), but these effects remain relatively mild.

### Long-timescale dynamics of DNA

Results above demonstrate that, in general terms, the trajectories obtained from MD simulations are robust with respect to choices of water or ion force fields, the ionic strength or the nature of the salt. We next decided to extend one of our trajectories (S&D, minimal salt (22,46)) to 10 μs to explore slow conformational changes that might be not visible in shorter simulations. The entire 10 μs DDD trajectory samples canonical B-DNA conformations which are very close to both X-Ray (21,74–76) and NMR (77) structures (Figure 6). The average helical parameters (twist, roll, tilt, shift, slide, rise) derived from the MD simulation (including terminal bases) is (35.2, 2.8, 0.2, 0.0, −0.2, 3.3), which compares extremely well with the results derived from NMR ensembles (PDB ID: 1NAJ (35.9, 2.3, 0.0, 0.0, −0.2, 3.2)) or X-Ray crystallography (PDB IDs: 1BNA, 2BNA, 7BNA and 9BNA: (35.0, −0.3, −0.2, −0.1, −0.2, 3.3), confirming the ability of PARMBSC1 to reproduce the overall conformational properties of DNA (see Table 2 for more details) in simulations that are at least an order of magnitude longer than today's 'state-of-the-art'.

This long trajectory is also able to capture some subtle details, such as the lower $\chi$ values sampled by guanosines compared with the other nucleotides, sugar phase angle distributions sampling the correct South–South East regions (Supplementary Table S11), the sequence variability of helical twist (7,51) and BI/BII populations (Supplementary

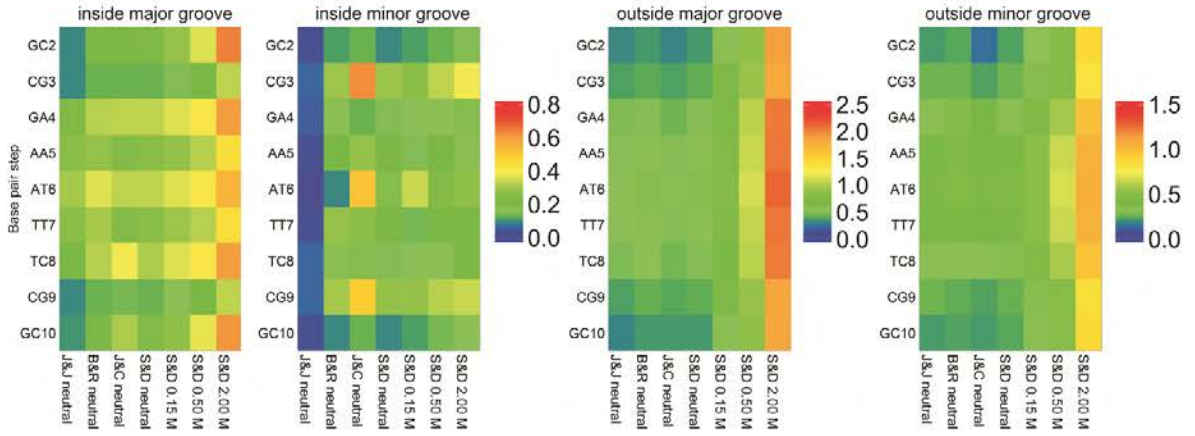



**Figure 4.** Average $K^+$ populations inside and outside the DNA grooves. Populations inside the major and the minor groove (left two panels), and outside both grooves (right two panels). The populations were measured for each bps removing the terminal ones (see 'Materials and Methods' section). See Supplementary Figure S1 for the CHC partitioning scheme used to divide the grooves.



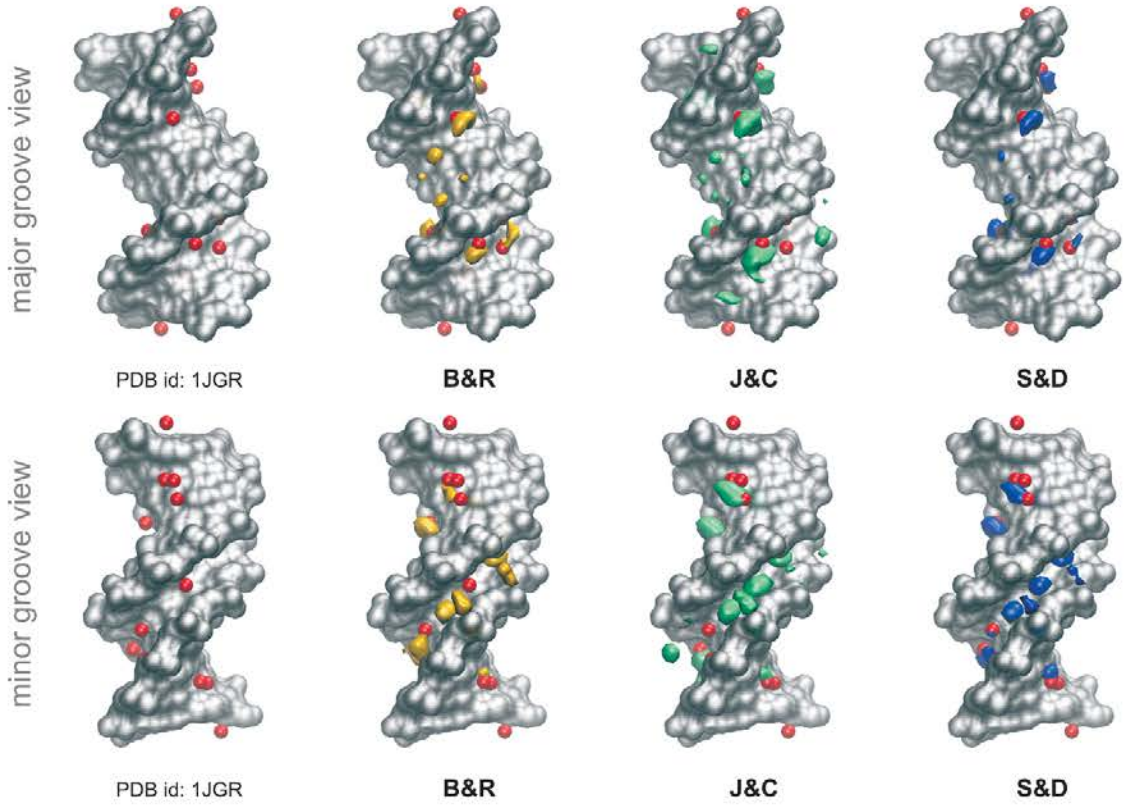



**Figure 5.** Potassium distributions along the helix. Cartesian $K^+$ isomolarity surfaces at 5.0 M reconstructed from the CHC histograms with respect to the average structure (shown as a silver surface). For comparison purposes, neutral systems have been overlapped with the $Tl^+$ cations (red spheres) that co-crystallized with the DNA (PDB ID: 1JGR). Note that thallium cations are used as a replacement of potassium in diffraction experiments (1). The distribution with the J&J parameters are not shown since any visible density was observed at 5.0 M concentration.

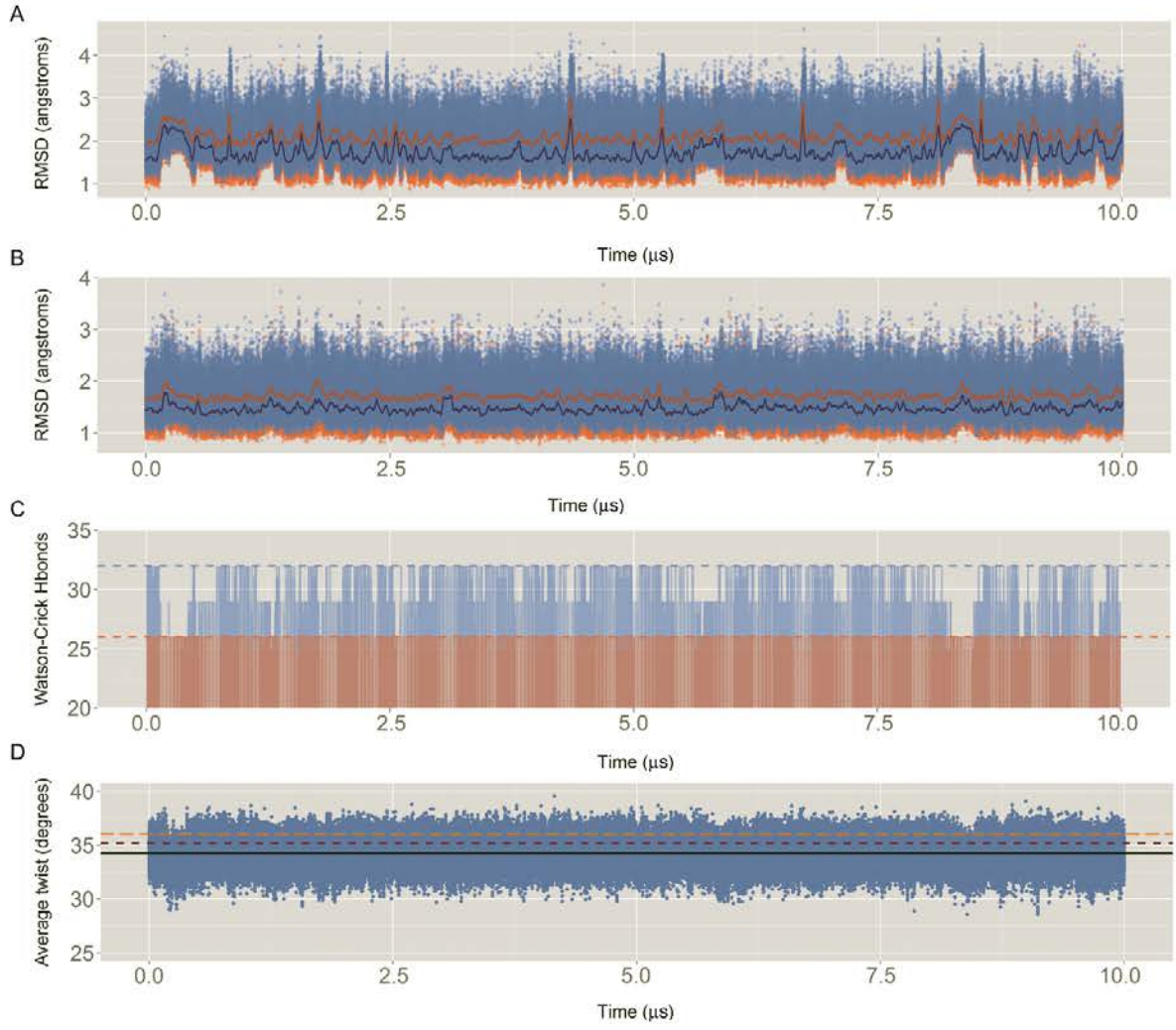



**Figure 6.** Descriptors of the quality of the simulation: (**A**) Root mean square deviation (RMSD) of all the heavy atoms of the Drew–Dickerson dodecamer (DDD) respect to the average experimental value. In blue, RMSD against an average of the X-ray structures (PDB IDs: 1BNA, 2BNA, 7BNA and 9BNA); In orange, RMSD against an average of the NMR ensemble with 5 structures (PDB ID: 1NAJ). For the sake of clarity, running averages every 20 ns are shown in dark orange and dark blue, for X-ray and NMR respectively. (**B**) Same than (**A**) but without considering the capping base pairs (i.e. removing all the heavy atoms of base pairs C1:G34 and G12:C13). (**C**) Evolution of the total number of Watson–Crick hydrogen bonds (Hbonds) with time. Considering a perfect interaction between the 12 bp would lead to a total of 32 Hbonds (light blue dashed line). Without considering the capping base pairs (light red), the ideal total number of Hbonds is 26 (light red dashed line). Hbonds were considered formed if the distance between the donor–acceptor atoms was $\leq$3.5 Å. (**D**) Sequence averaged twist for all the base pair steps (bps), excluding the terminals, with time. The average MD value is shown with a black line, while the experimental references are shown in dark red and orange dashed lines, for X-ray and NMR respectively.

Figure S17). Backbone torsions follow the expected behavior for a duplex B-DNA, preferentially exploring the canonical B-DNA substate characterize by $\alpha$ in *gauche*− (g−) and $\gamma$ in *gauche*+ (g+) (Supplementary Figure S18 and Supplementary Table S12), with $\epsilon$ in *trans* (t) and $\zeta$ in g− (i.e. the BI state).

Very interestingly, despite the canonical $\alpha\gamma$ state being the most populated (in agreement with previous PARMBSC0 simulations (10,22)), all the non-canonical conformations found in experimental protein–DNA complexes (78) are also detected here (Figure 7), thus improving on PARMBSC0 behavior. As shown in Supplementary Figure S18, $\alpha\gamma$ transitions are common (on average 460 transitions per μs per nucleotide) and fast (average residence times being around 3 ps); although only 0.01% of the non-canonical $\alpha\gamma$ states have non-negligible survival times of up to 1.2 ns (averaging across the four nucleotide types). Long-lived $\gamma$-flips of around 100 ns from the g+ to the non-canonical t state where not observed. Note that these flips were recently suggested to be a source of convergence prob-

**Table 2.** Sequence-averaged conformational parameters obtained from the 10 μs Drew–Dickerson dodecamer simulation[a]

| Parameter | Average | SD | Range | Minimum | Maximum | NMR[b] | Xray[c] |
|---|---|---|---|---|---|---|---|
| Shear | 0.00 | 0.30 | 6.43 | −3.62 | 2.81 | 0.00 | 0.03 |
| Stretch | 0.02 | 0.12 | 3.27 | −0.78 | 2.48 | −0.29 | 0.19 |
| Stagger | 0.10 | 0.38 | 4.79 | −2.15 | 2.66 | 0.02 | 0.21 |
| Buckle | 0.0 | 9.7 | 92.9 | −48.3 | 48.0 | 0.0 | −0.5 |
| Propeller | −9.2 | 8.4 | 81.6 | −49.6 | 39.2 | −17.4 | −14.4 |
| Opening | 1.3 | 4.0 | 71.1 | −29.4 | 56.8 | 1.1 | 1.6 |
| Xdisp | −0.58 | 1.05 | 10.42 | −6.05 | 4.36 | −0.01 | −0.15 |
| Ydisp | 0.00 | 0.84 | 9.16 | −4.59 | 4.58 | 0.02 | 0.52 |
| Inclination | 2.2 | 5.7 | 56.0 | −26.1 | 31.0 | 1.7 | −0.6 |
| Tip | 0.1 | 6.9 | 58.6 | −41.7 | 41.2 | 0.0 | −2.6 |
| Shift | −0.01 | 0.81 | 7.31 | −3.71 | 3.63 | 0.00 | −0.07 |
| Slide | −0.24 | 0.53 | 5.50 | −3.17 | 2.34 | −0.22 | 0.14 |
| Rise | 3.32 | 0.29 | 3.32 | 1.96 | 5.30 | 3.20 | 3.35 |
| Tilt | −0.1 | 4.7 | 44.6 | −22.6 | 23.0 | 0.0 | −0.4 |
| Roll | 1.5 | 5.5 | 60.1 | −31.8 | 31.7 | 2.3 | −0.7 |
| Twist | 34.3 | 5.5 | 52.3 | 3.4 | 57.1 | 36.0 | 35.2 |
| α[d] | −72.6 | 18.2 | 314.4 | | | −60.2 | −57.5 |
| β | 166.8 | 21.7 | 251.1 | | | 171.0 | 166.4 |
| γ | 55.6 | 23.2 | 243.6 | | | 49.6 | 48.3 |
| δ | 135.2 | 14.5 | 119.7 | | | 125.7 | 126.3 |
| ϵ | −159.9 | 23.5 | 169.1 | | | −170.8 | −164.3 |
| ζ | −111.4 | 36.1 | 203.1 | | | −103.5 | −112.1 |
| χ | −111.7 | 16.2 | 138.6 | | | −111.5 | −113.5 |
| Phase | 152.0 | 27.2 | 267.3 | | | 135.0 | 135.7 |
| Amplitude | 41.3 | 6.5 | 61.4 | | | 34.0 | 41.1 |

[a]Capping base pairs were removed from the analysis.
[b]Computed from the ensemble of structures (PDB ID: 1NAJ).
[c]Computed from the X-ray structures with PDB IDs: 1BNA, 2BNA, 7BNA and 9BNA.
[d]For the dihedral angles only the Watson strand was considered.

lems when using PARMBSC0 (10). As expected (7,22), C and G nucleotides show longer-lived and more frequent αγ transitions than A or T (Supplementary Figures S18). However, the occurrence of non-canonical αγ states is not cooperative and does not lead to the destructuring of the double helix that was found with older force fields (18). On average, at any given moment, less than one (0.86) of the 24 nucleotides is in an unusual αγ state. Extrapolating to polymeric DNA implies that 3.6% of nucleotides will exhibit an unusual αγ conformation. This could be a factor favoring recognition by specific proteins, given that crystal structures of protein-DNA complexes show a significant percentage of such states (10%, (67)).

As expected from previous experimental and theoretical studies (7,8,22,79,80), C·G base pairs show a significant propensity for BI/BII transitions (Figure 7), as evidenced by the concerted changes in ϵ and ζ from t to g− and from g− to t respectively (Supplementary Figure S19). BI/BII relative populations are notably improved with respect to PARMBSC0, and now reproduce more accurately NMR experiments (Table 3 and Supplementary Figure S17). As previously suggested (8,46,81), BI/BII polymorphism is directly connected to two sources of structural polymorphism, namely, the twist bimodality found for CG steps and the slide polymorphism observed for RpR steps. Concerted movements of the backbone and the bases allow the formation of an intra-molecular hydrogen bond of the type C8H8-O3′ between RpR steps, that was proved to be casually connected to BII populations (8,46). These concerted movements seems to be linked to twist polymorphism, the low twist state being driven by the presence of cations specifically binding in the minor groove of CG steps (46). Indeed, a systematic increase in the low twist state is observed upon increasing the amount of added K+Cl− from 0.15 to 2.0M (Supplementary Table S13). Note also that the weighted-average twist obtained with the BIC method (82) is higher with PARMBSC1 as a consequence of the higher population of the high twist state (0.66 versus 0.52 in PARMBSC0; Supplementary Table S13). PARMBSC1 is able to correctly reproduce this complex choreography: twist is correlated with ζ, with BI/BII, with the formation of the CH-O interaction and with the slide polymorphism in the neighboring step (Supplementary Figure S20). This is expected to be particularly important in understanding indirect recognition of DNA by proteins (83,84,85), and also the mechanism of DNA intercalation by small organic compounds (46,86).

We have also used our long (10 μs) DDD simulation to investigate base opening events. It is experimentally known that base opening (understood as conformational states where bases are unpaired for significant periods of time) happens on the millisecond timescale (87), at least for coding nucleobases (88), and thus far beyond our simulation window. Terminal base pairs, where stacking interactions are weaker should show slightly higher opening frequencies, but available experimental data do not support the presence of long-lived open states for terminal base pairs (20) and strongly warn against long-lived non-canonical conformations of terminal base pairs stabilized by interactions with the DNA grooves, as found in previous simulations with other force fields (20,89). The PARMBSC1 10 μs trajectory shows that, as expected, the terminal C·G pairs are more labile that the central ones and it is not rare to lose some of the hydrogen bonds for short periods of time (see Figure 6). Fraying events are common, and unusual arrangements of the terminal bases are visited, but they quickly revert to canonical Watson–Crick pairing (see Figure 8) as experimentally expected (20). The tWC/SE state (where cytosine

**Table 3.** Percentage of BI substates obtained with PARMBSC1 compared with PARMBSC0 and NMR experiments[a]

| Bps | PARMBSC0 | PARMBSC1 | NMR1 | NMR2 |
|---|---|---|---|---|
| GC | 48 | 73 | 75 | 53 |
| CG | 87 | 78 | 75 | 66 |
| GA | 53 | 45 | 53 | 44 |
| AA | 91 | 82 | 75 | 63 |
| AT | 99 | 98 | 100 | 78 |
| TT | 100 | 99 | 99 | 93 |
| TC | 98 | 93 | 89 | 92 |
| CG | 79 | 62 | 75 | 77 |
| GC | 71 | 61 | 79 | 35 |

[a]BI percentage for the bps in the DD dodecamer, obtained by averaging the difference between ϵ and ζ angles at the 3′-junction of the Watson strand of each base pair. NMR1 and NMR2 are values obtained using phosphorus chemical shifts as detailed in the works of Schwieters *et al.* (NMR1 (80)) and Tian *et al.* (NMR2, (79)), respectively.

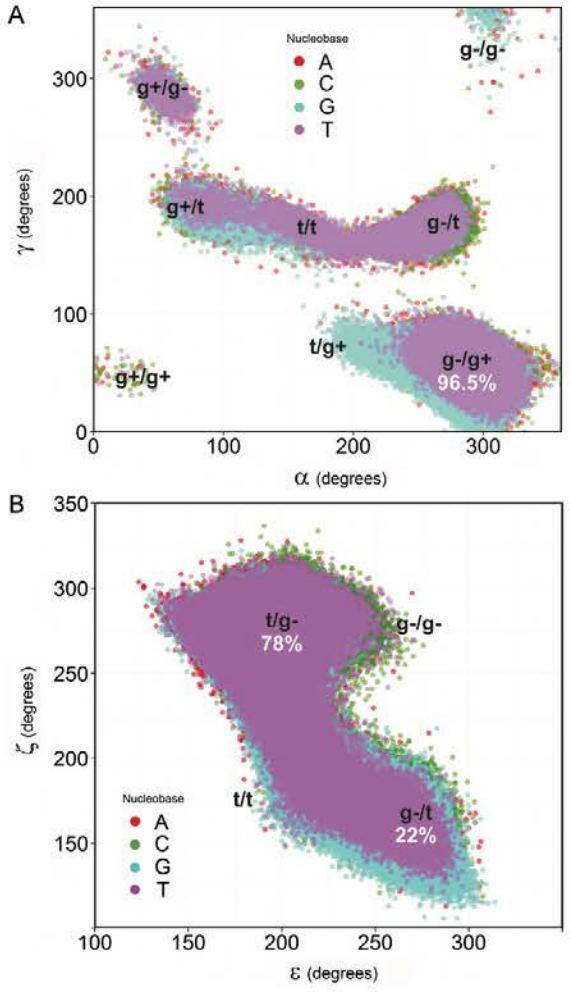



**Figure 7.** Major substates of the backbone observed during the simulation of DDD. (**A**) Scatter plot of α and γ angles grouped by nucleobase. To obtain the distribution of A in the αγ-plane the dynamics of the nucleobases A5 and A6 were considered together (similarly: C3/C9, G4/G10 and T7/T8 were used to build the C, G and T ensembles respectively). (**B**) Same as (A), but for ϵ and ζ angles. The global percentages (considering both strands) of the major canonical states are shown in white.

is turned around the glycosidic bond from the *anti* to *syn* conformation to form a non-WC pair resembling the *trans* Watson–Crick/Sugar Edge C·G pair well-known in RNA structures (90)), now occurs very rarely (probability $<10^{-5}$ and with ps lifetimes), in line with NMR experiments, but in contrast to earlier simulations (7,89). The time evolution of the base pair opening parameter shows that most of the fraying is due to transient sampling of large opening angles (Figure 8). We observe very few events where the glycosidic torsion (χ) goes from *anti* to *syn* in the terminal cytosine (one such event is highlighted by the blue circle in Figure 8). These rare and reversible events are connected with the formation of aO2···5′OH intra-cytosine hydrogen bond, which in turn stabilizes the anomalous tWC/SE conformation. Our simulations suggest that rare and short-lived tWC/SE conformations do not affect neighboring base steps and have no impact on DNA structure and dynamics on multi-microsecond timescales. In addition, no through-the-groove interactions between terminal and inner bases are observed in our long trajectory.

Finally, the 10 μs trajectory allowed us to analyze convergence issues in unprecedented detail; significantly extending previous studies (10,91,92). For this reason, we performed principle component analysis on segments of 1, 2 and 5 μs extracted from the 10 μs trajectory. Visual inspection of Supplementary Figure S21 (supplementary material) shows slight divergence between 1 μs segments, but no significant differences are observed between the 2 and 5 μs segments. The smoothed histograms of the main principal components clearly overlap, suggesting that DNA is sampling the same conformational space. Similarly, differences in entropy values for segments of 2 μs are smaller than 2% with respect to the entropy of the whole 10 μs trajectory, independently of the method used to calculate the entropy (see Supplementary Figure S22 in the supplementary material). This analysis suggest that PARMBSC1 is able to sample in 2 μs, what the user can expect to see in terms of conformational ensembles in 10 μs (91), and for most purposes (ion distribution being an exception) 2 μs trajectories can be considered to be converged.

## CONCLUSIONS

The availability of PARMBSC1, a new and accurate force field for DNA, has allowed us to explore the long timescale dynamics of the DDD, a prototypical B-DNA duplex, in aqueous solution. An unprecedented degree of sampling (involving more than 43 μs of simulation, including a sin-

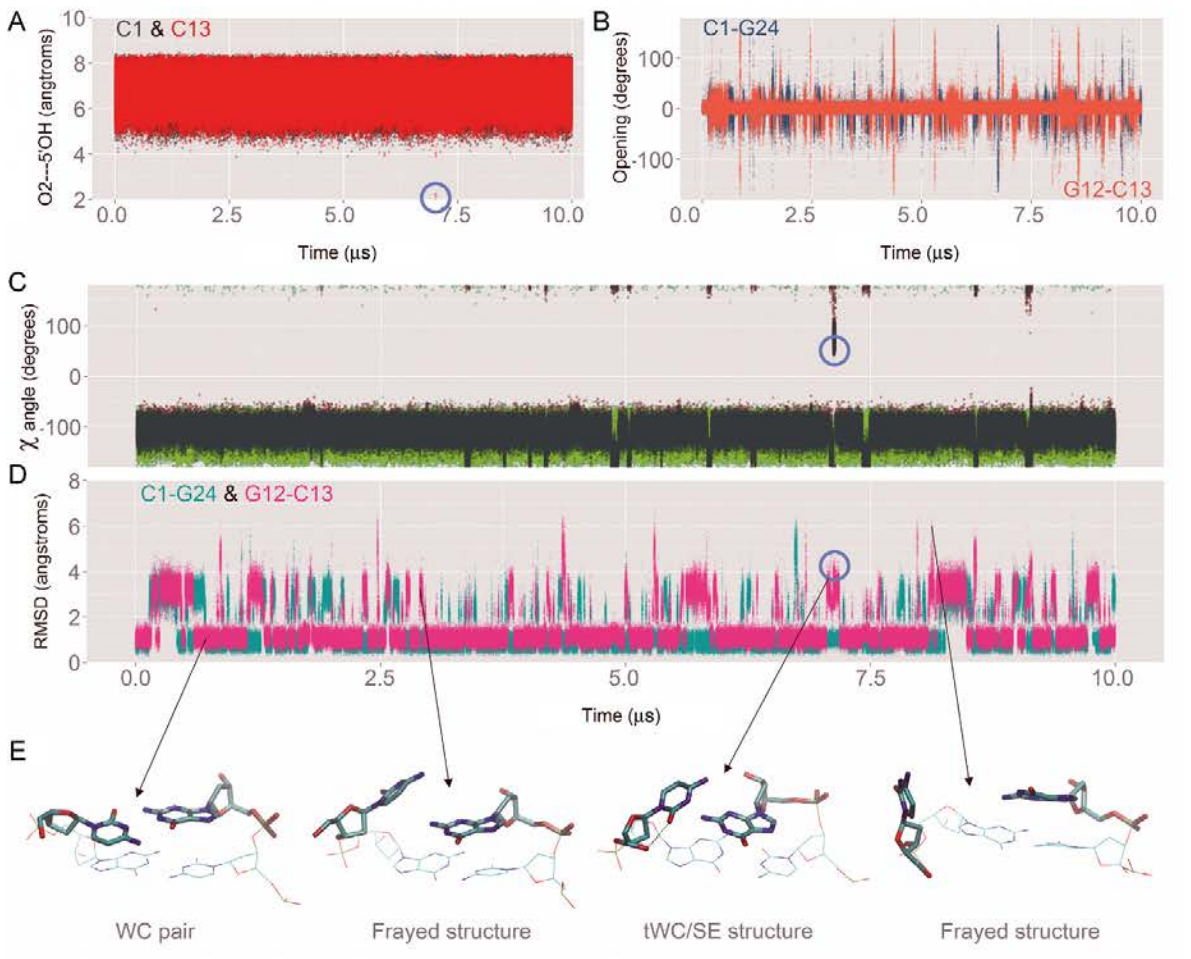



**Figure 8.** Analysis of the base-pair fraying at the ends of the dodecamer. Note that we used the same metrics described elsewhere (78) to analyze the fraying of DDD simulated with previous force fields. (**A**) Formation of an anomalous intra-cytosine hydrogen bond observed with PARMBSC0 and PARMBSC0-χOL4 (one formation event is highlighted by a blue circle). (**B**) Time evolution of the opening parameter. (**C**) χ angle during simulation is shown in red, light blue, dark green and light green for C1, G12, C13 and G24 respectively. (**D**) Mass-weighted RMSD of the capping base pairs respect to the first structure of the simulation. (**E**) Representative structures of the four mayor conformations found are depicted bellow. A similar behavior was observed in the other simulations performed changing the environment conditions (data not shown).

gle 10 μs trajectory) has allowed us to test the impact of different solvent and ion models, and of different salts, on DNA and to characterize the ion atmosphere around the double helix. We have also analyzed the detailed interplay between ions and local and global conformational changes, the prevalence of non-harmonic distortions and we have obtained reliable estimates for conformational jumps that can be important in explaining the specific binding of proteins or ligands to DNA. Lastly, the simulations of the DDD with PARMBSC1 are shown to accurately reproduce experimental data and to represent a clear improvement over the results obtained with earlier force fields.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.


## FUNDING

MINECO Severo Ochoa Award of Excellence (Government of Spain) (to IRB Barcelona); Spanish Ministry of Science [BIO2012-32868, 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.); Czech Republic Grant Agency [14-21893S to T.D., F.L.]; ANR grant CHROME [ANR-12-BSV5-0017-01 to R.L.]. Funding for open access charge: European Research Council (ERC SimDNA).

*Conflict of interest statement.* None declared.

## REFERENCES


1. Hud,N.V. (2008) Nucleic Acid-Metal Ion Interactions. *Biomolecular Sciences*. The Royal Society of Chemistry, Cambridge.
2. Geggier,S. and Vologodskii,A. (2010) Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15421–15426.
3. Huguet,J.M., Bizarro,C. V, Forns,N., Smith,S.B., Bustamante,C. and Ritort,F. (2010) Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15431–15436.
4. Heng,J.B., Aksimentiev,A., Ho,C., Marks,P., Grinkova,Y.V., Sligar,S., Schulten,K. and Timp,G. (2006) The electromechanics of DNA in a synthetic nanopore. *Biophys. J.*, **90**, 1098–1106.
5. Strick,T.R., Dessinges,M.-N., Charvin,G., Dekker,N.H., Allemand,J.-F., Bensimon,D. and Croquette,V. (2003) Stretching of macromolecules and proteins. *Reports Prog. Phys.*, **66**, 1–45.
6. Pérez,A., Lankas,F., Luque,F.J. and Orozco,M. (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.*, **36**, 2379–2394.
7. Dršata,T., Pérez,A., Orozco,M., Morozov,A. V, Sponer,J. and Lankaš,F. (2013) Structure, stiffness and substates of the dickerson-drew dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
8. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
9. Lankas,F., Sponer,J., Hobza,P. and Langowski,J. (2000) Sequence-dependent elastic properties of DNA. *J. Mol. Biol.*, **299**, 695–709.
10. Dršata,T. and Lankaš,F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102.
11. Pérez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.*, **45**, 196–205.
12. Arcella,A., Dreyer,J., Ippoliti,E., Ivani,I., Portella,G., Gabelica,V., Carloni,P. and Orozco,M. (2015) Structure and dynamics of oligonucleotides in the gas phase. *Angew. Chem. Int. Ed. Engl.*, **54**, 467–471.
13. Portella,G., Germann,M.W., Hud,N.V. and Orozco,M. (2014) MD and NMR analyses of choline and TMA binding to duplex DNA: on the origins of aberrant sequence-dependent stability by alkyl cations in aqueous and water-free solvents. *J. Am. Chem. Soc.*, **136**, 3075–3086.
14. Arcella,A., Portella,G., Collepardo-Guevara,R., Chakraborty,D., Wales,D.J. and Orozco,M. (2014) Structure and properties of DNA in apolar solvents. *J. Phys. Chem. B*, **118**, 8540–8548.
15. Dans,P.D., Walther,J., Gómez,H. and Orozco,M. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45.
16. Sponer,J., Cang,X. and Cheatham,T.E. (2012) Molecular dynamics simulations of G-DNA and perspectives on the simulation of nucleic acid structures. *Methods*, **57**, 25–39.
17. Sim,A.Y.L., Minary,P. and Levitt,M. (2012) Modeling nucleic acids. *Curr. Opin. Struct. Biol.*, **22**, 273–278.
18. Várnai,P. and Zakrzewska,K. (2004) DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.*, **32**, 4269–4280.
19. 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 alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
20. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2015) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
21. Drew,H.R., Wing,R.M., Takano,T., Broka,C., Tanaka,S., Itakura,K. and Dickerson,R.E. (1981) Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 2179–2183.
22. Pérez,A., Luque,F.J. and Orozco,M. (2007) Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.*, **129**, 14739–14745.
23. Trieb,M., Rauch,C., Wellenzohn,B., Wibowo,F., Loerting,T. and Liedl,K.R. (2004) Dynamics of DNA: B I and B II Phosphate Backbone Transitions. *J. Phys. Chem. B*, **108**, 2470–2476.
24. Arnott,S. and Hukins,D.W. (1972) Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.*, **47**, 1504–1509.
25. Howerton,S., Sines,C., VanDerveer,D. and Williams,L.D. (2001) Locating monovalent cations in the grooves of B-DNA. *Biochemistry*, **40**, 10023–10031.
26. Smith,D.E. and Dang,L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
27. Joung,I.S. and Cheatham,T.E. (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, **112**, 9020–9041.
28. Jensen,K.P. and Jorgensen,W.L. (2006) Halide, ammonium, and alkali metal ion parameters for modeling aqueous solutions. *J. Chem. Theory Comput.*, **2**, 1499–1509.
29. Beglov,D. and Roux,B. (1994) Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. *J. Chem. Phys.*, **100**, 9050–9063.
30. Shields,G.C., Laughton,C.A. and Orozco,M. (1997) Molecular dynamics simulations of the d(T·A·T) triple helix. *J. Am. Chem. Soc.*, **119**, 7463–7469.
31. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. **81**, 3684–3690.
32. Mor,A., Ziv,G. and Levy,Y. (2008) Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats. *J. Comput. Chem.*, **29**, 1992–1998.
33. Nosé,S. and Klein,M.L. (2006) Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, **50**, 1055–1076.
34. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J.. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
35. Darden,T., York,D. and Pedersen,L. (1993) Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
36. Salomon-Ferrer,R., Götz,A.W., Poole,D., Le Grand,S. and Walker,R.C. (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.
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. Roe,D.R. and Cheatham,T.E. (2013) PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, **9**, 3084–3095.
39. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47–W55.
40. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
41. Zuo,X., Cui,G., Merz,K.M., Zhang,L., Lewis,F.D. and Tiede,D.M. (2006) X-ray diffraction "fingerprinting" of DNA structure in solution for quantitative evaluation of molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 3534–3539.
42. Park,S., Bardhan,J.P., Roux,B. and Makowski,L. (2009) Simulated x-ray scattering of protein solutions using explicit-solvent models. *J. Chem. Phys.*, **130**, 134114.
43. Nguyen,H.T., Pabit,S.A., Meisburger,S.P., Pollack,L. and Case,D.A. (2014) Accurate small and wide angle x-ray scattering profiles from atomic models of proteins and nucleic acids. *J. Chem. Phys.*, **141**, 22D508.
44. Zuo,X. and Tiede,D.M. (2005) Resolving conflicting crystallographic and NMR models for solution-state DNA with solution X-ray diffraction. *J. Am. Chem. Soc.*, **127**, 16–17.
45. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–8149.
46. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
47. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–2423.

48. Schlitter,J. (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.*, **215**, 617–621.
49. Andricioaei,I. and Karplus,M. (2001) On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.*, **115**, 6289–6292.
50. Hess,B. (2000) Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E. Stat. Phys. Plasmas. Fluids. Relat. Interdiscip. Topics*, **62**, 8438–8448.
51. Pérez,A., Blas,J.R., Rueda,M., López-Bes,J.M., de la Cruz,X. and Orozco,M. (2005) Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.*, **1**, 790–800.
52. Orozco,M., Pérez,A., Noy,A. and Luque,F.J. (2003) Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.*, **32**, 350–364.
53. Noy,A. and Golestanian,R. (2012) Length Scale Dependence of DNA Mechanical Properties. *Phys. Rev. Lett.*, **109**, 228101.
54. Zheng,G., Czapla,L., Srinivasan,A.R. and Olson,W.K. (2010) How stiff is DNA? *Phys. Chem. Chem. Phys.*, **12**, 1399–1406.
55. Mecklin,C.J. (2007) Shapiro-Wilk Test for Normality. In: Salkind,NJ and Rasmussen,K (eds). *Encyclopedia of Measurement and Statistics*. SAGE Publications, Inc., Thousand Oaks, pp. 884–887.
56. R Core Team. (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.
57. Wickham,H. (2009) *ggplot2*. Springer, NY.
58. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
59. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
60. Savelyev,A. and MacKerell,A.D. (2015) Differential impact of the monovalent ions Li(+), Na(+), K(+), and Rb(+) on DNA conformational properties. *J. Phys. Chem. Lett.*, **6**, 212–216.
61. 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.
62. 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.
63. Noy,A., Soteras,I., Luque,F.J. and Orozco,M. (2009) The impact of monovalent ion force field model in nucleic acids simulations. *Phys. Chem. Chem. Phys.*, **11**, 10596–10607.
64. Pérez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.*, **45**, 196–205.
65. Savelyev,A. and MacKerell,A.D. (2015) Differential impact of the monovalent ions $Li^+$, $Na^+$, $K^+$, and $Rb^+$ on DNA conformational properties. *J. Phys. Chem. Lett.*, **6**, 212–216.
66. Levitt,M. (2001) The birth of computational structural biology. *Nat. Struct. Biol.*, **8**, 392–393.
67. Bixon,M. and Lifson,S. (1967) Potential functions and conformations in cycloalkanes. *Tetrahedron*, **23**, 769–784.
68. Mazur,A.K. and Maaloum,M. (2014) Atomic force microscopy study of DNA flexibility on short length scales: smooth bending versus kinking. *Nucleic Acids Res.*, **42**, 14006–14012.
69. Noy,A. and Golestanian,R. (2010) The chirality of DNA: elasticity cross-terms at base-pair level including A-tracts and the influence of ionic strength. *J. Phys. Chem. B*, **114**, 8022–8031.
70. de Leeuw,S.W., Perram,J.W. and Smith,E.R. (1980) Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, **373**, 27–56.
71. Sagui,C. and Darden,T.A. (1999) Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 155–179.
72. Shui,X., McFail-Isom,L., Hu,G.G. and Williams,L.D. (1998) The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*, **37**, 8341–8355.
73. Shui,X., Sines,C.C., McFail-Isom,L., VanDerveer,D. and Williams,L.D. (1998) Structure of the potassium form of CGCGAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. *Biochemistry*, **37**, 16877–16887.
74. Drew,H., Samson,S. and Dickerson,R.E. (1982) Structure of a B-DNA dodecamer at 16 K. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 4040–4044.
75. Westhof,E. (1987) Re-refinement of the B-dodecamer d(CGCGAATTCGCG) with a comparative analysis of the solvent in it and in the Z-hexamer d(5BrCG5BrCG5BrCG). *J. Biomol. Struct. Dyn.*, **5**, 581–600.
76. Holbrook,S., Dickerson,R. and Kim,S.-H. (1985) Anisotropic thermal-parameter refinement of the DNA dodecamer CGCGAATTCGCG by the segmented rigid-body method. *Acta Crystallogr. B*, **41**, 255–262.
77. Wu,Z., Delaglio,F., Tjandra,N., Zhurkin,V. and Bax,A. (2003) Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and (31)P chemical shift anisotropy. *J. Biomol. Nmr*, **26**, 297–315.
78. Varnai,P. (2002) alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res.*, **30**, 5398–5406.
79. Tian,Y., Kayatta,M., Shultis,K., Gonzalez,A., Mueller,L.J. and Hatcher,M.E. (2009) 31P NMR investigation of backbone dynamics in DNA binding sites. *J. Phys. Chem. B*, **113**, 2596–2603.
80. Schwieters,C.D. and 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**, 1152–1166.
81. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
82. Schwarz,G., Annals,T. and Mar,N. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
83. Djuranovic,D. and Hartmann,B. (2004) DNA fine structure and dynamics in crystals and in solution: the impact of BI/BII backbone conformations. *Biopolymers*, **73**, 356–368.
84. Wecker,K. (2002) The role of the phosphorus BI-BII transition in protein-DNA recognition: the NF-kappaB complex. *Nucleic Acids Res.*, **30**, 4452–4459.
85. Madhumalar,A. and Bansal,M. (2005) Sequence preference for BI/BII conformations in DNA: MD and crystal structure data analysis. *J. Biomol. Struct. Dyn.*, **23**, 13–27.
86. Frederick,C.A., Williams,L.D., Ughetto,G., Van der Marel,G.A., Van Boom,J.H., Rich,A. and Wang,A.H.J. (1990) Structural comparison of anticancer drug-DNA complexes: adriamycin and daunomycin. *Biochemistry*, **29**, 2538–2549.
87. Fei,J. and Ha,T. (2013) Watching DNA breath one molecule at a time. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17173–17174.
88. Cubero,E., Sherer,E.C., Luque,F.J., Orozco,M. and Laughton,C.A. (1999) Observation of spontaneous base pair breathing events in the molecular dynamics simulation of a difluorotoluene-containing DNA oligonucleotide. *J. Am. Chem. Soc.*, **121**, 8653–8654.
89. Zgarbová,M., Otyepka,M., Šponer,J., Lankaš,F. and Jurečka,P. (2014) Base pair fraying in molecular dynamics simulations of DNA and RNA. *J. Chem. Theory Comput.*, **10**, 3177–3189.
90. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
91. Galindo-Murillo,R., Roe,D.R. and Cheatham,T.E. (2014) On the absence of intrahelical DNA dynamics on the μs to ms timescale. *Nat. Commun.*, **5**, 5152.
92. Galindo-Murillo,R., Roe,D.R. and Cheatham,T.E. (2015) Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta*, **1850**, 1041–1058.

# SUPPORTING MATERIAL

# LONG-TIMESCALE DYNAMICS OF THE DREW-DICKERSON DODECAMER

Pablo D. Dans[1,2], Linda Danilāne[1,2,3], Ivan Ivani[1,2], Tomáš Dršata[4], Filip Lankaš[4,5], Adam Hospital[1,2], Jürgen Walther[1,2], Ricard Illa Pujagut[1,2], Federica Battistini[1,2], Josep Lluis Gelpí[6], Richard Lavery[7], and Modesto Orozco[1,2,6*]

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

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

[3]School of Chemistry, University of East Anglia (UEA). Norwich Research Park, Norwich NR4 7TJ, UK.

[4]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic. Flemingovonám 2, 166 10 Prague, Czech Republic.

[5]Laboratory of Informatics and Chemistry, University of Chemistry and Technology Prague, Technická 5, 166 28 Prague, Czech Republic.

[6]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain.

[7]Bases Moléculaires et Structurales des Systèmes Infectieux, Univ. Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, Lyon 69367, France.

[*] To whom correspondence should be addressed: Prof. Modesto Orozco, Tel: +34 934037155, Fax: +34 934037157, Email: modesto.orozco@irbbarcelona.org.

**Table S1.** PDB structures of DDD used to build the experimental conformational space.

| PDB Code | Method | Resolution | Number models | Title (taken from the PDB metadata) |
|---|---|---|---|---|
| 109D | X-RAY | 2.00 | 1 | Variability In DNA Minor Groove Width Recognised By Ligand Binding: The Crystal Structure Of A Bis-Benzimidazole Compound Bound To The DNA Duplex d(CGCGAATTCGCG)2 |
| 127D | X-RAY | 2.00 | 1 | DNA (5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3') Complexed With Hoechst 33258 |
| 129D | X-RAY | 2.25 | 1 | DNA (5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3') Complexed With Hoechst 33342 |
| 166D | X-RAY | 2.20 | 1 | Drug-DNA Minor Groove Recognition: Crystal Structure Of Gamma-Oxapentamidine Complexed With d(CGCGAATTCGCG)2 |
| 171D | NMR | --- | 1 | Solution Structure Of A DNA Dodecamer Containing The Anti-Neoplastic Agent Arabinosylcytosine: Combined Use Of NMR. Restrained Molecular Dynamics And Full Relaxation Matrix Refinement |
| 1BNA | X-RAY | 1.90 | 1 | Structure Of A B-DNA Dodecamer. Conformation And Dynamics |
| 1D30 | X-RAY | 2.40 | 1 | The Structure Of Dapi Bound To DNA |
| 1D43 | X-RAY | 2.00 | 1 | DNA Dodecamer C-G-C-G-A-A-T-T-C-G-C-G/Hoechst 33258 Complex: 0 Degrees C. Piperazine Up |
| 1D44 | X-RAY | 2.00 | 1 | DNA Dodecamer C-G-C-G-A-A-T-T-C-G-C-G/Hoechst 33258 Complex: 0 Degrees C. Piperazine Down |
| 1D45 | X-RAY | 1.90 | 1 | DNA Dodecamer C-G-C-G-A-A-T-T-C-G-C-G/Hoechst 33258 Complex:-25 Degrees C. Piperazine Down |
| 1D46 | X-RAY | 2.00 | 1 | DNA Dodecamer C-G-C-G-A-A-T-T-C-G-C-G/Hoechst 33258 Complex:-100 Degrees C. Piperazine Down |
| 1D64 | X-RAY | 2.10 | 1 | Crystal Structure Of A Pentamidine-Oligonucleotide Complex: Implications For DNA-Binding Properties |
| 1D86 | X-RAY | 2.20 | 1 | Structural Consequences Of A Carcinogenic Alkylation Lesion On DNA: Effect Of O6-Ethyl-Guanine On The Molecular Structure Of d(CGC[E6G]AATTCGCG)-Netropsin Complex |
| 1DNH | X-RAY | 2.25 | 1 | The Molecular Structure Of The Complex Of Hoechst 33258 And The DNA Dodecamer d(CGCGAATTCGCG) |
| 1DOU | X-RAY | 1.82 | 1 | Monovalent Cations Sequester Within The A-Tract Minor Groove Of [d(CGCGAATTCGCG)]2 |
| 1DUF | NMR | --- | 5 | The NMR Structure Of DNA Dodecamer Determined In Aqueous Dilute Liquid Crystalline Phase |
| 1EEL | X-RAY | 2.40 | 1 | Structure Of A Complex Between The DNA Sequence CGCGAATTCGCG And Bis[Piperidino-Ethyl]-Furamidine |
| 1FMQ | X-RAY | 2.00 | 1 | Cyclo-Butyl-Bis-Furamidine Complexed With CGCGAATTCGCG |
| 1FMS | X-RAY | 1.90 | 1 | Structure Of Complex Between Cyclohexyl-Bis-Furamidine And d(CGCGAATTCGCG) |
| 1FQ2 | X-RAY | 1.20 | 1 | Crystal Structure Analysis Of The Potassium Form Of B-DNA Dodecamer CGCGAATTCGCG |
| 1FTD | X-RAY | 2.00 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'-Symmetric Bis-Benzimidazole Complex |
| 1G3X | X-RAY | 2.70 | 1 | Intercalation Of An 9acridine-Peptide Drug In A DNA Dodecamer |
| 1GIP | NMR | --- | 21 | The NMR Structure Of DNA Dodecamer Determined In Aqueous Dilute Liquid Crystalline Phase |
| 1JGR | X-RAY | 1.20 | 1 | Crystal Structure Analysis Of The B-DNA Dodecamer CGCGAATTCGCG With Thallium Ions. |
| 1LEX | X-RAY | 2.25 | 1 | Structure Of A Dicationic Monoimidazole Lexitropsin Bound To DNA (Orientation 1) |
| 1LEY | X-RAY | 2.25 | 1 | Structure Of A Dicationic Monoimidazole Lexitropsin Bound To DNA (Orientation 2) |
| 1M6F | X-RAY | 1.78 | 1 | Strong Binding In The DNA Minor Groove By An Aromatic Diamidine With A Shape That Does Not Match The Curvature Of The Groove |
| 1NAJ | NMR | --- | 5 | High Resolution NMR Structure Of DNA Dodecamer Determined In Aqueous Dilute Liquid Crystalline Phase |
| 1PRP | X-RAY | 2.10 | 1 | Crystal Structure Of d(CGCGAATTCGCG) Complexed With Propamidine. A Short-Chain Homologue Of The Drug Pentamidin |

| | | | | |
|---|---|---|---|---|
| 1QV4 | X-RAY | 2.50 | 1 | B-DNA Dodecamer CGTGAATTCACG Complexed With Minor Groove Binder Methylproamine |
| 1QV8 | X-RAY | 2.50 | 1 | B-DNA Dodecamer d(CGCGAATTCGCG)2 Complexed With Proamine |
| 1QXB | NMR | --- | 1 | NMR Structure Determination Of The Self Complementary DNA Dodecamer CGCGAATT*CGCG In Which A Ribose Is Inserted Between The 3'-OH Of T8 And The 5'-Phosphate Group Of C9 |
| 1VZK | X-RAY | 1.77 | 1 | A Thiophene Based Diamidine Forms A ""Super"" AT Binding Minor Groove Agent |
| 1ZPH | X-RAY | 1.80 | 1 | Crystal Structure Analysis Of The Minor Groove Binding Quinolinium Quaternary Salt SN 8315 Complexed With CGCGAATTCGCG |
| 1ZPI | X-RAY | 1.60 | 1 | Crystal Structure Analysis Of The Minor Groove Binding Quinolinium Quaternary Salt SN 8224 Complexed With CGCGAATTCGCG |
| 227D | X-RAY | 2.20 | 1 | A Crystallographic And Spectroscopic Study Of The Complex Between d(CGCGAATTCGCG)2 And 2.5-Bis(4-Guanylphenyl)Furan. An Analogue Of Berenil. Structural Origins Of Enhanced DNA-Binding Affinity |
| 289D | X-RAY | 2.20 | 1 | Targeting The Minor Groove Of DNA: Crystal Structures Of Two Complexes Between Furan Derivatives Of Berenil And The DNA Dodecamer d(CGCGAATTCGCG)2 |
| 298D | X-RAY | 2.20 | 1 | Targeting The Minor Groove Of DNA: Crystal Structures Of Two Complexes Between Furan Derivatives Of Berenil And The DNA Dodecamer d(CGCGAATTCGCG)2 |
| 2B0K | X-RAY | 1.64 | 1 | Crystal Structure Of The DB921-d(CGCGAATTCGCG)2 Complex. |
| 2B3E | X-RAY | 1.36 | 1 | Crystal Structure Of DB819-d(CGCGAATTCGCG)2 Complex. |
| 2BNA | X-RAY | 2.70 | 1 | Structure Of A B-DNA Dodecamer At 16 Kelvin |
| 2DAU | NMR | --- | 1 | Dickerson-Drew DNA Dodecamer. NMR. Minimized Average Structure |
| 2DBE | X-RAY | 2.50 | 1 | Crystal Structure Of A Berenil-Dodecanucleotide Complex: The Role Of Water In Sequence-Specific Ligand Binding |
| 2DYW | X-RAY | 1.13 | 1 | A Backbone Binding DNA Complex |
| 2GVR | X-RAY | 1.65 | 1 | Crystal Structure Of The Berenil-d(CGCGAATTCGCG)2 Complex At 1.65 A Resolution. |
| 2GYX | X-RAY | 1.86 | 1 | Crystal Structure Of DB884- d(CGCGAATTCGCG)2 Complex At 1.86 A Resolution. |
| 2I2I | X-RAY | 1.63 | 1 | Crystal Structure Of The DB293-d(CGCGAATTCGCG)2 Complex. |
| 2I5A | X-RAY | 1.65 | 1 | Crystal Structure Of A DB1055-d(CGCGAATTCGCG)2 Complex |
| 2L7D | NMR | --- | 5 | Ribonucleotide Perturbation Of DNA Structure: Solution Structure Of [d(CGC)R(G)d(AATTCGCG)]2 |
| 2NLM | X-RAY | 2.05 | 1 | Crystal Structure Of The DB 911- d(CGCGAATTCGCG)2 Complex At 2.05 A Resolution. |
| 302D | X-RAY | 2.20 | 1 | Meta-Hydroxy Analogue Of Hoechst 33258 ('Hydroxyl In' Conformation) Bound To d(CGCGAATTCGCG)2 |
| 303D | X-RAY | 2.20 | 1 | Meta-Hydroxy Analogue Of Hoechst 33258 ('Hydroxyl Out' Conformation) Bound To d(CGCGAATTCGCG)2 |
| 311D | X-RAY | 2.20 | 1 | The Role Of Hydrogen Bonding In Minor-Groove Drug-DNA Recognition. Structure Of A Bis-Amidinium Derivative Of Hoechst 33258 Complexed To The Dodecanucleotide d(CGCGAATTCGCG)2 |
| 328D | X-RAY | 2.60 | 1 | Structure Of A d(CGCGAATTCGCG)2-Sn7167 Complex |
| 355D | X-RAY | 1.40 | 1 | The B-DNA Dodecamer At High Resolution |
| 360D | X-RAY | 1.85 | 1 | Structure Of 2.5-Bis{[4-(N-Ethylamidino)Phenyl]}Furan Complexed To 5'-d(CGCGAATTCGCG)-3'. A Minor Groove Drug Complex. Showing Patterns Of Groove Hydration |
| 3OIE | X-RAY | 1.90 | 1 | Crystal Structure Of The DB1880-d(CGCGAATTCGCG)2 Complex |
| 3U05 | X-RAY | 1.27 | 1 | Crystal Structure Of DB1804-d(CGCGAATTCGCG)2 Complex |
| 3U08 | X-RAY | 1.25 | 1 | Crystal Structure Of DB1963-d(CGCGAATTCGCG)2 Complex At 1.25 A Resolution |
| 3U0U | X-RAY | 1.24 | 1 | Crystal Structure Of The DB1883-d(CGCGAATTCGCG)2 Complex At 1.24 A Resolution |
| 3U2N | X-RAY | 1.25 | 1 | Crystal Structure Of DNA(CGCGAATTCGCG)2 At 1.25 Angstroms |
| 428D | X-RAY | 1.75 | 1 | Structure Of The Potassium Form Of CGCGAATTCGCG: DNA Deformation By Electrostatic Collapse Around Small Cations |
| 442D | X-RAY | 1.60 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'. Benzimidazole Derivative Complex |
| 443D | X-RAY | 1.60 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'/ Benzimidazole Derivative |

| | | | | |
|---|---|---|---|---|
| | | | | Complex |
| 444D | X-RAY | 2.40 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'. Benzimidazole Derivative Complex |
| 445D | X-RAY | 2.60 | 1 | 5'-d(*CP*GP*CP*GP*AP*AP*TP*TP*CP*GP*CP*G)-3'. Benzimidazole Derivative Complex |
| 447D | X-RAY | 2.20 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3' |
| 448D | X-RAY | 2.20 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'. Benzimidazole Derivative Complex |
| 449D | X-RAY | 2.10 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'. Benzimidazole Derivative Complex |
| 453D | X-RAY | 1.80 | 1 | 5'-d(*Cp*Gp*Cp*Gp*Ap*Ap*Tp*Tp*Cp*Gp*Cp*G)-3'-Benzimidazole Complex |
| 455D | X-RAY | 1.43 | 1 | A6/A18 Inter-Strand Dithiobis(Propane)-Crosslinked Dodecamer (CGCGAATTCGCG)2 |
| 4AGZ | X-RAY | 1.25 | 1 | Crystal Structure Of The Db 985-d(CGCGAATTCGCG)2 Complex At 1.25 A Resolution. |
| 4C64 | X-RAY | 1.32 | 1 | Ultra High Resolution Dickerson-Drew Dodecamer B-DNA |
| 4U8A | X-RAY | 1.48 | 1 | Crystal Structure Of d(CGCGAATTCGCG)2 Complexed With BPH-1503 |
| 4U8B | X-RAY | 1.31 | 1 | Crystal Structure Of d(CGCGAATTCGCG)2 Complexed With BPH-1358 |
| 4U8C | X-RAY | 1.24 | 1 | Crystal Structure Of d(CGCGAATTCGCG)2 Complexed With BPH-1409 |
| 5BNA | X-RAY | 2.60 | 1 | The Primary Mode Of Binding Of Cisplatin To A B-DNA Dodecamer: C-G-C-G-A-A-T-T-C-G-C-G |
| 7BNA | X-RAY | 1.90 | 1 | Anisotropic Thermal-Parameter Refinement Of The DNA Dodecamer CGCGAATTCGCG By The Segmented Rigid-Body Method |
| 8BNA | X-RAY | 2.20 | 1 | Binding Of Hoechst 33258 To The Minor Groove Of B-DNA |

**Table S2.** Different metrics showing the quality of simulations performed.

| System | RMSd | RMSd (without ends) | RMSd per bp (without ends) | % Watson-Crick Hbonds |
|---|---|---|---|---|
| Na+ / J&C / neutrality / SPCE | 2.00 | 1.85 | 0.19 | 98 |
| Na+ / S&D / neutrality / SPCE | 1.91 | 1.77 | 0.18 | 98 |
| Na+ / S&D / neutrality / TIP3P | 1.90 | 1.60 | 0.16 | 97 |
| Na+ / J&C / neutrality / TIP3P | 2.01 | 1.78 | 0.18 | 96 |
| Na+Cl- / S&D / 0.15 M / TIP3P | 2.06 | 1.70 | 0.17 | 93 |
| Na+Cl- / S&D / 0.5 M / TIP3P | 2.12 | 1.83 | 0.18 | 95 |
| Na+Cl- / S&D / 2.0 M / TIP3P | 2.24 | 1.96 | 0.20 | 92 |
| K+ / J&C / neutrality / SPCE | 2.29 | 1.59 | 0.16 | 91 |
| K+ / S&D / neutrality / SPCE | 1.86 | 1.65 | 0.17 | 98 |
| K+ / S&D / neutrality / TIP3P | 1.95 | 1.75 | 0.18 | 95 |
| K+Cl- / S&D / 0.15 M / SPCE | 2.23 | 1.99 | 0.20 | 96 |
| K+Cl- / S&D / 0.5 M / SPCE | 1.90 | 1.65 | 0.17 | 95 |
| K+Cl- / S&D / 2.0 M / SPCE | 1.97 | 1.72 | 0.17 | 95 |
| K+ / J&J / neutrality / SPCE | 2.18 | 1.80 | 0.18 | 97 |
| K+ / B&R / neutrality / SPCE | 1.82 | 1.63 | 0.16 | 96 |

**Table S3.** Sequence-averaged conformational parameters obtained from all the Dickerson-Drew dodecamer simulations performed with sodium.

| Systems (Na+) | J&C, neutral, SPCE | | S&D, neutral, SPCE | | J&C, neutral, TIP3P | | S&D, 0.15M, TIP3P | | S&D, 0.5M, TIP3P | | S&D, 2.0M, TIP3P | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | Avg | SD | Avg | SD | Avg | SD | Avg | SD | Avg | SD | Avg | SD |
| Shear | 0.00 | 0.30 | 0.00 | 0.30 | 0.00 | 0.30 | -0.01 | 0.33 | 0.00 | 0.30 | 0.00 | 0.32 |
| Stretch | 0.02 | 0.11 | 0.02 | 0.11 | 0.02 | 0.11 | 0.02 | 0.12 | 0.02 | 0.11 | 0.02 | 0.12 |
| Stagger | 0.09 | 0.38 | 0.08 | 0.38 | 0.10 | 0.38 | 0.11 | 0.41 | 0.11 | 0.38 | 0.13 | 0.39 |
| Buckle | 0.2 | 9.4 | 0.1 | 9.4 | 0.1 | 9.7 | 0.4 | 10.0 | -0.1 | 9.7 | 0.0 | 9.8 |
| Propeller | -9.5 | 8.2 | -9.1 | 8.2 | -9.3 | 8.4 | -9.0 | 8.4 | -9.6 | 8.3 | -9.7 | 8.6 |
| Opening | 1.2 | 3.8 | 1.3 | 3.9 | 1.3 | 3.9 | 1.4 | 4.1 | 1.3 | 4.0 | 1.4 | 4.1 |
| Xdisp | -0.3 | 1.0 | -0.4 | 1.0 | -0.6 | 1.0 | -0.5 | 1.1 | -0.5 | 1.1 | -0.4 | 1.1 |
| Ydisp | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.9 |
| Inclination | 1.7 | 5.7 | 1.8 | 5.6 | 2.3 | 5.6 | 2.0 | 5.7 | 2.1 | 5.7 | 1.8 | 5.8 |
| Tip | 0.3 | 5.2 | 0.2 | 5.3 | 0.1 | 6.8 | -0.2 | 11.9 | -0.1 | 5.3 | 0.1 | 7.2 |
| Shift | -0.05 | 0.83 | -0.02 | 0.83 | 0.00 | 0.82 | 0.02 | 0.80 | 0.01 | 0.81 | -0.02 | 0.80 |
| Slide | -0.16 | 0.52 | -0.18 | 0.52 | -0.24 | 0.53 | -0.23 | 0.53 | -0.21 | 0.53 | -0.19 | 0.53 |
| Rise | 3.31 | 0.29 | 3.31 | 0.28 | 3.32 | 0.29 | 3.33 | 0.30 | 3.32 | 0.29 | 3.32 | 0.29 |
| Tilt | -0.2 | 4.8 | -0.1 | 4.8 | 0.0 | 4.7 | 0.1 | 4.8 | 0.0 | 4.6 | -0.1 | 4.6 |
| Roll | 1.1 | 5.4 | 1.3 | 5.4 | 1.5 | 5.5 | 1.5 | 5.4 | 1.4 | 5.4 | 1.1 | 5.6 |
| Twist | 34.6 | 5.5 | 34.6 | 5.4 | 34.3 | 5.5 | 34.5 | 5.5 | 34.5 | 5.4 | 34.6 | 5.5 |
| α | -72.6 | 16.6 | -72.7 | 18.2 | -72.8 | 18.5 | -73.0 | 19.4 | -73.3 | 20.5 | -72.4 | 19.5 |
| β | 166.5 | 19.8 | 166.6 | 21.5 | 166.3 | 22.6 | 165.3 | 22.2 | 165.5 | 24.6 | 166.2 | 21.9 |
| γ | 54.7 | 18.1 | 55.3 | 21.7 | 56.1 | 24.3 | 57.8 | 24.8 | 56.9 | 27.6 | 55.6 | 23.9 |
| δ | 136.9 | 13.1 | 136.7 | 13.3 | 135.4 | 14.4 | 135.9 | 14.2 | 135.8 | 14.2 | 135.8 | 14.1 |
| ε | -157.4 | 25.0 | -158.3 | 24.4 | -159.6 | 23.5 | -159.4 | 23.5 | -159.0 | 23.5 | -159.0 | 24.0 |
| ζ | -117.0 | 39.1 | -115.0 | 38.4 | -111.2 | 36.4 | -109.7 | 38.2 | -111.2 | 36.3 | -112.9 | 36.5 |
| χ | -110.0 | 15.4 | -110.5 | 15.8 | -112.0 | 16.3 | -111.8 | 16.4 | -111.6 | 16.8 | -111.1 | 16.1 |
| Phase | 153.6 | 24.4 | 153.4 | 24.6 | 152.0 | 26.8 | 152.9 | 26.5 | 152.6 | 26.8 | 152.4 | 26.7 |
| Amplitude | 41.5 | 6.5 | 41.5 | 6.5 | 41.3 | 6.4 | 41.3 | 6.3 | 41.4 | 6.4 | 41.5 | 6.5 |

**Table S4.** Sequence-averaged conformational parameters obtained from all the Dickerson-Drew dodecamer simulations performed with potassium.

| Systems (K+) | J&C, neutral, SPCE | | S&D, neutral, SPCE | | S&D, 0.5M, SPCE | | S&D, 0.15M, SPCE | | S&D, 2.0M, SPCE | | S&D, neutral, TIP3P | | J&J, neutral, SPCE | | B&R, neutral, SPCE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | Avg | SD | Avg | SD | Avg | SD | Avg | SD | Avg | SD | Avg | SD | Avg | SD | Avg | SD |
| Shear | -0.01 | 0.32 | 0.00 | 0.30 | -0.01 | 0.31 | 0.00 | 0.29 | 0.00 | 0.31 | 0.0 | 0.30 | 0.00 | 0.30 | 0.00 | 0.30 |
| Stretch | 0.02 | 0.12 | 0.02 | 0.11 | 0.02 | 0.12 | 0.02 | 0.11 | 0.02 | 0.12 | 0.02 | 0.12 | 0.02 | 0.11 | 0.02 | 0.11 |
| Stagger | 0.17 | 0.38 | 0.10 | 0.38 | 0.12 | 0.38 | 0.11 | 0.37 | 0.13 | 0.37 | 0.11 | 0.38 | 0.08 | 0.38 | 0.11 | 0.38 |
| Buckle | 0.9 | 9.5 | -0.1 | 9.5 | 0.1 | 9.6 | 0.1 | 9.4 | 0.0 | 9.6 | 0.0 | 9.6 | 0.0 | 9.7 | 0.1 | 9.5 |
| Propeller | -9.1 | 8.3 | -8.7 | 8.1 | -8.8 | 8.2 | -8.4 | 8.3 | -8.9 | 8.3 | -8.5 | 8.3 | -8.2 | 8.1 | -8.4 | 8.2 |
| Opening | 0.9 | 4.0 | 1.1 | 3.9 | 1.1 | 4.1 | 1.0 | 3.9 | 1.1 | 4.0 | 1.1 | 4.0 | 1.3 | 4.0 | 1.2 | 3.9 |
| Xdisp | -0.4 | 1.0 | -0.5 | 1.0 | -0.4 | 1.0 | -0.4 | 1.0 | -0.4 | 1.0 | -0.5 | 1.0 | -0.8 | 1.0 | -0.5 | 1.0 |
| Ydisp | -0.2 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.8 |
| Inclination | 0.2 | 5.8 | 1.8 | 5.5 | 1.6 | 5.6 | 1.5 | 5.6 | 1.4 | 5.5 | 1.8 | 5.6 | 2.7 | 5.6 | 1.9 | 5.5 |
| Tip | 0.3 | 6.0 | 0.0 | 5.4 | 0.0 | 5.6 | -0.1 | 5.3 | 0.0 | 6.0 | 0.0 | 5.6 | 0.1 | 5.8 | 0.0 | 5.4 |
| Shift | 0.03 | 0.75 | 0.00 | 0.77 | -0.01 | 0.77 | -0.01 | 0.77 | 0.00 | 0.76 | 0.00 | 0.78 | 0.00 | 0.81 | 0.01 | 0.79 |
| Slide | -0.16 | 0.50 | -0.23 | 0.51 | -0.21 | 0.51 | -0.20 | 0.51 | -0.19 | 0.51 | -0.21 | 0.51 | -0.28 | 0.53 | -0.21 | 0.51 |
| Rise | 3.32 | 0.29 | 3.31 | 0.29 | 3.31 | 0.29 | 3.31 | 0.29 | 3.31 | 0.29 | 3.31 | 0.29 | 3.32 | 0.29 | 3.31 | 0.29 |
| Tilt | 0.3 | 4.6 | 0.0 | 4.5 | 0.0 | 4.5 | 0.0 | 4.6 | 0.0 | 4.5 | 0.0 | 4.6 | 0.0 | 4.6 | 0.0 | 4.6 |
| Roll | 0.5 | 5.3 | 1.3 | 5.3 | 1.2 | 5.3 | 1.2 | 5.3 | 1.1 | 5.3 | 1.4 | 5.4 | 1.6 | 5.5 | 1.3 | 5.3 |
| Twist | 34.9 | 5.5 | 34.5 | 5.5 | 34.6 | 5.6 | 34.6 | 5.5 | 34.8 | 5.5 | 34.4 | 5.6 | 34.1 | 5.5 | 34.6 | 5.4 |
| α | -73 | 22 | -72 | 17 | -72 | 17 | -72 | 18 | -72 | 17 | -73 | 17 | -72 | 18 | -72 | 17 |
| β | 164 | 22 | 168 | 20 | 167 | 21 | 166 | 21 | 168 | 19 | 167 | 21 | 168 | 20 | 167 | 20 |
| γ | 59 | 27 | 55 | 19 | 56 | 21 | 56 | 21 | 55 | 18 | 56 | 21 | 55 | 19 | 55 | 19 |
| δ | 137 | 12 | 136 | 14 | 136 | 14 | 136 | 13 | 136 | 13 | 136 | 14 | 135 | 14 | 136 | 13 |
| ε | -159 | 22 | -160 | 23 | -159 | 23 | -159 | 23 | -160 | 23 | -160 | 23 | -161 | 23 | -160 | 23 |
| ζ | -113 | 34 | -114 | 36 | -114 | 36 | -115 | 36 | -114 | 35 | -113 | 36 | -112 | 36 | -114 | 35 |
| χ | -110 | 15 | -111 | 15 | -111 | 16 | -111 | 15 | -111 | 15 | -111 | 16 | -112 | 16 | -111 | 15 |
| Phase | 154 | 23 | 152 | 26 | 153 | 25 | 153 | 25 | 152 | 25 | 153 | 26 | 152 | 27 | 153 | 25 |
| Amplitude | 41.3 | 6.2 | 41.3 | 6.5 | 41.3 | 6.4 | 41.4 | 6.4 | 41.3 | 6.4 | 41.3 | 6.4 | 41.3 | 6.5 | 41.4 | 6.4 |

**Table S5**. Peak positions inferred from experimental and computational solution scattering profiles.[a]

| | | Computed | | | |
|---|---|---|---|---|---|
| Peak | Exp | 1BNA | 1GIP | BSC1 avg | BSC1 MD |
| P1 | 0.456[b] | --- | 0.460 (0.004)[c] | 0.410 (0.046) | 0.448 (0.008) ± 0.014 |
| P2 | 0.750 | 0.750 (0.000) | 0.770 (0.020) | 0.720 (0.030) | 0.725 (0.025) ± 0.042 |
| P3 | 1.127 | 1.145 (0.018) | 1.150 (0.023) | 1.110 (0.017) | 1.083 (0.044) ± 0.055 |
| P4 | 1.513 | 1.520 (0.007) | 1.530 (0.017) | 1.510 (0.003) | 1.506 (0.007) ± 0.017 |
| P5 | 1.834 | 1.900 (0.066) | 1.850 (0.016) | 1.910 (0.076) | 1.864 (0.030) ± 0.013 |

[a] Values are reported in $Å^{-1}$. [b] Peak positions were determined from zero crossing points in the first derivative [Savelyev A., MacKerell Jr. D. *J. Phys. Chem. Letter* **2015**, 6, 212]. [c] In bracket we report the absolute difference respect to the experimental value.

**Table S6.** Global DNA flexibility in the Cartesian and Helical space comparing parmbsc1 simulations with parmbsc0.[a]

| Force-field | Water model | Ion model | Ions | Entropy all heavy | | Entropy backbone | | First 3 eigenval. | Eigenval. 10, 20 and 30 | Self-similarity | P.Length (nm)[b] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Parmbsc0 | TIP3P | S&D | Na+, neutral | 2.21 | *2.11* | 1.40 | *1.30* | 153,113,66 | 17,6,3 | 0.89 | 51.4±32.2 |
| Parmbsc1 | TIP3P | S&D | Na+, neutral | 2.29 | *2.19* | 1.50 | *1.40* | 140,116,63 | 18,7,4 | 0.95 | 49.1±31.5 |
| Parmbsc1 | TIP3P | S&D | Na+Cl-, 0.5M | 2.17 | *2.07* | 1.16 | *1.06* | 136,107,61 | 15,6,3 | 0.91 | 53.2±33.0 |
| Parmbsc1 | TIP3P | S&D | Na+Cl-, 0.15M | 2.18 | *2.08* | 1.15 | *1.06* | 138,110,63 | 16,6,3 | 0.90 | 52.7±30.5 |
| Parmbsc1 | TIP3P | J&C | Na+, neutral | 2.24 | *2.14* | 1.46 | *1.36* | 131,113,62 | 17,7,4 | 0.85 | 52.0±32.6 |
| Parmbsc1 | SPCE | J&C | Na+, neutral | 2.19 | *2.08* | 1.42 | *1.32* | 142,107,46 | 14,6,3 | 0.86 | 52.5±33.7 |
| Parmbsc1 | SPCE | J&C | K+, neutral | 2.21 | *2.25* | 1.22 | *1.16* | 149,133,69 | 16,6,3 | 0.80 | 56.7±35.0 |
| Parmbsc1 | SPCE | S&D | K+Cl-, 0.5M | 2.18 | *2.08* | 1.20 | *1.10* | 132,101,50 | 16,6,3 | 0.85 | 55.4±34.7 |
| Parmbsc1 | SPCE | S&D | K+Cl-, 0.15M | 2.19 | *2.09* | 1.19 | *1.10* | 138,108,55 | 17,6,3 | 0.89 | 48.7±30.5 |
| Parmbsc1 | SPCE | S&D | K+, neutral | 2.19 | *2.09* | 1.18 | *1.10* | 131,96,51 | 17,6,3 | 0.96 | 48.9±30.5 |
| Parmbsc1 | SPCE | J&J | K+, neutral | 2.22 | *2.26* | 1.21 | *1.15* | 127,104,54 | 21,7,4 | 0.84 | 51.3±31.8 |
| Parmbsc1 | SPCE | B&R | K+, neutral | 2.15 | *2.20* | 1.16 | *1.10* | 134,100,48 | 15,6,4 | 0.88 | 54.5±34.9 |
| Parmbsc1 | TIP3P | S&D | Na+Cl-, 2M | 2.30 | *2.34* | 1.28 | *1.21* | 171,123,69 | 22,8,4 | 0.84 | 51.2±32.9 |
| Parmbsc1 | SPCE | S&D | K+Cl-, 2M | 2.21 | *2.25* | 1.21 | *1.15* | 136,110,50 | 18,7,3 | 0.85 | 56.5±35.2 |
| Parmbsc1 | SPCE | J&C | K+, neutral | 2.21 | *2.25* | 1.22 | *1.16* | 149,133,69 | 16,6,3 | 0.80 | 57.4±35.3 |

[a] Each line represents a different simulation (see Table 1 for a complete description of the simulated systems). Entropies (in kcal/mol K) were determined using Schlitter's (in roman) and Andricioaei-Karplus' (in italic) methods [Schlitter J., *Chem. Phys. Lett.* **1993**, *215*, 617–621. Andricioaei I., and Karplus M., *J. Chem. Phys.* **2001**, *115*, 6289]. [b] Computed using a local implementation of a coarse-grained Monte Carlo algorithm [Ivani I., et al. *Nature Methods* **2016**, 10.1038/nmeth.3658].

**Table S7**. Concentration in Molarity (and ion population) of K+ ions *inside* the minor groove of DDD.

| | | J&J neutrality | B&R neutrality | J&C neutrality | S&D neutrality | S&D 0.15M | S&D 0.5M | S&D 2.0M |
|---|---|---|---|---|---|---|---|---|
| 1 | CG | 0.14 (0.02) | 0.44 (0.07) | 3.10 (0.12) | 0.33 (0.05) | 0.41 (0.08) | 0.93 (0.15) | 1.11 (0.19) |
| 2 | GC | 0.26 (0.04) | 0.75 (0.12) | 0.88 (0.14) | 0.64 (0.10) | 0.71 (0.12) | 0.89 (0.14) | 1.04 (0.16) |
| 3 | CG | 0.49 (0.07) | 2.11 (0.30) | 3.92 (0.63) | 2.07 (0.30) | 1.90 (0.27) | 2.36 (0.34) | 2.73 (0.39) |
| 4 | GA | 0.43 (0.06) | 1.27 (0.18) | 0.80 (0.14) | 1.11 (0.16) | 1.23 (0.17) | 1.14 (0.17) | 1.30 (0.19) |
| 5 | AA | 0.31 (0.04) | 1.73 (0.21) | 2.30 (0.29) | 1.63 (0.20) | 1.40 (0.16) | 1.75 (0.21) | 1.48 (0.18) |
| 6 | AT | 0.18 (0.02) | 0.84 (0.10) | 4.37 (0.52) | 1.59 (0.19) | 3.03 (0.35) | 2.11 (0.25) | 2.40 (0.28) |
| 7 | TT | 0.32 (0.04) | 2.39 (0.29) | 2.27 (0.27) | 1.83 (0.22) | 1.25 (0.15) | 1.68 (0.20) | 1.67 (0.20) |
| 8 | TC | 0.51 (0.07) | 1.22 (0.18) | 1.29 (0.19) | 1.11 (0.16) | 1.15 (0.17) | 1.18 (0.17) | 1.38 (0.20) |
| 9 | CG | 0.51 (0.07) | 2.17 (0.31) | 3.52 (0.50) | 2.00 (0.29) | 1.99 (0.29) | 2.32 (0.33) | 2.41 (0.35) |
| 10 | GC | 0.26 (0.04) | 0.64 (0.10) | 0.87 (0.14) | 0.63 (0.10) | 0.78 (0.12) | 0.96 (0.15) | 1.07 (0.17) |
| 11 | CG | 0.17 (0.03) | 0.43 (0.07) | 1.12 (0.18) | 0.41 (0.07) | 0.56 (0.09) | 0.69 (0.11) | 1.20 (0.21) |

**Table S8**. Concentration in Molarity (and ion population) of K+ ions *inside* the major groove of DDD.

| | | J&J neutrality | B&R neutrality | J&C neutrality | S&D neutrality | S&D 0.15M | S&D 0.5M | S&D 2.0M |
|---|---|---|---|---|---|---|---|---|
| 1 | CG | 0.15 (0.05) | 0.19 (0.06) | 1.19 (0.08) | 0.17 (0.06) | 0.36 (0.13) | 0.48 (0.15) | 0.89 (0.28) |
| 2 | GC | 0.30 (0.10) | 0.61 (0.20) | 0.67 (0.22) | 0.70 (0.23) | 0.93 (0.29) | 1.14 (0.37) | 1.98 (0.65) |
| 3 | CG | 0.33 (0.10) | 0.44 (0.14) | 0.48 (0.14) | 0.46 (0.14) | 0.50 (0.16) | 0.66 (0.21) | 1.12 (0.34) |
| 4 | GA | 0.67 (0.24) | 0.95 (0.33) | 1.06 (0.34) | 0.99 (0.34) | 1.03 (0.37) | 1.20 (0.42) | 1.77 (0.61) |
| 5 | AA | 0.75 (0.27) | 0.83 (0.29) | 0.74 (0.25) | 0.77 (0.27) | 0.77 (0.29) | 0.93 (0.33) | 1.28 (0.45) |
| 6 | AT | 0.83 (0.31) | 1.02 (0.37) | 0.94 (0.34) | 0.92 (0.34) | 0.99 (0.37) | 1.15 (0.42) | 1.54 (0.56) |
| 7 | TT | 0.75 (0.27) | 0.88 (0.31) | 0.73 (0.25) | 0.76 (0.27) | 0.85 (0.30) | 0.94 (0.33) | 1.28 (0.45) |
| 8 | TC | 0.63 (0.22) | 0.97 (0.33) | 1.11 (0.39) | 0.91 (0.32) | 1.06 (0.37) | 1.19 (0.42) | 1.79 (0.61) |
| 9 | CG | 0.33 (0.10) | 0.46 (0.14) | 0.49 (0.15) | 0.45 (0.14) | 0.56 (0.17) | 0.69 (0.21) | 1.12 (0.34) |
| 10 | GC | 0.33 (0.11) | 0.62 (0.20) | 0.98 (0.32) | 0.73 (0.24) | 0.85 (0.28) | 1.16 (0.38) | 1.91 (0.62) |
| 11 | CG | 0.15 (0.06) | 0.22 (0.07) | 0.23 (0.07) | 0.20 (0.06) | 0.26 (0.08) | 0.41 (0.13) | 0.95 (0.31) |

**Table S9.** Concentration in Molarity (and ion population) of K+ ions *outside* the major groove of DDD.

| | | J&J neutrality | B&R neutrality | J&C neutrality | S&D neutrality | S&D 0.15M | S&D 0.5M | S&D 2.0M |
|---|---|---|---|---|---|---|---|---|
| 1 | CG | 0.24 (0.24) | 0.31 (0.29) | 0.99 (0.19) | 0.30 (0.28) | 0.46 (0.48) | 0.84 (0.77) | 1.84 (1.78) |
| 2 | GC | 0.33 (0.31) | 0.37 (0.36) | 0.32 (0.30) | 0.37 (0.36) | 0.59 (0.55) | 0.92 (0.88) | 1.97 (1.89) |
| 3 | CG | 0.43 (0.38) | 0.46 (0.41) | 0.45 (0.38) | 0.47 (0.42) | 0.67 (0.61) | 1.04 (0.92) | 2.10 (1.84) |
| 4 | GA | 0.50 (0.49) | 0.52 (0.51) | 0.46 (0.45) | 0.52 (0.51) | 0.72 (0.72) | 1.06 (1.05) | 2.08 (2.06) |
| 5 | AA | 0.53 (0.52) | 0.55 (0.55) | 0.52 (0.51) | 0.55 (0.55) | 0.73 (0.76) | 1.08 (1.09) | 2.07 (2.07) |
| 6 | AT | 0.55 (0.54) | 0.59 (0.59) | 0.57 (0.57) | 0.59 (0.59) | 0.78 (0.79) | 1.13 (1.13) | 2.13 (2.14) |
| 7 | TT | 0.54 (0.53) | 0.56 (0.56) | 0.52 (0.51) | 0.55 (0.55) | 0.74 (0.75) | 1.08 (1.08) | 2.06 (2.07) |
| 8 | TC | 0.50 (0.49) | 0.53 (0.53) | 0.46 (0.46) | 0.51 (0.51) | 0.71 (0.71) | 1.05 (1.04) | 2.08 (2.05) |
| 9 | CG | 0.42 (0.38) | 0.49 (0.43) | 0.45 (0.40) | 0.47 (0.42) | 0.69 (0.61) | 1.04 (0.91) | 2.09 (1.85) |
| 10 | GC | 0.32 (0.30) | 0.38 (0.36) | 0.37 (0.36) | 0.38 (0.36) | 0.58 (0.55) | 0.92 (0.88) | 1.94 (1.86) |
| 11 | CG | 0.24 (0.25) | 0.30 (0.30) | 0.29 (0.28) | 0.30 (0.29) | 0.48 (0.46) | 0.82 (0.78) | 1.85 (1.81) |

**Table S10**. Concentration in Molarity (and ion population) of K+ ions outside the minor groove of DDD.

| | | J&J neutrality | B&R neutrality | J&C neutrality | S&D neutrality | S&D 0.15M | S&D 0.5M | S&D 2.0M |
|---|---|---|---|---|---|---|---|---|
| 1 | CG | 0.30 (0.14) | 0.37 (0.16) | 0.80 (0.10) | 0.35 (0.16) | 0.45 (0.24) | 0.88 (0.39) | 1.77 (0.82) |
| 2 | GC | 0.49 (0.22) | 0.54 (0.24) | 0.31 (0.15) | 0.53 (0.23) | 0.72 (0.32) | 1.07 (0.47) | 1.94 (0.87) |
| 3 | CG | 0.67 (0.27) | 0.69 (0.27) | 0.51 (0.23) | 0.67 (0.27) | 0.90 (0.35) | 1.20 (0.47) | 2.07 (0.81) |
| 4 | GA | 0.81 (0.34) | 0.85 (0.36) | 0.80 (0.38) | 0.81 (0.34) | 1.11 (0.44) | 1.38 (0.58) | 2.32 (0.99) |
| 5 | AA | 0.97 (0.39) | 1.10 (0.44) | 0.92 (0.39) | 1.04 (0.42) | 1.36 (0.51) | 1.67 (0.66) | 2.66 (1.06) |
| 6 | AT | 1.01 (0.42) | 1.14 (0.46) | 0.93 (0.37) | 1.09 (0.44) | 1.29 (0.50) | 1.69 (0.68) | 2.71 (1.08) |
| 7 | TT | 0.98 (0.40) | 1.02 (0.41) | 0.95 (0.38) | 1.05 (0.42) | 1.25 (0.50) | 1.64 (0.66) | 2.67 (1.06) |
| 8 | TC | 0.80 (0.34) | 0.79 (0.33) | 0.79 (0.34) | 0.82 (0.35) | 1.05 (0.45) | 1.39 (0.59) | 2.32 (0.98) |
| 9 | CG | 0.66 (0.27) | 0.63 (0.25) | 0.61 (0.23) | 0.66 (0.26) | 0.89 (0.35) | 1.22 (0.48) | 2.10 (0.83) |
| 10 | GC | 0.50 (0.22) | 0.52 (0.23) | 0.50 (0.22) | 0.54 (0.24) | 0.76 (0.34) | 1.08 (0.48) | 1.96 (0.87) |
| 11 | CG | 0.32 (0.16) | 0.35 (0.16) | 0.34 (0.15) | 0.36 (0.17) | 0.56 (0.26) | 0.87 (0.40) | 1.79 (0.84) |

**Table S11.** Phase angle substates sampled by parmbsc1, reported in percentages, in comparison to parmbsc0 and the experimental conformational space.[a]

| base | source | C3'n | C4'x | O1'n | C1'x | C2'n | C3'x | C4'n |
|---|---|---|---|---|---|---|---|---|
| G2 | parmbsc1 | 0 | 0 | 0 | 17 | 64 | 19 | 0 |
| | parmbsc0 | 0 | 0 | 0 | 21 | 73 | 6 | 0 |
| | Xray/NMR | 2 | 0 | 0 | 12 | 83 | 3 | 0 |
| C3 | parmbsc1 | 2 | 4 | 5 | 27 | 56 | 7 | 0 |
| | parmbsc0 | 1 | 3 | 11 | 39 | 41 | 4 | 0 |
| | Xray/NMR | 4 | 38 | 13 | 28 | 14 | 3 | 0 |
| G4 | parmbsc1 | 0 | 0 | 0 | 25 | 63 | 11 | 0 |
| | parmbsc0 | 1 | 1 | 3 | 40 | 47 | 8 | 0 |
| | Xray/NMR | 5 | 0 | 0 | 1 | 88 | 5 | 0 |
| A5 | parmbsc1 | 3 | 1 | 1 | 14 | 63 | 17 | 0 |
| | parmbsc0 | 3 | 2 | 8 | 26 | 47 | 15 | 0 |
| | Xray/NMR | 2 | 0 | 1 | 11 | 84 | 4 | 0 |
| A6 | parmbsc1 | 1 | 1 | 2 | 24 | 57 | 14 | 0 |
| | parmbsc0 | 1 | 1 | 11 | 45 | 36 | 6 | 0 |
| | Xray/NMR | 0 | 0 | 0 | 46 | 49 | 4 | 0 |
| T7 | parmbsc1 | 1 | 2 | 4 | 46 | 46 | 2 | 0 |
| | parmbsc0 | 1 | 3 | 18 | 57 | 21 | 1 | 0 |
| | Xray/NMR | 0 | 0 | 13 | 73 | 10 | 4 | 0 |
| T8 | parmbsc1 | 1 | 3 | 5 | 39 | 51 | 2 | 0 |
| | parmbsc0 | 0 | 3 | 16 | 53 | 27 | 1 | 0 |
| | Xray/NMR | 0 | 0 | 15 | 59 | 19 | 6 | 0 |
| C9 | parmbsc1 | 3 | 3 | 4 | 36 | 49 | 3 | 0 |
| | parmbsc0 | 3 | 6 | 18 | 44 | 29 | 1 | 0 |
| | Xray/NMR | 0 | 0 | 5 | 37 | 55 | 3 | 0 |
| G10 | parmbsc1 | 0 | 0 | 0 | 20 | 59 | 20 | 0 |
| | parmbsc0 | 1 | 1 | 7 | 33 | 44 | 14 | 0 |
| | Xray/NMR | 0 | 0 | 0 | 13 | 85 | 2 | 0 |
| C11 | parmbsc1 | 1 | 1 | 2 | 25 | 66 | 5 | 0 |
| | parmbsc0 | 0 | 0 | 1 | 37 | 59 | 2 | 0 |
| | Xray/NMR | 0 | 0 | 1 | 18 | 69 | 10 | 2 |

[a] O1'x, C1'n and C2'x were not included in the table since a value of 0 was found for all the bases and for both force-fields and the experimental conformational space.

**Table S12.** α/γ substates sampled by parmbsc1, reported in percentages, in comparison to parmbsc0 and the experimental conformational space.

| base | source | g-/g- | g-/t | g-/g+ | t/g- | t/t | t/g+ | g+/g- | g+/t | g+/g+ |
|---|---|---|---|---|---|---|---|---|---|---|
| G2 | parmbsc1 | 0 | 0 | 97 | 0 | 0 | 2 | 0 | 1 | 0 |
| | parmbsc0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Xray/NMR | 0 | 0 | 94 | 0 | 0 | 3 | 3 | 0 | 0 |
| C3 | parmbsc1 | 0 | 10 | 90 | 0 | 0 | 0 | 0 | 0 | 0 |
| | parmbsc0 | 0 | 3 | 96 | 0 | 0 | 0 | 0 | 1 | 0 |
| | Xray/NMR | 0 | 0 | 94 | 0 | 2 | 0 | 4 | 0 | 0 |
| G4 | parmbsc1 | 0 | 1 | 96 | 0 | 0 | 1 | 0 | 2 | 0 |
| | parmbsc0 | 0 | 0 | 97 | 0 | 0 | 1 | 0 | 2 | 0 |
| | Xray/NMR | 0 | 0 | 98 | 0 | 0 | 2 | 0 | 0 | 0 |
| A5 | parmbsc1 | 0 | 2 | 98 | 0 | 0 | 0 | 0 | 0 | 0 |
| | parmbsc0 | 0 | 0 | 98 | 0 | 0 | 2 | 0 | 0 | 0 |
| | Xray/NMR | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 1 | 0 |
| A6 | parmbsc1 | 0 | 1 | 98 | 0 | 0 | 0 | 0 | 1 | 0 |
| | parmbsc0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Xray/NMR | 0 | 0 | 97 | 0 | 0 | 0 | 3 | 0 | 0 |
| T7 | parmbsc1 | 0 | 1 | 99 | 0 | 0 | 0 | 0 | 0 | 0 |
| | parmbsc0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Xray/NMR | 1 | 0 | 94 | 0 | 0 | 0 | 5 | 0 | 0 |
| T8 | parmbsc1 | 0 | 2 | 96 | 0 | 2 | 0 | 0 | 0 | 0 |
| | parmbsc0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Xray/NMR | 0 | 0 | 90 | 0 | 2 | 0 | 5 | 0 | 3 |
| C9 | parmbsc1 | 0 | 4 | 96 | 0 | 0 | 0 | 0 | 0 | 0 |
| | parmbsc0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Xray/NMR | 0 | 0 | 96 | 0 | 0 | 0 | 3 | 1 | 0 |
| G10 | parmbsc1 | 0 | 0 | 98 | 0 | 0 | 1 | 0 | 1 | 0 |
| | parmbsc0 | 0 | 0 | 99 | 0 | 0 | 1 | 0 | 0 | 0 |
| | Xray/NMR | 2 | 0 | 98 | 0 | 0 | 0 | 0 | 0 | 0 |
| C11 | parmbsc1 | 0 | 3 | 97 | 0 | 0 | 0 | 0 | 0 | 0 |
| | parmbsc0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Xray/NMR | 0 | 0 | 90 | 0 | 2 | 6 | 2 | 0 | 0 |

**Table S13.** Twist weighted averages and BIC components for the C3pG4 and C9pG10 steps.[a]

| | 1st component | | | 2nd component | | | Weighted Average |
|---|---|---|---|---|---|---|---|
| Systems | Avg | SD | Weight | Avg | SD | Weight | |
| J&C, Na+, neutral, SPCE | 24.5 | 4.6 | 0.19 | 37.7 | 4.6 | 0.81 | 35.2 |
| S&D, Na+, neutral, SPCE | 29.8 | 7.2 | 0.44 | 37.8 | 4.0 | 0.56 | 34.3 |
| J&C, Na+, neutral, TIP3P | 23.9 | 4.5 | 0.29 | 36.6 | 4.4 | 0.71 | 32.9 |
| S&D, Na+Cl-, 0.15M, TIP3P | 24.9 | 4.8 | 0.32 | 37.7 | 4.3 | 0.68 | 33.6 |
| S&D, Na+Cl-, 0.5M, TIP3P | 24.8 | 4.6 | 0.36 | 36.9 | 4.4 | 0.64 | 32.6 |
| S&D, Na+Cl-, 2.0M, TIP3P | 25.1 | 4.9 | 0.30 | 37.3 | 4.2 | 0.70 | 33.7 |
| J&C, K+, neutral, SPCE | 24.5 | 4.7 | 0.29 | 38.8 | 4.1 | 0.71 | 34.7 |
| S&D, K+, neutral, SPCE | 22.7 | 4.0 | 0.36 | 36.3 | 4.4 | 0.64 | 31.4 |
| S&D, K+Cl-, 0.15M, SPCE | 23.4 | 4.2 | 0.28 | 37.0 | 4.3 | 0.72 | 33.1 |
| S&D, K+Cl-, 0.5M, SPCE | 23 | 4.2 | 0.35 | 36.9 | 4.4 | 0.65 | 32.0 |
| S&D, K+Cl-, 2.0M, SPCE | 22.6 | 4.1 | 0.41 | 36.6 | 4.4 | 0.59 | 30.9 |
| S&D, K+, neutral, TIP3P | 23.6 | 4.4 | 0.32 | 36.9 | 4.5 | 0.68 | 32.7 |
| J&J, K+, neutral, SPCE | 23.5 | 4.2 | 0.27 | 36.5 | 4.7 | 0.73 | 33.0 |
| B&R, K+, neutral, SPCE | 22.8 | 4.0 | 0.30 | 36.5 | 4.4 | 0.70 | 32.4 |

[a] Values reported are averages between the two CpG steps present in the DDD sequence.

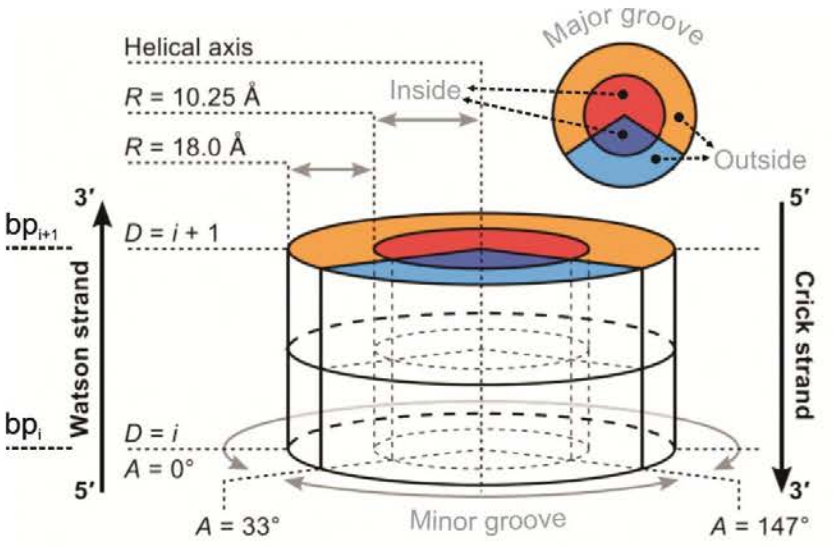



**Figure S1**. Schematic view of a base-pair step. Untwisted view of a base-pair step detailing the CHC space-partitioning scheme used to measure ion populations around DNA. See Methods and the recent work by Lavery and coworkers [Lavery R., Maddocks J.H., Pasi M. and Zakrzewska K. *Nucleic Acids Res*. **2014**, *42*, 8138–49.] for more details. Adapted from [Pasi M., Maddocks J.H. and Lavery,R. *Nucleic Acids Res.* **2015**, *43*, 2412–23].

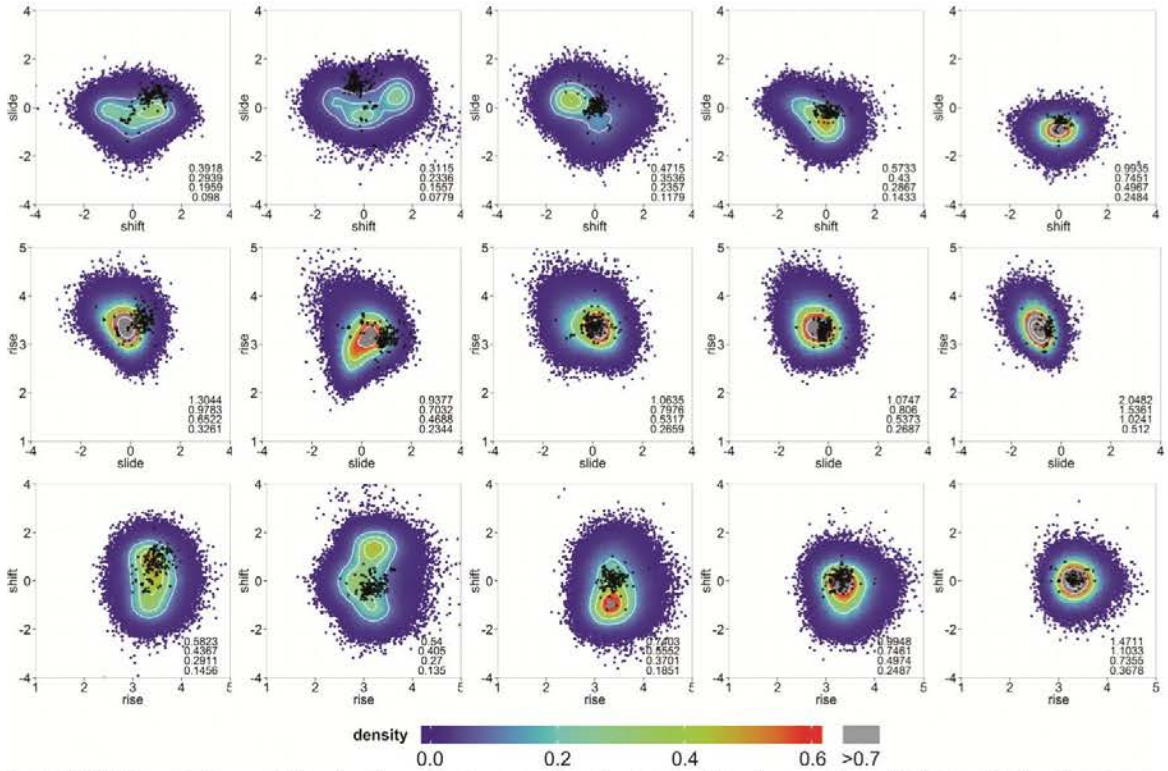



**Figure S2.** Comparison at the bps level between the theoretical and experimental translational spaces. Translational parameters are reported in angstroms. All distinct bps found in DDD are shown (removing the ends): GC (first column), CG (second column), GA (third column), AA (fourth column), and AT (fifth column). Smoothed 2D densities, estimated by fitting the observed distributions to a bivariated normal kernel (evaluated on a square grid of 90x90 bins), are depicted by coloring the points coming from the MD simulations with a color gradient from low (blue) to high (red) density. Iso-density curves are shown in white and are quantified on the bottom right side of each plot. Experimental conformations are shown as black dots (supplementary Table S1).

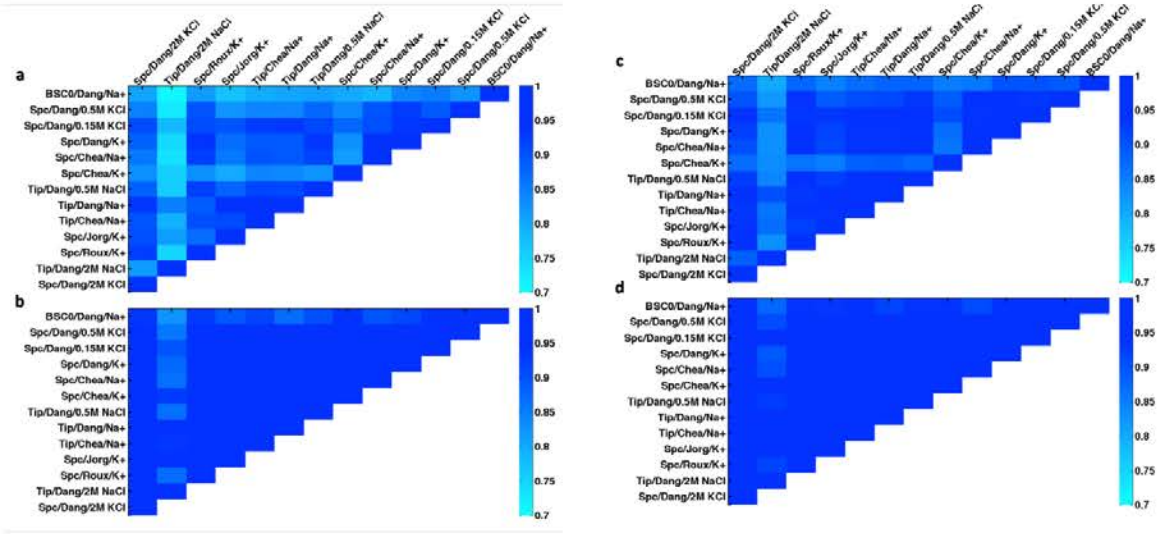
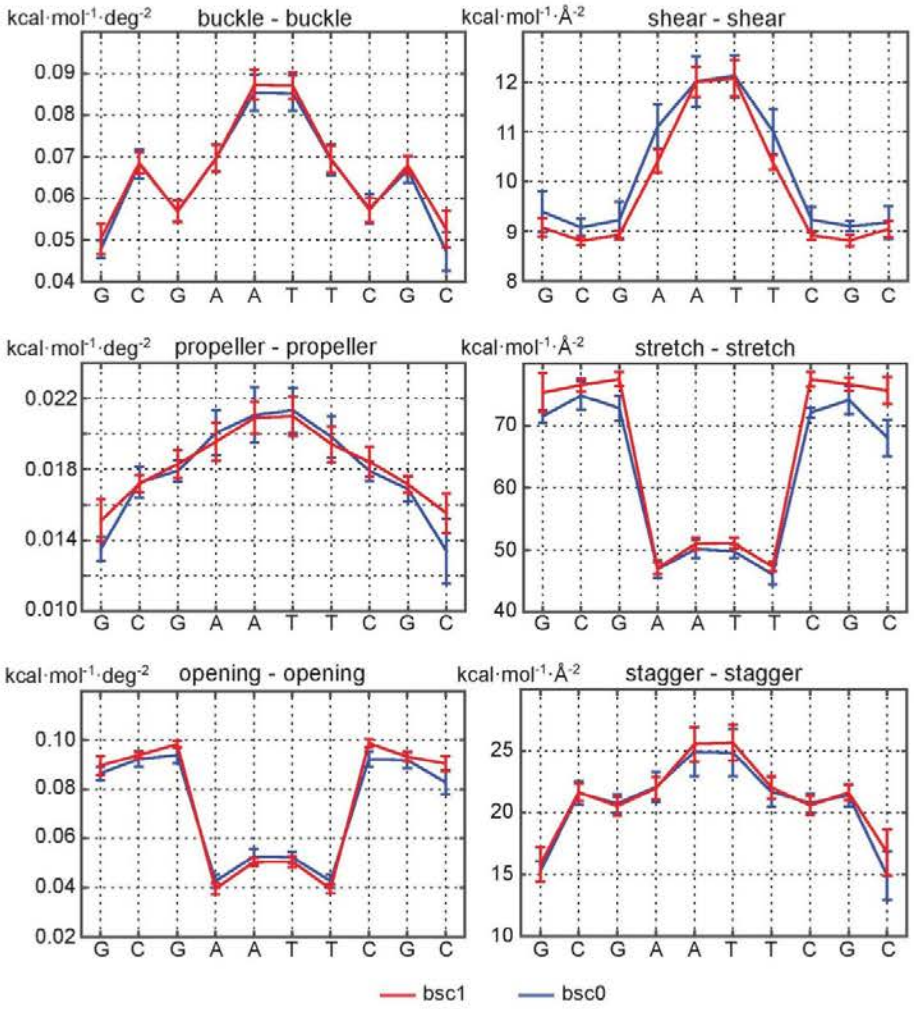



**Figure S3.** Similarities between the DNA dynamics from the simulations performed with parmbsc1 and the reference parmbsc0. Absolute (**a**), relative (**b**), energy-weighted (**c**), and relative energy-weighted (**d**) similarity index matrixes between trajectories of DDD in different environments. See the work of Pérez and coworkers for more details on similarity indexes calculation [A. Pérez, J. R. Blas, M. Rueda, J. M. López-Bes, X. de la Cruz, and M. Orozco. *J. Chem. Theory Comput.* **2005**, *1*, 790–800].




**Figures S4.** Force constants associated with the deformation of a single helical degree of freedom per base pair obtained from DDD trajectories in different environments with parmbsc1 and parmbsc0 force-fields. The rotational values are given in kcal/mol deg$^2$ and translational ones in kcal/mol Å$^2$. Only diagonal entries of the full non-local stiffness matrix are reported. The two meta-trajectories were built from all the bsc1 and bsc0 simulations described in Table 1, respectively. The error bars denote the standard deviation between the simulations within each force-field.

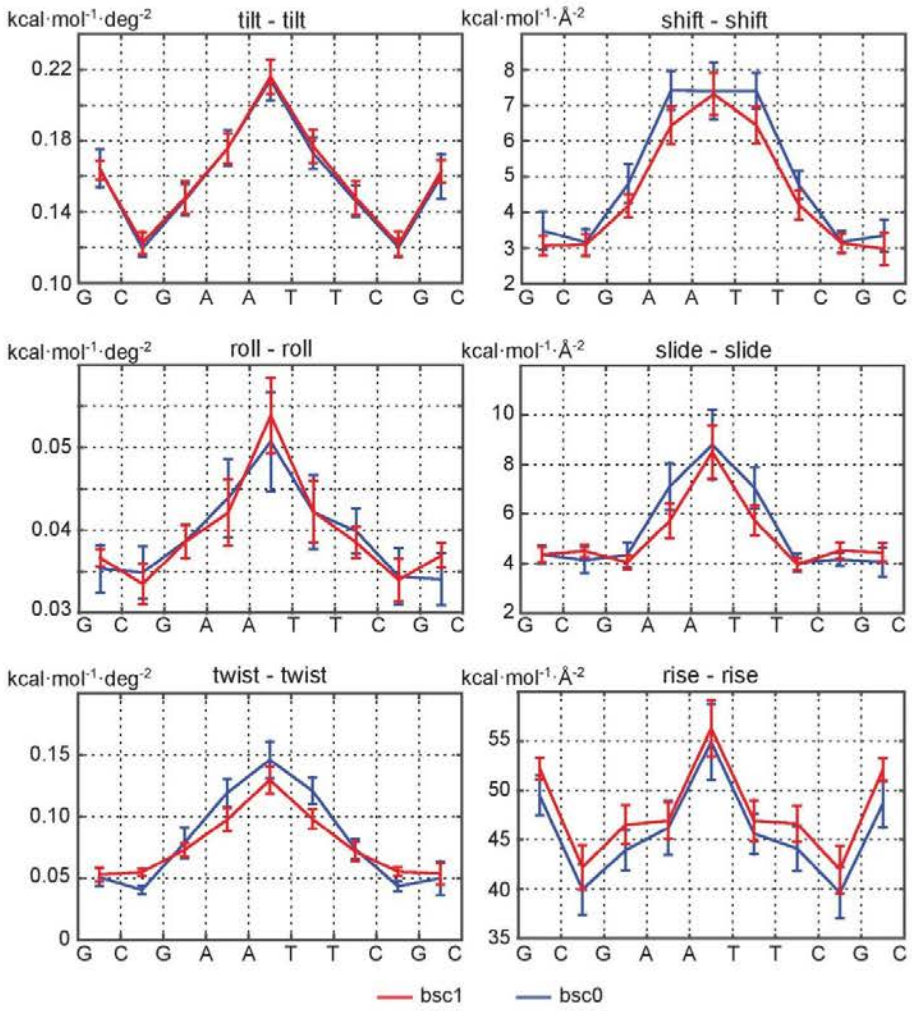



**Figure S5.** Force constants associated with the deformation of a single helical degree of freedom per base pair steps obtained from DDD trajectories in different environments with parmbsc1 and parmbsc0 force-fields. The rotational values are given in kcal/mol deg$^2$ and translational ones in kcal/mol Å$^2$. Only diagonal entries of the full non-local stiffness matrix are reported. The two meta-trajectories were built from all the bsc1 and bsc0 simulations described in Table 1, respectively. The error bars denote the standard deviation between the simulations within each force-field.

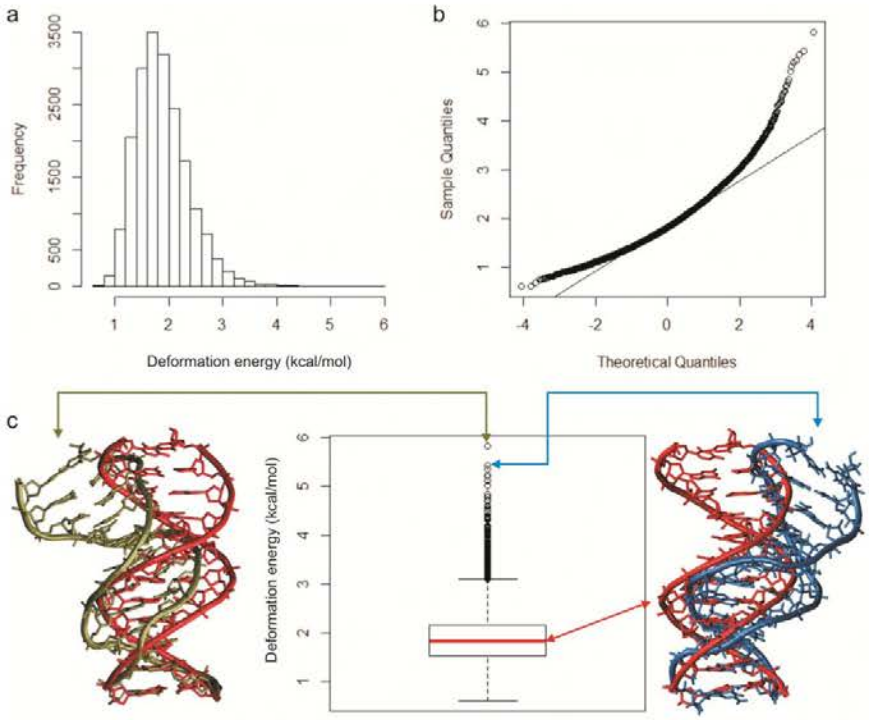



**Figure S6**. Graphical statistical analysis of the deformation energy distribution. a) Histogram of deformation energy for the complete set of trajectories. b) Q-Q plot comparing the distribution of quintiles from MD with a theoretical normal distribution (straight line). c) Boxplot showing the distribution in quartiles (median is depicted in red), and the outliers points noticeable above the last decile. The structure of the DNA in the two most extreme cases is overlapped with a representative structure of the median, to highlight the deformation of the double helix beyond the harmonic regime.

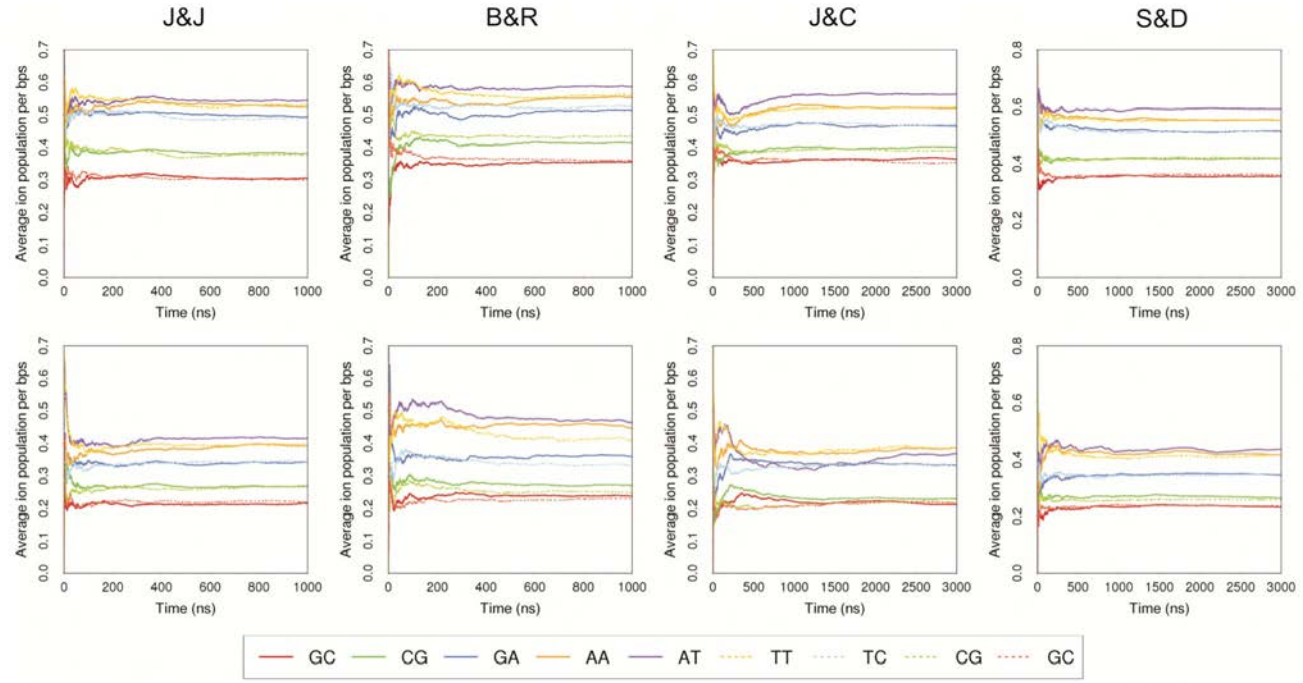



**Figure S7.** Time evolution of the average ion population per bps outside the major (first row) and minor (second row) grooves. The terminal bases were removed from the analysis. See Figure S4 and the Methods section for a detailed description of the space-partitioning scheme.

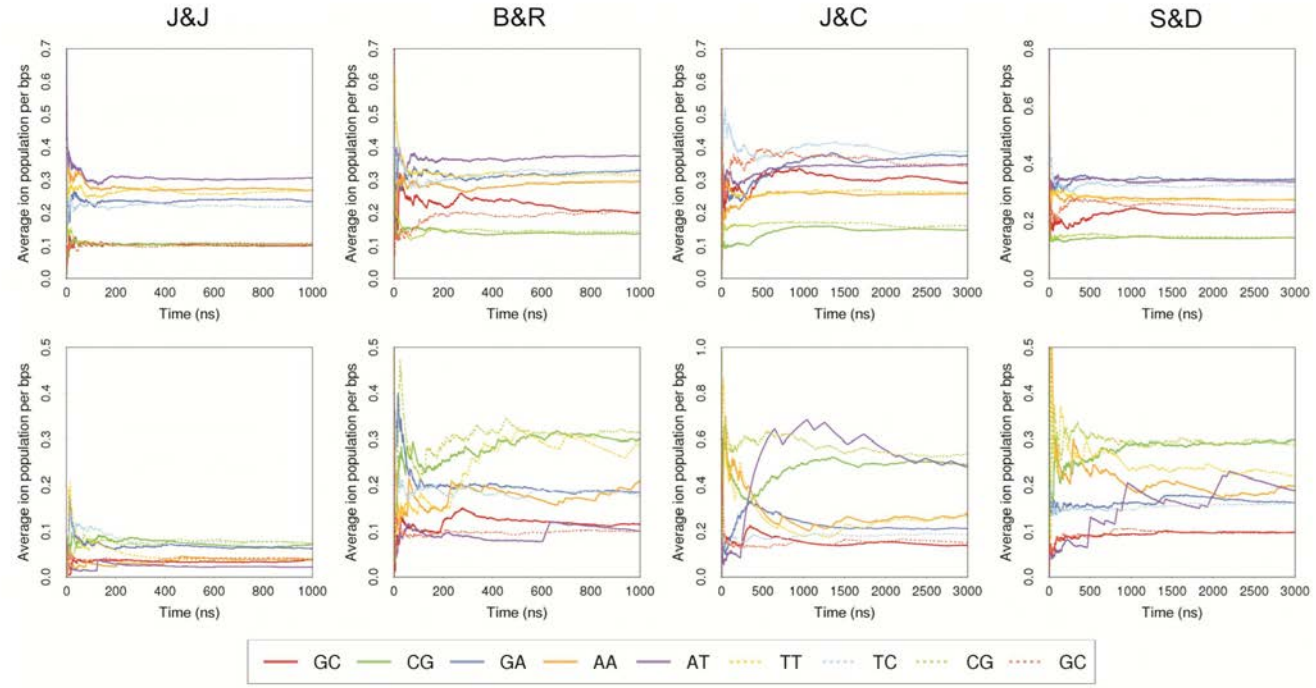



**Figure S8.** Time evolution of the average ion population per bps inside the major (first row) and minor (second row) grooves. The terminal bases were removed from the analysis. See Figure S4 and the Methods section for a detailed description of the space-partitioning scheme.

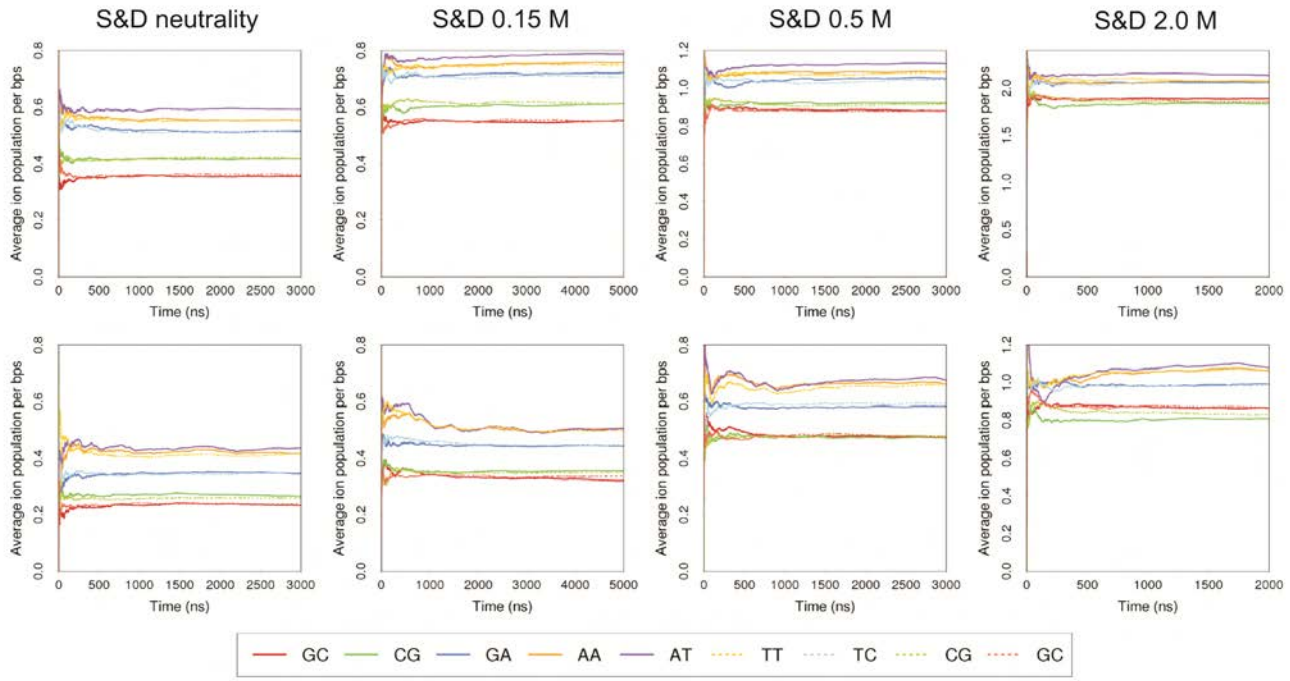



**Figure S9.** Time evolution of the average ion population per bps outside the major (first row) and minor (second row) grooves. The terminal bases were removed from the analysis. See Figure S4 and the Methods section for a detailed description of the space-partitioning scheme.

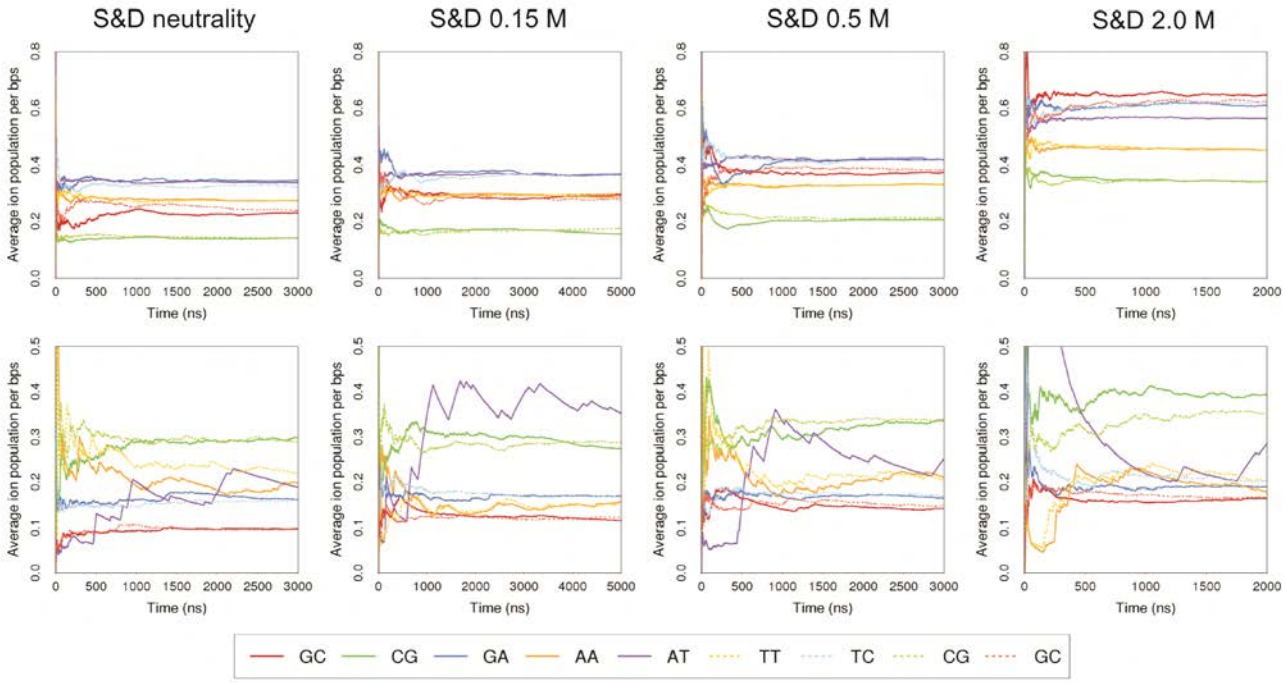



**Figure S10.** Time evolution of the average ion population per bps inside the major (first row) and minor (second row) grooves. The terminal bases were removed from the analysis. See Figure S4 and the Methods section for a detailed description of the space-partitioning scheme.

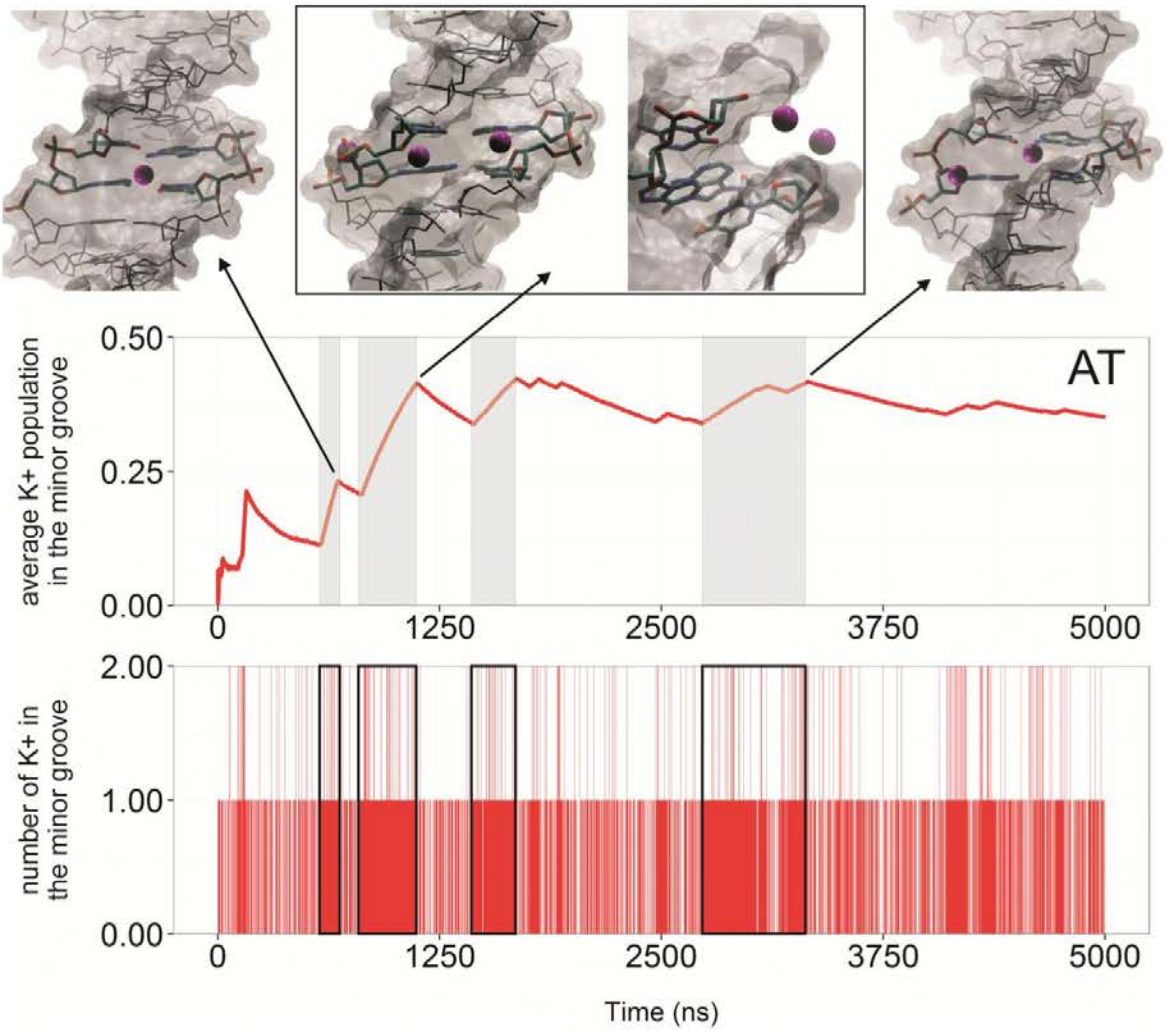



**Figure S11**. Time evolution of the $K^+$ dynamics in the minor groove of the AT bps. In the first plot we show the accumulative population average along time, evidencing with gray regions the saw tooth-like increases. As shown in the count of ions inside the groove (second plot), these frustration events are produced by very long residence times (highlighted with black boxes), with the concomitant entrance of two $K^+$ in the minor groove.

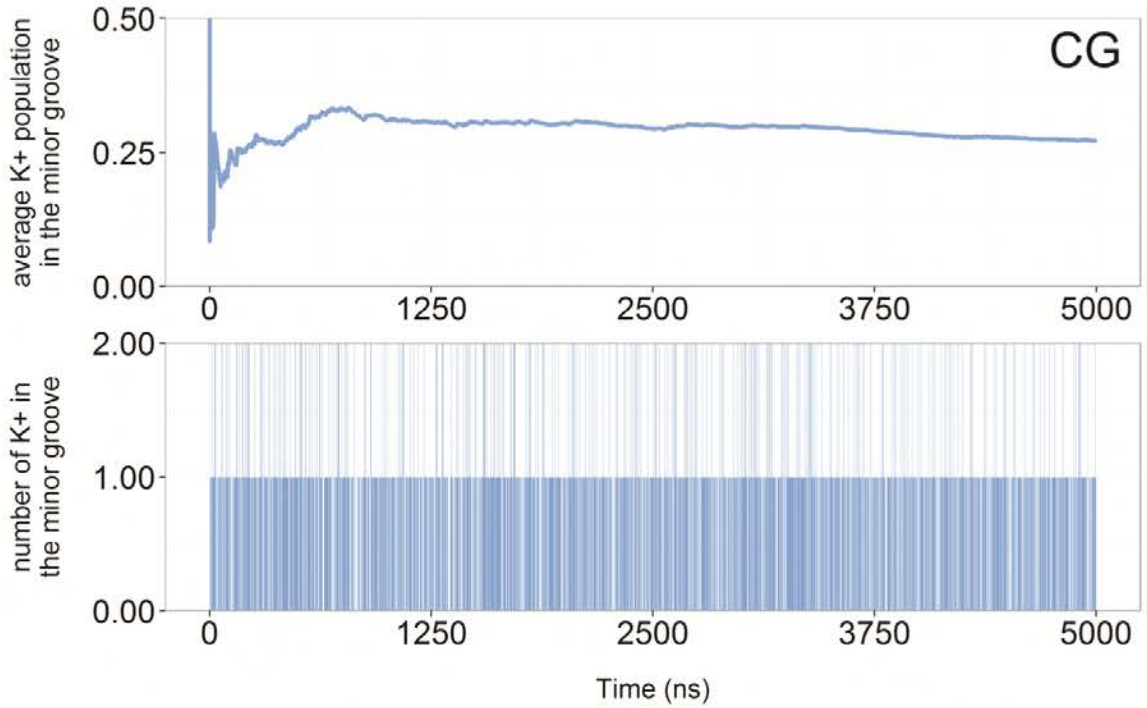



**Figure S12**. Time evolution of the $K^+$ dynamics in the minor groove of the CG bps. In the first plot we show the accumulative population average along time, evidencing smooth convergence. As shown in the count of ions inside the groove (second plot), the entrance and leaving of cations inside the minor groove is almost instantaneous, if compared with the AT bps (see Figure S11).

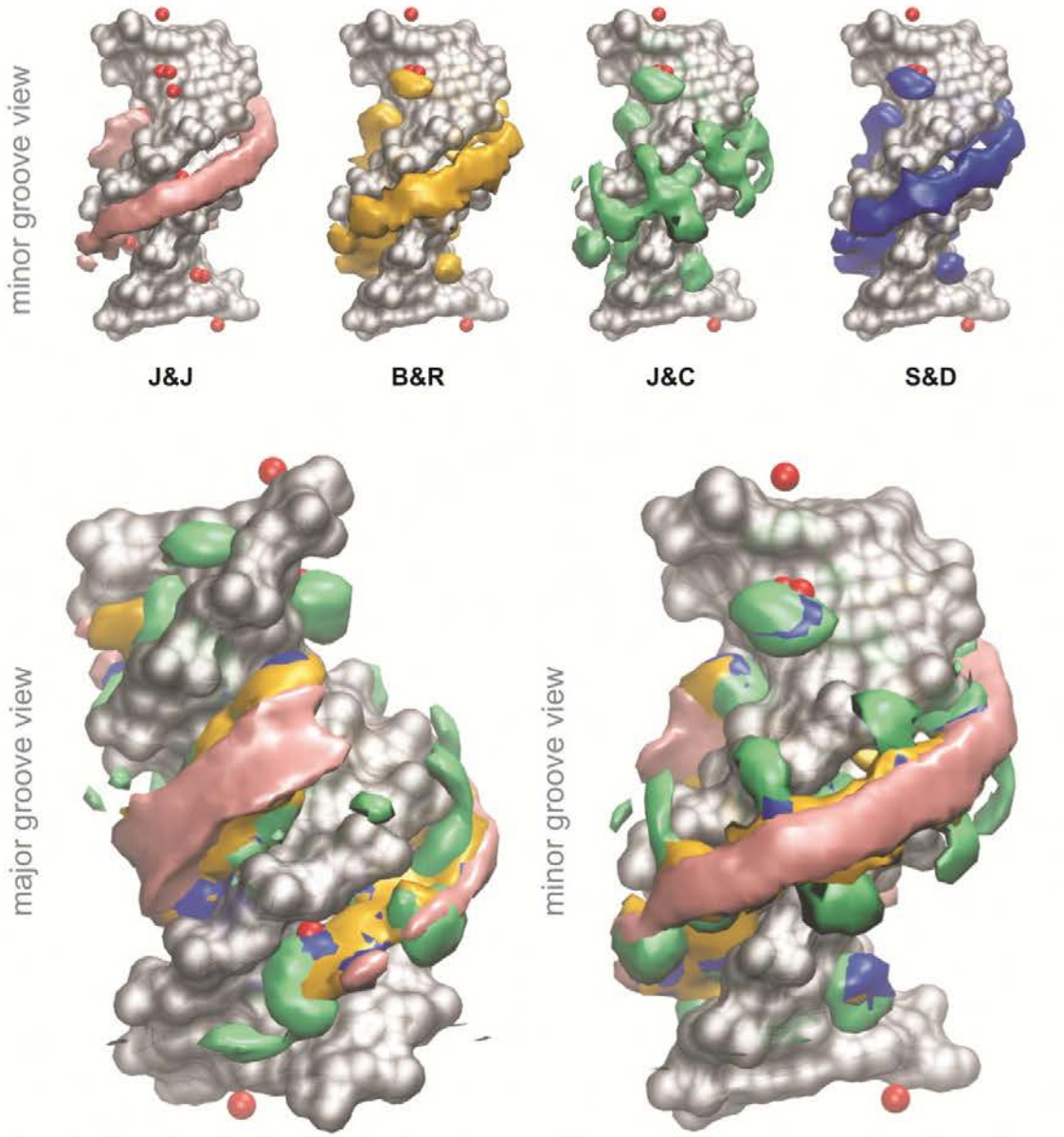



**Figure S13**. Potassium distributions along the helix. Cartesian K+ isomolarity surfaces at 1.5 M reconstructed from the CHC histograms with respect to the average structure (shown with a silver sheet). For comparison purposes, neutral systems have been overlapped with the Tl+ cations (red spheres) that co-crystallized with the DNA (PDB code 1JGR). Note that Thallium cations are used as a replacement of Potassium in diffraction experiments [Hud N. V and Engelhart A.E. Chapter 3 Sequence-specific DNA-Metal Ion Interactions. In *Nucleic Acid-Metal Ion Interactions* **2009**. The Royal Society of Chemistry, pp. 75–117].

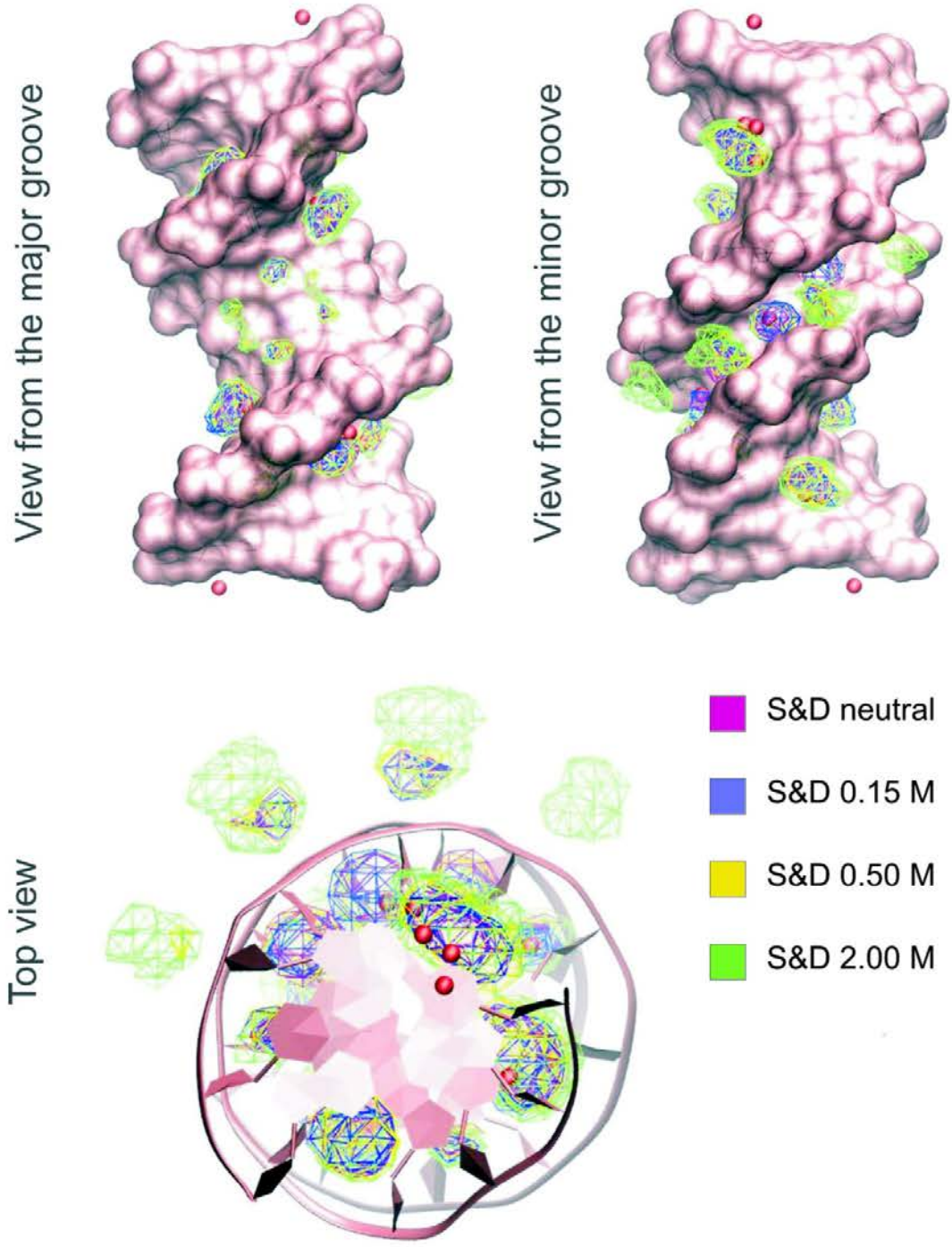



**Figure S14**. K+ distributions along the helix. Cartesian K+ isomolarity surfaces at 5.0 M reconstructed from the CHC histograms with respect to the average structure (shown with a pink sheet). For comparison purposes, densities coming from S&D simulations with increasing added salt, have been overlapped with the Tl+ cations (red spheres) that co-crystallized with the DNA (PDB code 1GJR).

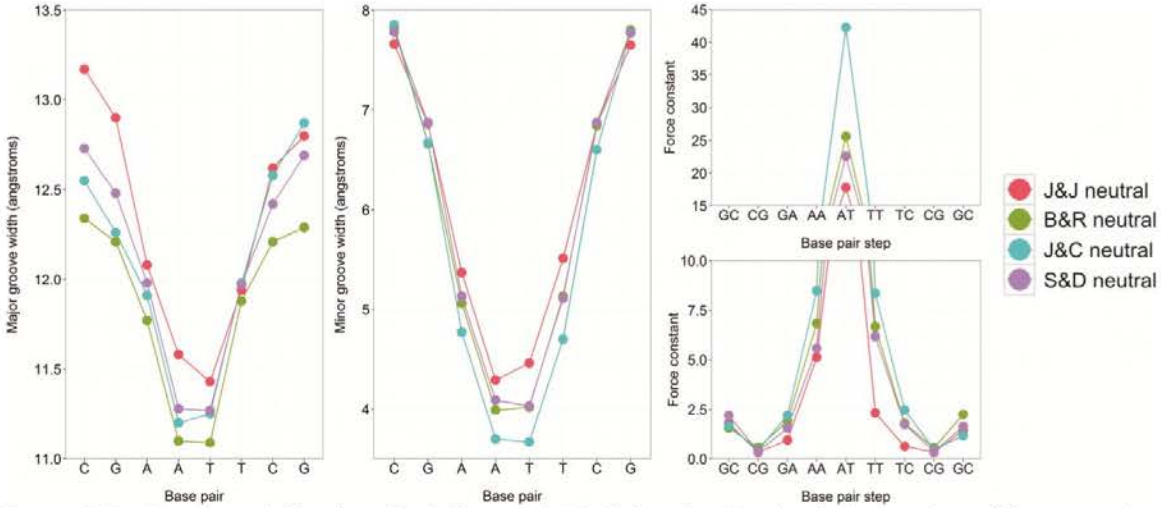



**Figure S15.** Grooves width, bps flexibility, and BI % for the K+ simulations using different cation models. From LEFT to RIGHT: Major groove width; minor groove width; and bps flexibility obtained by multiplying the diagonal entries of the 6x6 inter molecular stiffness matrices. Note that the simulation with J&C predicts the stiffest oligomer as previously shown (see Main text).

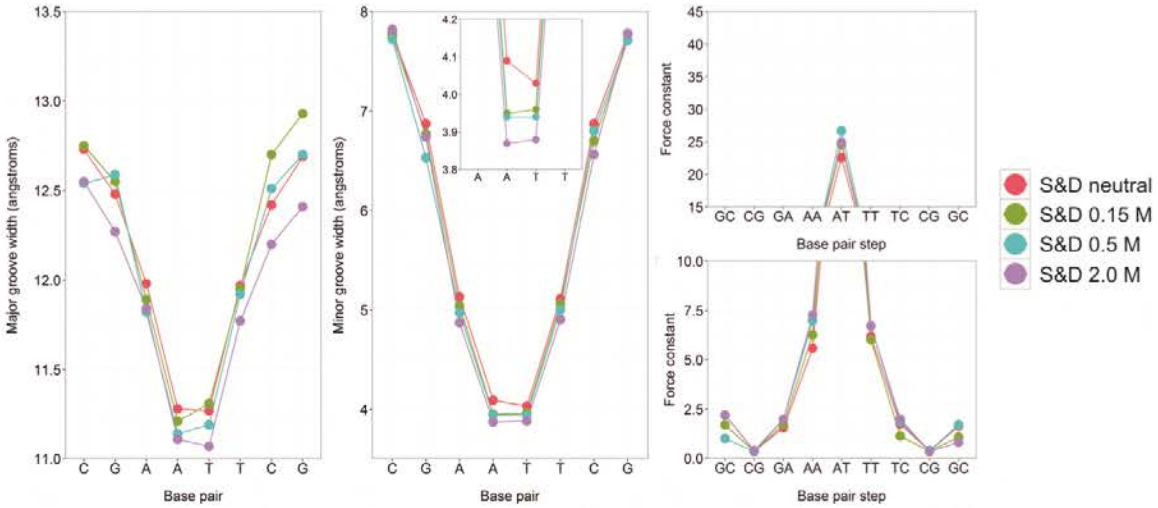



**Figure S16.** Grooves width, bps flexibility, and BI % for the K+ simulations using added salt. From LEFT to RIGHT: Major groove width; minor groove width; and bps flexibility obtained by multiplying the diagonal entries of the 6x6 inter molecular stiffness matrices. Note that the simulation with J&C predicts the stiffest oligomer as previously shown (seeMain text).

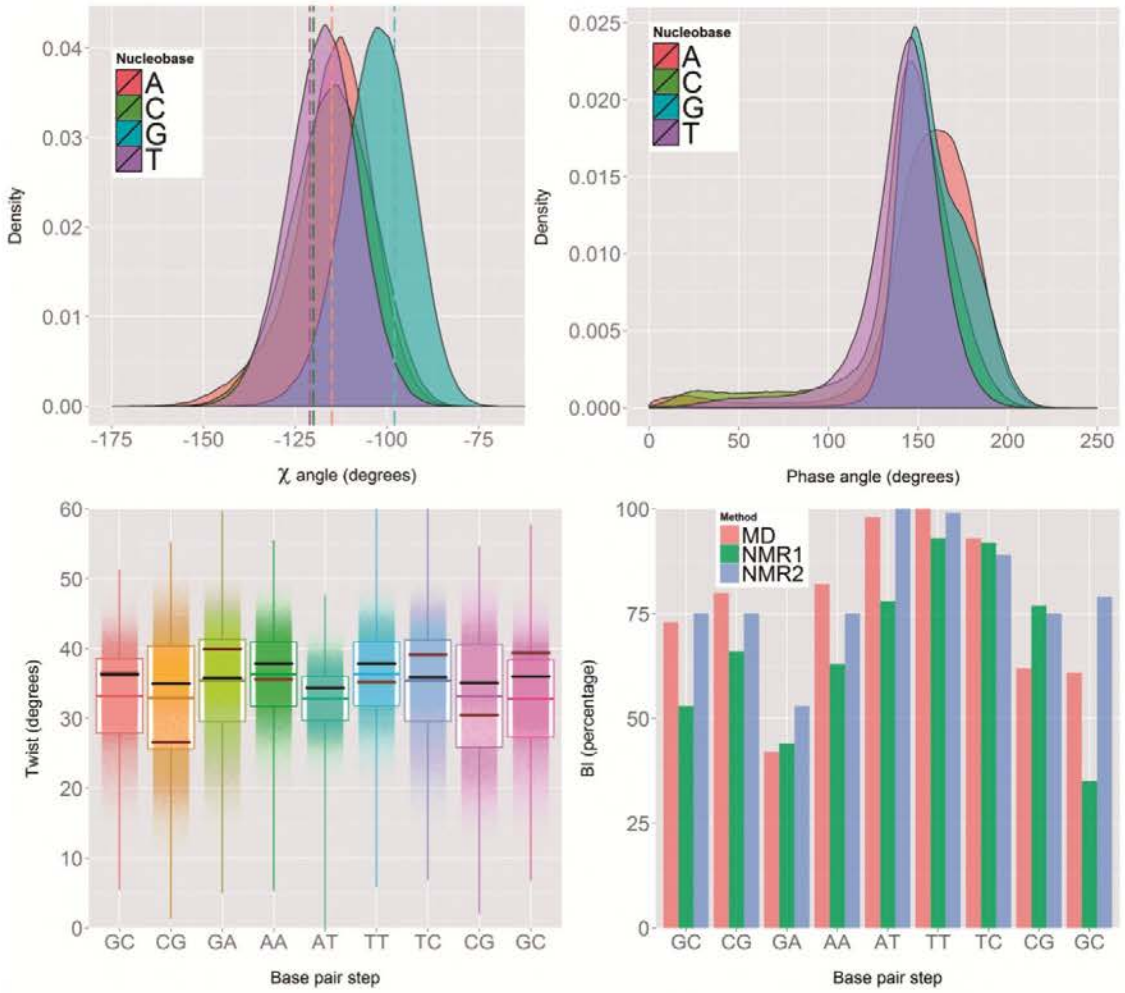



**Figure S17**. From LEFT to RIGHT, from TOP to BOTTOM: **i)** Distribution of the χ dihedral angle grouped by nucleobase (i.e. to obtain the distribution for Guanine (G) an ensemble with the instantaneous χ values along the 10 microseconds was created using the dynamics of G2, G4 and G10). The average experimental values (averaging the Xray and NMR structures with PDB codes: 1BNA, 2BNA, 7BNA, 9BNA, and 1NAJ), grouped in the same way, are shown in dashed lines. Note that only the Watson strand is shown (the Crick strand is totally symmetric). **ii)** Same than (i), for the Phase pseudo angle (see Table S3 to compare with experimental data). **iii)** Box plot showing the average, the standard deviation, and the max and min values of the twist helical parameter for the base pair steps number 2 to 10 in DDD. A one-dimensional representation of the twist distribution is also shown for each bps using the 10 microseconds of simulation. The averaged experimental values are shown with a dark red and a black line, for Xray and NMR respectively. **iv)** BI percentage for each bps, obtained by averaging the difference between ε and ζ angles at the 3'-junction of the Watson strand of each base pair. NMR1 and NMR2 are values obtained using phosphorus chemical shifts as detailed in the works of Tian *et al.*, and Schwieters *et al.* respectively (see Main text).

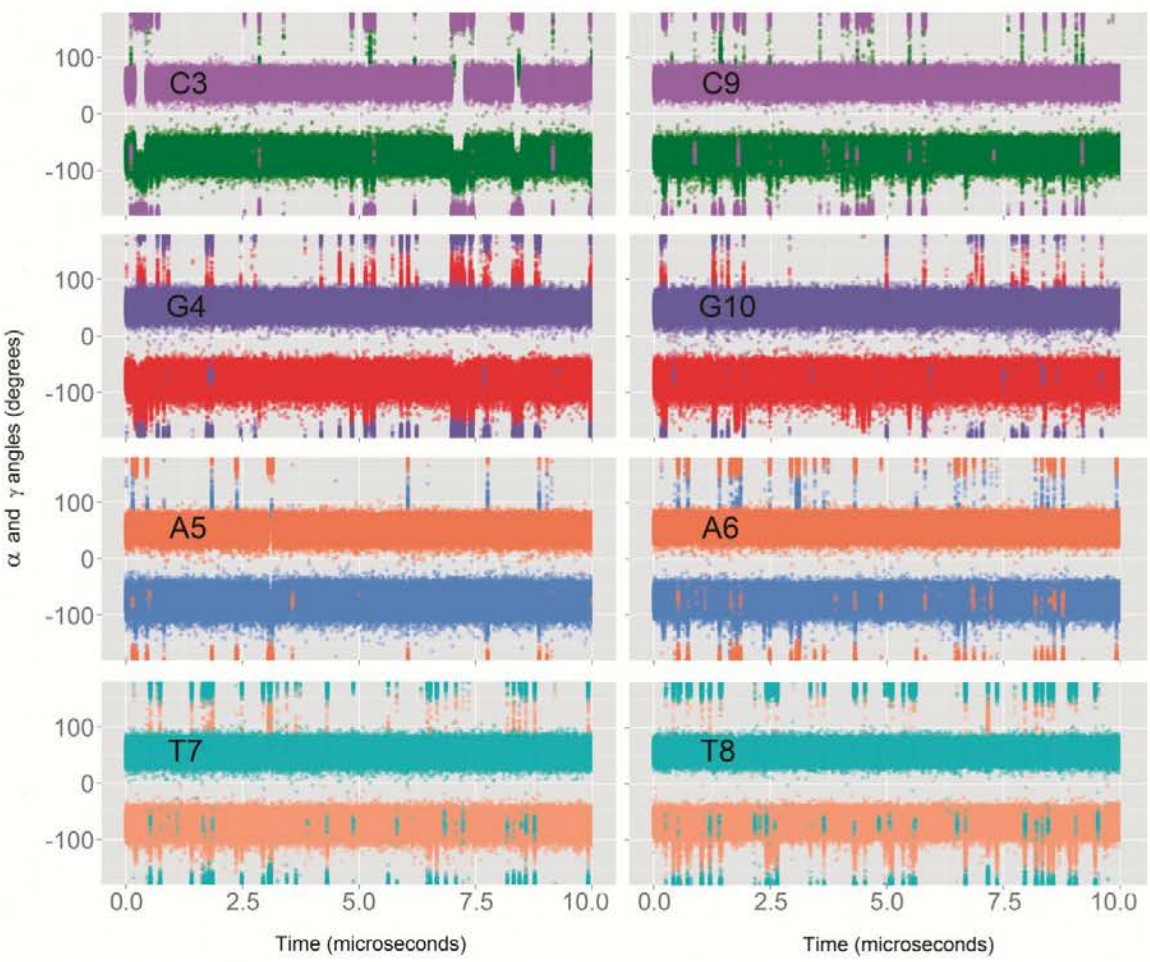



**Figure S18**. Time evolution of α and γ torsion angles in Drew-Dickerson dodecamer. From TOP to BOTTOM: α and γ angles are depicted in green and purple for cytosines, red and violet for guanines, light blue and orange for adenines, and light orange and cyan for thymines. Note that in all the cases both angles preferably explore the canonical B-DNA substate characterize by α in *gauche-* (g-) and γ in *gauche+* (g+).

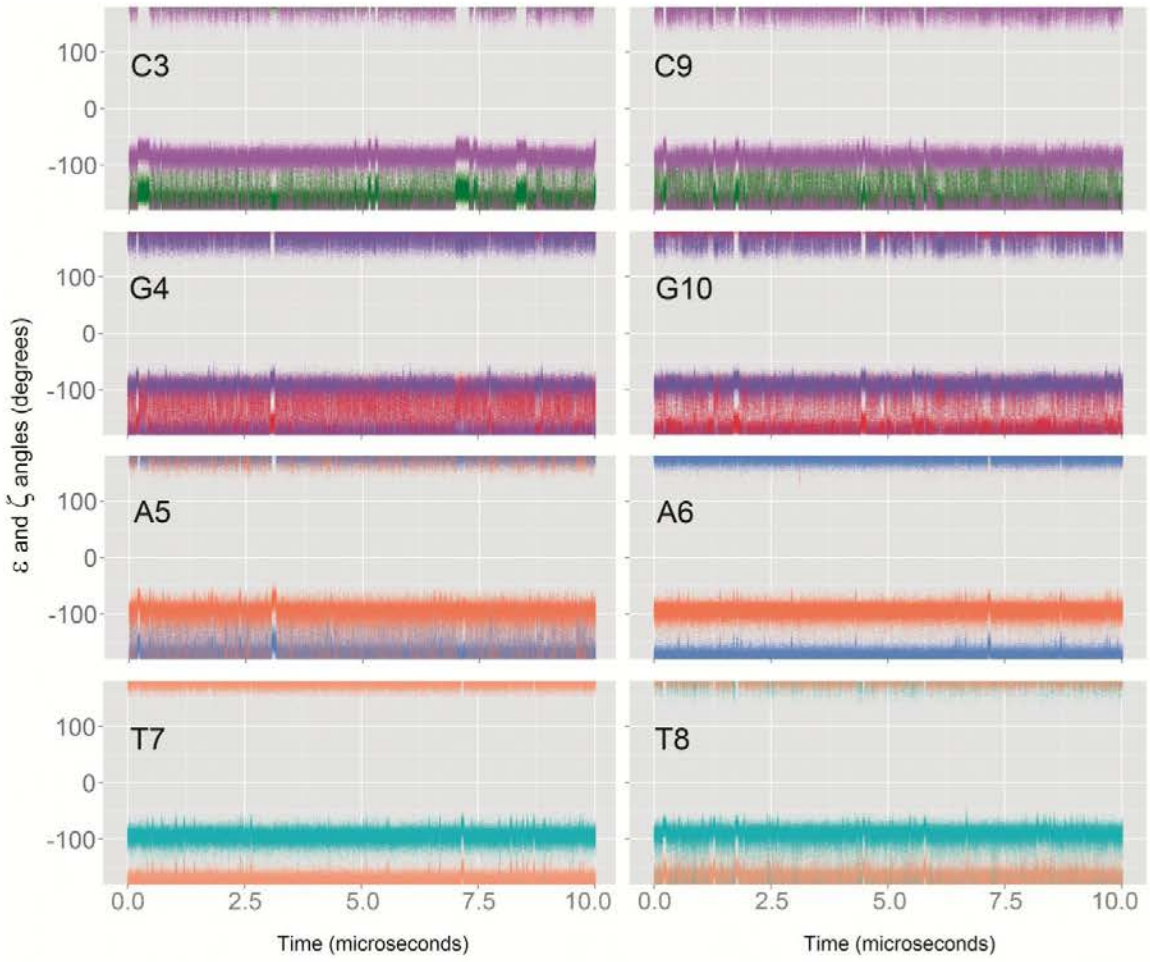



**Figure S19**. Time evolution of ε and ζ torsion angles in Drew-Dickerson dodecamer. From TOP to BOTTOM: ε and ζ angles are depicted in green and purple for cytosines, red and violet for guanines, light blue and orange for adenines, and light orange and cyan for thymines. Note that in all the cases both angles spend most of the time in the canonical B-DNA substate characterize by ε in *trans* (t) and ζ in *gauche-* (g-), also known as BI state. Cytosine and guanine show more propensity to BI → BII transitions, as evidenced by the concerted changes in ε and ζ from t→g- and from g-→t respectively.

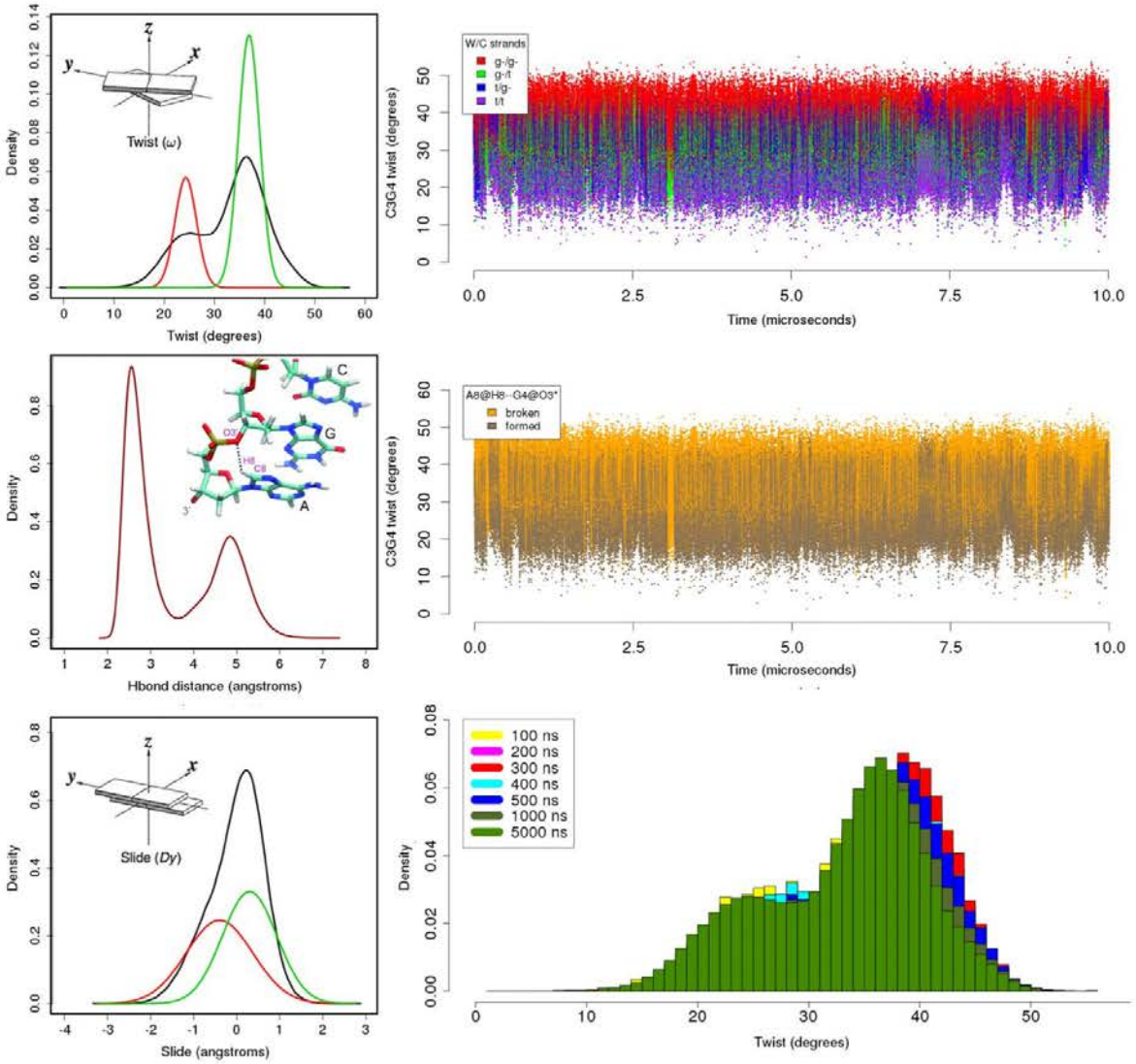



**Figure S20**. From LEFT to RIGHT, from TOP to BOTTOM: **i)** Twist distribution of the C3pG4 step (black) and the two components obtained with BIC ( ); LT component in red and HT component in green. **ii)** Coupling between the twist at the C3pG4 step of DDD and the ζ angle substates at the 3'-side of the bps. **iii)** Distribution of the C8H8-O3' hydrogen bond between G4 and A5. **iv)** Coupling between the twist at the C3pG4 step of DDD and the formation of the C8H8-O3' hbond. **v)** Slide distribution of the G4pA5 step (black) and the two components obtained with BIC; LS (low slide) component in red and HS component in green. **vi)** Twist distribution of the C3pG4 step at increasing time.

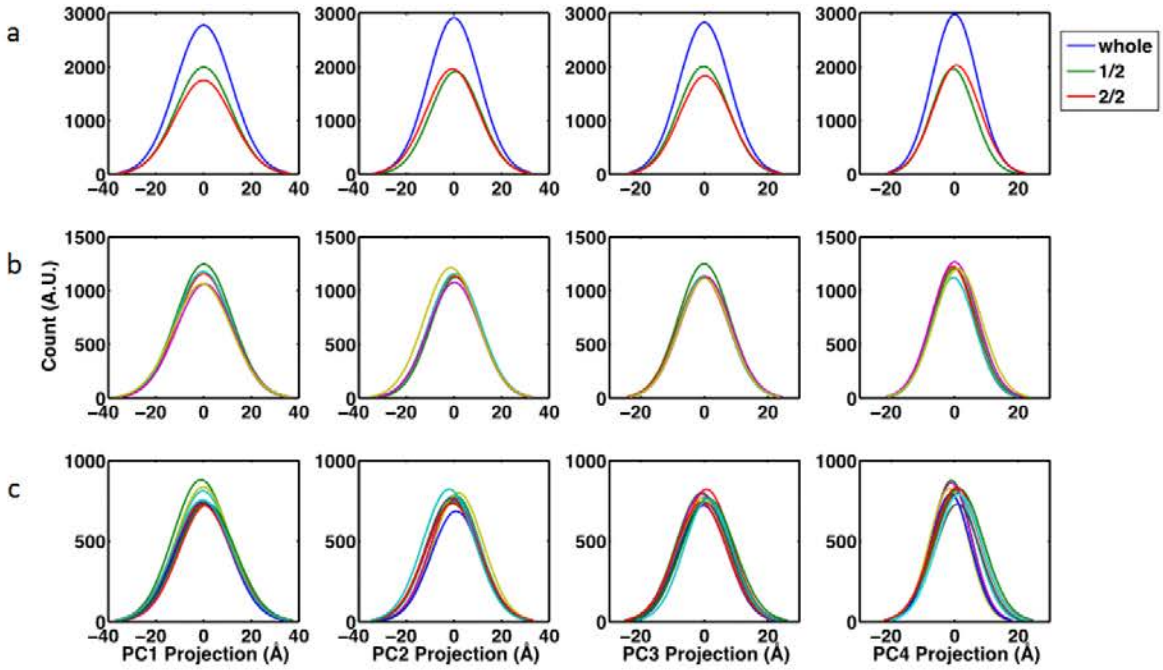
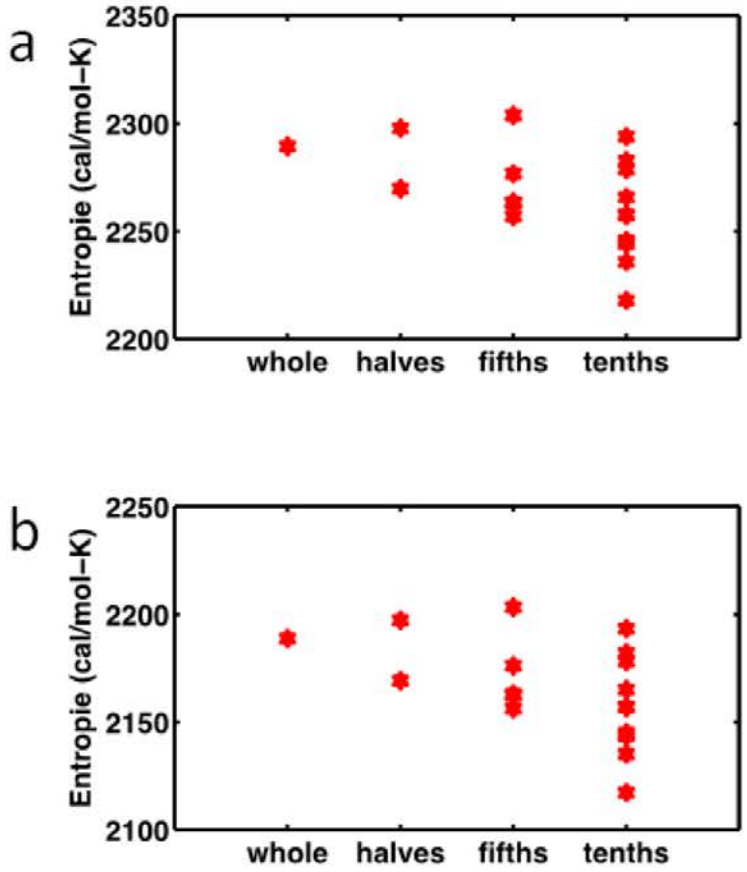



**Figure S21**. Principal Components histograms for first 4 principal components calculated for all bases excluding capping ones. PCA was done for segments of halves (a), fifths (b) and tenths (c) of the 10 μs trajectory, where each color represents different segments, while blue curve in (a) represents histograms of the whole trajectory. Analysis was done using CPPTRAJ software package (see Main text).




**Figures S22.** Entropy values calculated for whole and segments of halves, fifths and tenths of the 10 μs trajectory of DDD. Entropies were calculated using Schlitter's method (a) and Andricioaei-Karplus' method (b) [Schlitter J., *Chem. Phys. Lett.* **1993**, *215*, 617–621. Andricioaei I., and Karplus M., *J. Chem. Phys.* **2001**, *115*, 6289].

## 4.3 DNA force field "blind" benchmark (Publication 3)

During the development of parmbsc1 force field different modifications have been developed to alleviate the problems of parmbsc0, namely from the Czech's consortium, also known as the Olomouc group (OL family of force fields). OL1 (Zgarbová et al. 2013) patch tried to improve ε/ζ representation and the $B_I/B_{II}$ equilibrium in canonical B-DNA, while OL4 patch (Krepl et al. 2012) had an objective to correct χ conformation for DNA, in order to better represent unusual forms of DNA, such as Z-DNA and quadruplexes. An interesting approach was applied in RNA simulation world, where Chen and Garcia scaled down van der Waals interactions (besides correcting χ) of parmbsc0 force field, in order to fold some known RNA hairpin motifs (Chen & García 2013). Following the assumption that DNA force fields generally produce over-stacking, some authors took the same approach implementing same correction for DNA simulation in order to study free energy of WC to Hoogsteen pairing transition (Yang et al. 2015). These and other tailor-made modifications have allowed the study of some exotic forms of DNA in the multi-nanosecond regime, but have produced also a notable confusion in the field, since they are not additive and it is unclear when they should be used. After the publication of parmbsc1 force field, Jurečka's group published their latest version of OL force field, called OL15, which included previous OL1 and OL4 corrections with an additional correction of the β torsion (Zgarbová et al. 2015). Meanwhile, CHARMM family of force fields had seen recent advances (Hart et al. 2011) implemented into the universal CHARMM36 force field, as well as the advances in the all atom polarizable force field (Savelyev & MacKerell 2014).

Recent publication from the group of Thomas Cheatham assessed the current state of DNA AMBER force fields, where they compared the performance of parmbsc1 and OL15 force fields. Their study included 5 sequences containing PDB entries: DDD (1BNA/1NAJ), 2 poly-A tracts (1FZX and 1SK5), a duplex in sub-Å resolution (3GGI), and a small Z-DNA duplex (1I0T). Besides 1NAJ structure, all the other structures used for comparison are crystallographic ones. Additionally, to test the convergence of the two force fields, they simulated DDD using 100 independent MD simulations, each extended to 10 μs, concatenating them into a 1 ms long trajectory for the two force fields. Their results conclude that for DDD both force fields yield a sub-1 Å agreement with the average NMR structure, and no deviations in the ms timescale.

In the parmbsc1 publication (see Chapter 4.1) we showed a small benchmark of force fields for DDD sequence (note that OL15 appeared more than 1 year after parmbsc1 was developed, so it was not considered in our original benchmarking). We decided to thoroughly test all the recent important corrections of parmbsc0, together with CHARMM general and polarized force fields, comparing the results with *de novo* NMR experiments for 3 B-DNA duplex sequences: SEQ1:

d[CGCGCAATTCGCG]$_2$ (DDD), SEQ2: d[GCTAGCGAGTCC]$_2$ and SEQ3: d[GGAGACCAGAGG]$_2$. The first sequence was selected as a control to determine the reliability of the experimental models derived from NMR data, while the other two were selected are real "blind" tests, as their structures were unknown prior to our NMR study. We did 2 μs simulations of each sequence using 8 different force fields, *parmbsc0*, *parmbsc1*, *parmbsc0_OL1*, *parmbsc0_OL1_OL4*, *parmbsc0_OL15*, *parmbsc0_ChenGarcia*, *CHARMM36* (C36) and the polarizable *CHARMM36_Drude* (C36dip) and the same environmental conditions.

Our results demonstrate that parmbsc1 provides the best fit to various experimental data. (see Tables 1-5 in the following publication) for the three oligos considered, including those for which experimental data is presented here for the first time. Parmbsc1 allows us to predict wide-angle scattering data, which was not considered at all at the stage of force field development. These findings strongly support the lack of "overtraining" artefacts related to the refinement of the force field. Close in accuracy to parmbsc1 is the latest force field OL15, while the other parmbsc0- based force field behave reasonably well, except the Chen-Garcia one, which provide poor results. CHARMM pair additive force field provides poorer results than those derived from amber-family of force fields, while the polarized version leads to structural corruption.

Parmbsc1 and OL15 force fields are also best in reproducing helical averages and profiles (see Figures 2 and 4, and Supplementary Tables S1-S6 in the following publication), while from the analysis of other AMBER force fields, we see that OL1 still experience the excessive terminal fraying problem of the parmbsc0, which is absent with the addition of OL4 correction (see Table 6 in the following publication). Major drawback of OL1+OL4 combination is the underestimation of twist (in average 2º less than the parmbsc1 average). "DNA-adopted" Chen-Garcia correction produces highly undertwisted structures with almost no sequence dependency in the helical profiles. Simulations with CHARMM36 force field showed big deviations in the terminal base pairs for all sequences, which affected the rest of the structure and its helical parameter profiles (except for SEQ3 where the results are comparable with AMBER family of force fields). Polarizable force field C36dip showed strong deviations from the canonical form of B-DNA duplex yielding "ladder-like" structures after, in average, 50 ns of the simulation, thus it was excluded from most of the helical analysis.

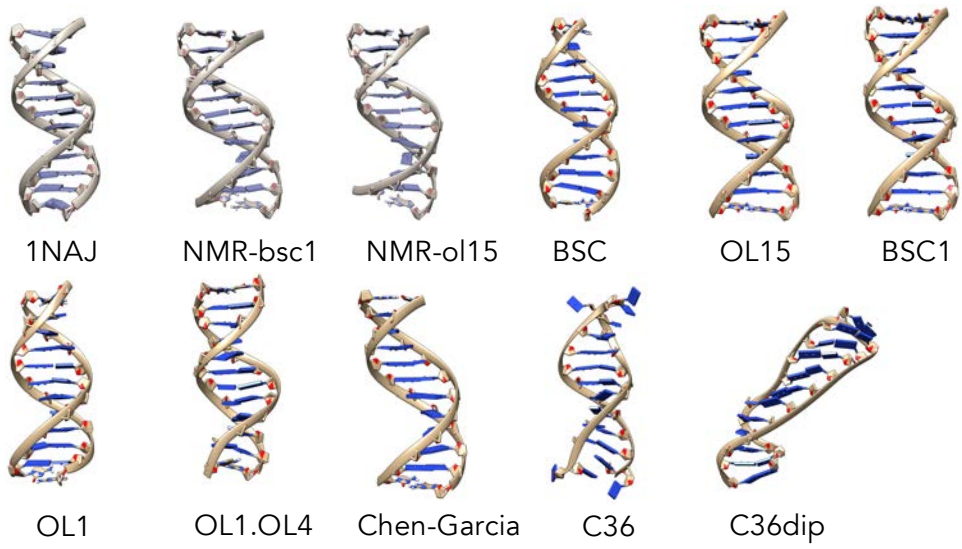



**Figure 4.3. Structural comparison of NMR and MD averaged structures.** NMR structures (PDB: 1NAJ) and two *de novo* obtained in the group by refinement using parmbsc1 and OL15 force fields (shown in grey; top left corner), while MD average structures are obtained from last 20 ns of the MD simulations using different force fields.

Critical analysis of NMR-derived data illustrate that the "experimental" models typically used as "gold standards" in force field validation are no so robust and small changes in transforming experimental restraints into structures can induce not-negligible changes in the final models, which are special evident when looking at sequence-dependent properties. Our results suggest that some caution is required before assuming a "structural" model as the "true", and points theoretical ensembles as an excellent alternative to experimentally derived solution structures. Never before MD simulations has been able to provide structural ensembles with the quality that can be obtained with current force fields.

# HOW ACCURATE ARE ACCURATE FORCE-FIELDS FOR DNA? (under preparation)

Pablo D. Dans[1,2,&], Ivan Ivani[1,2,&], Guillem Portella[1,2,3], Adam Hospital[1,2], Carlos González[4,*] and Modesto Orozco[1,2,5,*]

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

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

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

[4] Instituto Química Física Rocasolano. Consejo Superior de Investigaciones Científicas. Madrid, Spain.

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

[&] Equally contributing authors.

[*] Correspondence to: Prof. Modesto Orozco (modesto.orozco@irbbarcelona.org) or Prof. Carlos González (cgonzalez@iqfr.csic.es).

## ABSTRACT

Last generation of force-fields are raising expectations on the quality of molecular dynamics (MD) simulations of DNA, as well as to the belief that theoretical models can substitute experimental ones in many cases. However these claims are based on limited benchmarks, where MD simulations have shown just the ability to reproduce reasonably well already existing "experimental models" whose accuracy to represent DNA conformation in solution is sometimes unclear. We present here a reverse validation approach, by first running simulations on a series of DNA duplexes of unknown structure, and later solve them by using NMR spectroscopy. Our approach allowed us not only to check directly for experimental observables on duplexes previously not solved, removing consequently the risk of overtraining in theoretical simulations, but also to assess the reliability of "experimental structures" and its dependence on subtle details of the refinement procedure. Overall we found that simulations using last generation of AMBER force-fields have a very high predictive power, and can be safely use to reproduce global structure of DNA duplexes and even fine sequence-dependent details.

## INTRODUCTION

Since the first prototypes published in the seventies, DNA force-fields have been under continuum refinement. The accessibility of an increasing amount of experimental data and the possibility to perform high-level quantum mechanical (QM) calculations has provided the required reference data for force-field refinement, but the real engine behind the improvement of force-fields has been the continuum increase in hardware and software capabilities. Thus, as a new generation of hardware and software allowed the access to larger trajectory time scale, errors in the force-field that became hidden in shorter simulations emerged, forcing a community effort to solve them. In this sense, problems in twist emerging in sub-nanosecond scale parm94 simulations led to the development

of parm99, which was the dominant force-field until multi nanosecond trajectories reported the presence of artefactual α/γ transitions, which accumulated in time corrupting the entire duplex. These issues were solved by the parmbsc0 revision, which became the "gold standard" for almost a decade, until microsecond scale trajectories highlighted the existence of other errors, which required further recalibration of the force-field, leading to parmbsc1, and to the Czech's family of force-fields. A similar type of error-driven refinement happened for the CHARMM family of force-fields until the latest two-body and polarized versions were developed. Recent studies have shown that, for example, latest generation of AMBER force-fields are able to provide reasonable B-like duplex structures in molecular dynamics (MD) trajectories millisecond time scale, far beyond the usual requirement of MD users.

There is little doubt that last generation of force-fields provides reasonable pictures of regular DNA duplexes, but how accurate is the information that can be derived from MD trajectories? Is it, for a given oligonucleotide comparable with that derived in solution by experimental techniques? Can it be safely used to parameterize coarse grain models? It is easy to be overoptimistic on the quality of last-generation force-fields, but despite a general optimism, not convincing evidence on their quality exists, since benchmark studies are rare, sometimes limited to a prototypical duplex (the Drew-Dickerson dodecamer; DDD; with the obvious risk of overtraining in the force-field), or to performed just with the ambition to reproduce general properties of DNA for a long series of duplexes. More extensive benchmark considering several duplexes, like that reported in the parameterization of parmbsc1 are typically based on the comparison of MD trajectories with a reference "experimental model". The risks of this type of validation strategy have been largely underestimated as the quality of "experimental models" is taken as a dogma, while the PDB is full of artifacts. For example, DNA often crystallizes in the A-form, as a left handed Z-DNA helix, or as an H-like conformation, while none of these structures is significantly populated in physiological conditions. Similarly, NMR-derived structures should be taken also with some caution, since NMR spectroscopy does not provide direct information on the structure, but just observables that are then manipulated by mathematical models to derive geometrical restrains (for example average torsions or proton-proton distances) which are imposed to force-field based sampling algorithms to define the "experimental model". We cannot ignore that even when high quality NMR spectra are used not all the details of a NMR-solved structure are equally well defined, conflicting restrains may exist, and technical details in spectra acquisition and processing may impact dramatically in the final structural model. Experimental structures are typically taken as the "truth", but they are just models, sometimes of uncertain quality.

We present here a systematic unbiased validation of the last-generation force-fields for DNA. We choose first the well know DDD duplex, as this structure it is very well refined and helps as benchmark of the quality of experimental models derived by the current standard NMR procedures. For this purpose we collected "de novo" NMR data for DDD, using them into different resolution protocols to determine structural model whose quality can be checked by comparison with a myriad of ultra-high resolution X-Ray structures, a very-accurate NMR model (1NAJ), and

accurate wide angle scattering data (WAXS). This preliminary study provides two interesting types of results: i) the expected accuracy of the best NMR models that can be refined from typical NMR-data, and ii) the "optimal" processing method to transform NMR spectra into structural models. Once these two points were focused our attention in two other duplexes, of unknown experimental structure: d(GCTAGCGAGTCC)·d(GGACTCGCTAGC) and d(GGAGACCAGAGG)·d(CCTCTGGTCTCC). We collected NMR data in solution for both of them and solve their structure using the optimized refinement protocol determined before. In parallel, we collected unbiased MD trajectories for DDD (as control) and for the two other duplexes using the last generation of force-fields: parmbsc0, parmbsc0 including OL1 corrections, OL1+OL4 corrections, OL1+OL4+OL5 (OL15; corrections; &), parmbsc0 with Chen-Garcia modifications, parmbsc1, Charmm36, and the new polarizable Charmm36. Theoretical ensembles were then compared with "experimental models" as well as with direct experimental observables.

Results demonstrate that NMR-derived models are robust to the force-field used in refinement, but are more dependent than expected from other details on the refinement procedure. The global structure is very well defined from the NMR spectra, but sequence-dependent structural details might be in some cases bias due to compensatory variations along the duplex that can lead to sample extreme values of the expected distribution of helical parameters at the base-pair step resolution. Unbiased simulations demonstrate that not all force-fields provide results of the same quality and some of them can produce poor results. However, last generation AMBER force-fields, particularly OL15 and parmbsc1 provide structural data (both global and local) of very high quality for B-DNA duplexes. In fact, our results strongly suggest that expected quality of the ensembles obtained with last-generation AMBER force-fields can be similar to that of NMR-derived structural models.

## METHODS

**Force-field selection**. We evaluate here the most prominent AMBER and CHARMM families of DNA force-fields. From AMBER family force-fields we test the default parmbsc0 , the OL1, refined in ε/ζ backbone dihedrals; OL4, refined in χ torsion, coupled with OL1 (noted as OL1_OL4); recently published OL15, refined in β backbone dihedral and coupled with previous corrections OL1 and OL4, and the parmbsc1. All these force-fields share the same non-bonded part of the force-field that comes from the old parm94 force-field. Lastly, we tried DNA adapted version of Chen and Garcia's force-field for RNA (noted as CG*),* which follow AMBER definitions, with refined χ torsion, as well scaled vdW terms to reproduces weaker stacking interactions. From CHARMM force-field we benchmarked latest CHARMM36 for DNA (noted as *C36*) and recent polarized version of the same force-field based on classical Drude-particle oscillators (noted as *C36_pol*).

**System preparation**. All simulations done with AMBER14 package (parmbsc0, parmbsc1, OL1, OL1_OL4, OL15, CG) were prepared using *leap* extension of AMBER14 package. CHARMM36 simulations were prepared using *grompp* extension of GROMACS simulation package, while

simulations with polarized C36_pol force-field were prepared using Drude Prepper from CHARMM-GUI server. NMR derived structures were used as starting points for the simulations. All the systems were solvated in TIP3P box of water molecules with a minimum of 10 Å beyond the solute, neutralized with Na+ ions with additional 150 mM of NaCl. Ion parameters from Smith and Dang were used for AMBER family simulations, while the default CHARMM ion parameters were considered for that family.

**Molecular Dynamics Simulations**. We have performed 2 μs simulations of the 3 duplexes for each force-field, except for computationally demanding C36_pol force-field for which we have performed 1.2 μs simulation of SEQ1 and 100 ns of SEQ2 and SEQ3 (use of the polarized force-field increases with our computer resources around 10 time the cost of the equivalent pair-additive simulations). We have use Particle Mesh Ewald (PME) code from the programs AMBER14 or GROMACS, depending on the given simulation. For C36_pol simulations we used a special NAMD code. As described in a previous work no major deviations are expected from the use of different computer codes. Unless otherwise noted NPT conditions with default temperature and pressure setting, at 300 K and pressure of 1 atm were used. All simulations but C36_dip, used an integration step of 2 fs in conjunction with SHAKE to constrain X-H bonds with default tolerance. Long range electrostatic interactions were calculated using the PME method with default grid settings and tolerance. All structures were first optimized, thermalized and pre-equilibrated for 1 ns using our standard equilibrium protocol and were later equilibrated for 10 ns. We used default settings coming from CHARMM-GUI server for all C36_dip simulations, which is mainly different to other simulations in using 1 fs time step and dual-Langevin thermostat scheme.

**NMR analysis**. NMR spectroscopy studies were performed to obtain experimental constraints that can later be compared directly with equivalent observables from MD simulations, to determine in detail the behavior of certain structural details as well as to provide "structural models" that can be then used to benchmark unbiased simulations.

*NMR experiments.* Samples of DDD (SEQ1), SEQ2 and SEQ3 duplexes (~1.5 mM duplex concentration) were suspended in 500 μL of either $D_2O$ or $H_2O/D_2O$ 9:1 in 25 mM sodium phosphate buffer, 125 mM NaCl, pH 7. NMR spectra were acquired in a Bruker Advance spectrometer operating at 800 MHz, and processed with Topspin software. DQF-COSY, TOCSY and NOESY experiments were recorded in $D_2O$ and $H_2O/D_2O$ 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 was used for semiautomatic assignment of the NOESY cross-peaks and quantitative evaluation of the NOE intensities**.**

*NMR assignments and experimental constraints.* Sequential assignments of exchangeable and non-exchangeable proton resonances were performed following standard methods for right-handed, double-stranded nucleic acids, using DQF-COSY, TOCSY and 2D NOESY spectra. Complete assignment could be carried out with the exception of some H5'/H5" protons, and some guanine amino resonances which are not observed. Spectral assignment pathways

are shown in Figures SXX. Quantitative distance constraints were obtained from NOE intensities by using a complete relaxation matrix analysis with the program MARDIGRAS. Error bounds in the inter-protonic distances were estimated by carrying out several MARDIGRAS calculations with different initial models (standard A- and B-forms), mixing times (100, 200 and 300 ms) 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. No solvent exchange effects were taken into account in the analysis of NOE intensities in $H_2O$. Therefore, only upper limits were used in the distance constraints involving labile protons. In case of severe overlapping between cross-peaks, the NOE intensities are not considered reliable enough for the complete relaxation analysis and the only qualitative upper distance limits were set according to a visual classification of NOEs in strong, medium and weak. J-coupling constants were roughly estimated from DQF-COSY cross-peaks. In all cases, DQF-COSY cross-peaks were consistent with a South domain conformation.

*Refinement of experimental structures*. Two different approaches were used to derive ensembles of structures using atomistic force-fields based on the distance constraints obtained experimentally. The first approach, labeled as *Standard* in this work, refers to the classical and usual annealing procedure performed using parmbsc0 force-field. Accordingly, ideal fiber B-DNA and A-DNA structures are thermalized (298 K) and equilibrated for 100 ps each (using the same options described previously), applying harmonic restraints of 100 kcal/mol·Å$^2$ on the DNA. Then, a 500 ps MD simulation is performed where the global restraints are replaced by the specific NMR distance constraints obtained experimentally (each represented by a harmonic restraint of 20 kcal/mol·Å$^2$). To obtain the final ensemble, fifty structures (one every 10 ps) were chosen and minimized individually *in vacuo* at 0 ºK removing ions and waters but keeping the NMR constraints. In the second approach different starting structures were used (using the 3 reference force-fields: parmbsc0, parmbsc1 and OL15), as well as a different annealing protocol. The same thermalization and equilibration procedures were used but starting from equilibrated structures (taken after 1 μs of simulation time) obtained from the unbiased MD simulations described previously. Then, 3 MD simulations of 500 ps/50 ps/500 ps were performed in this way: i) The NMR constraints were smoothly applied (from 0 to 500 ps) scaling linearly the harmonic restraints from 2 to 20 kcal/mol·Å$^2$. ii) the system is then cooled down during 50 ps from 298 to 50 ºK maintaining the restraints. iii) Finally, the last segment of 500 ps of MD simulation at 50 ºK with the NMR constraints (20 kcal/mol·Å$^2$) are used to generate the ensemble of structures (50 structures, one every 10 ps). Depending on the origin of the initial structures, these ensembles were labeled in this work as: NMR-BSC0, NMR-BSC1 and NMR-BSC0$_{OL15}$.

**Analysis**. During production runs, data was typically collected every 1 ps, which allowed us to study infrequent, but fast movements. Geometrical analysis were carried out with AMBERTOOLS 15, GROMACS tools, MDWeb, NaFlex, and the Curves+ package. As in our recent work, the 3D-RISM model was used to compute the SAXS-WAXS spectra (*Small-Angle and Wide-Angle X-ray Scattering*) of the experimental structures 1BNA (X-ray), 1NAJ (NMR), 1GIP (NMR), the NMR structures derived in-house, and the average structure from the unbiased MD simulations (computed

with cpptraj from the last 200 ns of simulation). SAXS-WAXS spectra were only computed for the DDD sequence, from which the experimental solution scattering profile was available. The distribution function of waters and ions computed with RISM also considered a TIP3P solution with 150 mM of added NaCl. The statistical analysis was obtained with the R 3.0.1 statistical package and the ggplot2 library, or with MATLAB version 2014a. The molecular plots were generated using either VMD 1.9, or the UCSF Chimera package version 1.8.1.

## RESULTS AND DISCUSSION

**Are "experimental structures" accurate?** As described above, "experimental structures" are in reality just models which fulfill a series or geometrical retrains derived from the processing of some experimental observables. In particular, most "NMR-experimental structures" in solution are derived from MD simulations that incorporate three types of experimental restrains: i) the interchangeability of protons which provide a direct information on the hydrogen bonding scheme, ii) the J-couplings which provide direct information on certain torsional angles (for example those defining sugar puckering), and iii) the NOE intensities, which after processing yield average inter-proton distances. Additional restrains, such as the residual dipolar couplings (RDC) can be incorporated, but this is still not a common practice. Fortunately for our purposes, DDD was used as a model for a tour-of-force for NMR refinement and structure at PDB entry 1NAJ was refined considering all possible NMR-derived restrains, leading to what is supposed to be the most accurate model of DDD in solution. Comparison of our "de novo" NMR structure with 1NAJ provides us direct information on the errors expected in NMR-derived models obtained using the current standards for NMR structural refinement. Additional information on the accuracy of the new experimental structure of DDD can be obtained by comparing with high resolution X-Ray data (excluding terminal bases to reduce lattice artifacts), with other NMR-refined models, and finally with low resolution data derived from wide angle scattering spectroscopy in solution (WAXS).

The currently standard procedure to refine a DNA structure from NMR data starts with a series of NMR-restrained MD simulations, taking A- and B- helices starting structures. When convergence of the different trajectories is clear, an annealing process is done, yielding to a set of DNA conformations which are expected to define the structural ensemble in solution (see *Methods*). For DDD this procedure leads to a fast convergence of all trajectories to the B-basin, and to a narrow set of structures pertaining to the B-family and that reproduces well experimental restrains, which is typically considered to signal a good quality in the refined model (Tables 1 and 2). The optimized structures are globally similar to previously reported "experimental structures" for DDD (see Table 1), but looking in detail disturbing differences become evident. For example (see Figure 1) twist of central d(ApT) step is very low in ensembles obtained using the standard refinement procedure compared with 1NAJ, with crystal structures and with the expected values obtained from database analysis (see Suppl Figure S1). This under-twisting is corrected in neighboring d(ApA) steps, which adopt unusually large twist values (Figure 1 and Supp. Figure S1) leading to an overall correct helix. This sharp twist profile is not directly supported by specific NOEs in this

region, and is not due to errors in the force-field (pambsc0 as default; see below), but it is mostly related to equilibration artefacts probably produced by the sharp cooling of the system, which reduce dramatically NOEs violation, but at the expense of distorting locally the structure. In fact, when a more elaborated refinement procedure is used (see *Methods*, Table 1 and Figure 1) with exactly the same experimental restraints better helical profiles are obtained for all the force-fields (Tables 1, 2, Figure 1 and Suppl. Table S1). It is worth noting that the force-field used in refinement seems less relevant for the final quality of the model, as the NMR-structures refined using parmbsc0, parmbsc1 and OL15 are quite similar (see Tables 1 and 2). Finally, a few words of caution are needed on the use of the "NOE violations" as a direct undisputable measure of the quality of structural ensembles. Thus, NMR-refined ensembles in PDB codes 1NAJ or 1GIP, or X-Ray structures lead to a non-negligible number of violations of our NMR data, while these experimental structural models are close to ours (see Figure 1 and Table 2), while the structural model refined from the standard NMR- procedure (see above), which leads to unrealistic sequence dependent properties (Figure 1 and Table 2), shows structural ensembles with the best NOE violation metrics.

In summary, systematic analysis with a very well characterized B-DNA duplex, strongly suggests that while global structure can be safely recovered by NMR-restrained models, some caution is needed when going to details, since they are depended on the quality of the data and on the way in which they are processed, and that can contain non-negligible local errors, difficult to detect "a priori". This means that: i) some indulgence should be apply to simulation results, as significant local deviations between NMR-derived models and theoretical results are not always signaling a poor quality of the later, and ii) some caution should exist when helical parameters derived at the base-pair step level from NMR-data are transferred to reproduce structural properties of other duplexes.

**Do force fields corrupt duplex structure?** As described above, we have collected accurate NMR observables for three very different DNA duplexes, which in all the cases are found as stable B-type structures. We can then compare the structural models derived by imposing NMR restraints (using the mild refinement procedure outlined above, which seems to correct some of the errors of the standard procedure) with unbiased ms- scale simulations performed with the different force-fields. As shown in Figures 2 and 3, not all the force-fields provide samplings consistent with the experimental data. For example, the scaling down of van der Waals interactions in CG-force-field leads to structures which are far from those expected for a B-DNA duplex. CHARMM36 provides reasonable structures for the central portion of the duplex (perhaps with the exception of roll), but terminal fraying is too large distorting the geometry neighboring pairs (see Figure 3 and discussion below). The newly polarizable CHARMM36pol force-field represents, in our opinion, a milestone in the development of a new generation of force-fields, as it is able to maintain the duplex integrity for around 100 ns (which should be considered a major success for this type of experimental force-fields), but room for improvement exist in the balance of the different interactions, as all the duplexes simulated with this force-field are extremely distorted in the ms scale (see Figure 3). Trajectories obtained with parmbsc0, the

different patches added to parmbsc0 by Jurečka and coworkers, and parmbsc1 provide stable helices belonging in all the cases to the B-family (see Figures 2-4).

**What is the global quality of theoretical DNA ensembles?** Unbiased MD trajectories obtained from parmbsc1 and OL15 simulations provide samplings of the DDD conformational space that are globally hard to distinguish from the experimental models, as noted in RMSd in the range 1.3-1.7 Å (Table 3a) to the different experimental models, values which are not far from the range 1.0-1.5 Å found between the different experimental models (Table 1). Similarly, average helical parameters obtained from unbiased MD trajectories of DDD using parmbsc1 or OL15 force-fields are within the range of variability of experimental models (Table 4, Figure 2). All other BSC0-based force-fields behave also reasonably well in terms of general structure for DDD, while significantly larger RMSds and worse helical parameters are found for the rest of the tested potentials (see also Figure 2). As discussed above, it is important to not only evaluate the similarity between unbiased MD samplings and experimental structural models, but also the ability of unbiased ensembles to reproduce direct experimental observables. Not surprisingly, parmbsc1 simulations reproduce very well NMR restrains used in solving 1NAJ, and encouraging, also the new NMR observables collected here (Table 5). In fact, the parmbsc1 unbiased trajectories for DDD seem to be more consistent with NMR observables than many of the experimental models deposited in PDB (compare Table 2 and 5). Furthermore, parmbsc1 trajectories reproduce also very well the challenging WAXS spectrum, which is not well reproduced for most of the experimental models deposited in PDB (Suppl. Table S1). As expected from the previous sections the new OL15 functional provides also good estimates of experimental observables, while slightly, worse results are derived from parmbsc0, OL1 and OL1+OL4 force-fields. Large deviations between predicted and detected experimental observables are found in simulations performed with the other force-fields considered in this work.

In summary, last generation of AMBER family of force-fields (the 2014-developed parmbsc1 and the 2016-developed OL15) reproduces extremely well the general structure of DDD. However, DDD was always the guinea pig in force-field development, and accordingly this good agreement might just reflect over-training in the force-field. We could argue that overtraining cannot explain agreement on WAXS spectra, or with the new NMR data for DDD collected here. However, as discussed above, in order to have a complete unbiased estimate of the quality of recent force-fields we analyzed theoretically SEQ2 and SEQ3, duplexes for which no experimental information was available at the time of running the simulations. Results in Table 3b confirm the ability of parmbsc1 and OL15 to sample conformations globally close to the refined NMR ones (RMSd around 1.6-1.8 Å; see Table 3), providing average helical coordinates which are very close to the experimental ones (Table 4) for both duplexes (see also Figure 3). NOE violations (Table 5) obtained from parmbsc1 and OL15 samplings at room temperature are obviously larger than those obtained when the NMR restrains are included (see Suppl. Table S2) and temperature is reduced to 50° K, but close to the errors found in Table 2 for other high-quality structures of DDD. In summary, comparison of unbiased trajectories for SEQ2 and SEQ3 with previously unavailable

experimental data rules out hypothesis of overtraining in the development of the last generation of AMBER-family of force-fields and demonstrate the accuracy of these force-fields in terms of global structure in solution. We cannot evaluate here the ability of parmbsc1 and OL15 ensembles to reproduce WAXs spectra, but we provide estimates for future experimental testing (see Suppl. Table S3). As expected from DDD results, previous parmbsc0-based force-fields behave reasonably well, CHARM36 generate some moderate artifacts related to massive fraying (see below) and Chen-Garcia force-field yield to largely under-twisted structures (Tables 3-4). Finally, as discussed for DDD the polarized CHARMM36 behaves very well for dozens of nanoseconds, but later the helical structure is lost (see Figure 3 and Table 4)

**Are helix ends well represented by current force-fields?** The breathing of central base pairs is a very rare event, as it requires breaking dual stacking interactions, something very unlikely in microsecond-long simulations for coding bases (CITA ELENA DFT and references we cite there), but stacking interactions are less intense for terminal base pairs, which are then expected to open more frequently. In fact, there are 110 DNA duplexes in PDB, where at least one terminal base pair is broken (this represents 60% of the DNA structures in PDB). However, 40 of the 110 open pairs are d(A·T), and in 70% of the cases the open pair shows hydrogen bonding interactions with other nucleobases in the crystal lattice. In the three duplexes considered here the helices are caped with d(C·G) pairs, which means that we should expect slight breathing, but rare opening events in time scale of the simulations. Furthermore, the analysis of proton interchangeability indicates that protons at the terminal d(C·G) step have a similar accessibility to solvent than those of a central d(C·G) steep (Figure XX), and sequential NOEs, which provide direct information on local stacking (see Figure XY) have a similar intensity in central and terminal base pair steps. Altogether, even our NMR data cannot provide a quantitative estimate of the opening frequencies, they are inconsistent with a massive opening of the terminal pairs, in disagreement not only with CHARMM36, CHARMM36pol, but also with parmbsc0 OL1 and Chen-Garcia simulations (Table 4, Figure 3 and Figure 4). Trajectories collected with the last generation of AMBER force-fields are in much better agreement with NMR data, reporting a conservation of terminal hydrogen bonding above 96% of the simulation time. The conservation of terminal pairing is an important improvement since opened base often displaced towards the groove leading to a propagation of structural distortions in the central portion of the duplexes.

**Are sequence-dependent properties of DNA well reproduced by force-fields?** Massive initiatives, such as the Ascona B-DNA consortium are using MD simulations of a large number of duplexes to trace sequence-dependent properties of DNA, providing parameters that can be then implemented in coarse-grained helical models to simulate long DNA segments. Unfortunately, except for a few cases, the validity of sequence-dependent geometrical parameters derived from MD simulations has not been yet demonstrated. Figure 2 shows that all parmbsc0-based force-field are able to provide reasonable general profiles of helical properties along the central 10 bp part of the sequence, but a detailed analysis shows that parmbsc0 and OL1-OL4 generate good relative profiles, but underestimate the twist (see above). The OL1 and OL15 simulations

lead to very good profiles in the entire duplex (see Table 6 and Figure 2) except for a certain over-twist at the CG step which generates compensatory under-twist at the neighboring d(GA) and d(GC), an apparently incorrect balance in low/high twist populations at d(CG) step seems to be the responsible of this effect (see Suppl. Fig. S2), which was also present in more extended OL15 simulations by Cheatham and coworkers. Parmbsc1 provide helical profiles, which are in practice, indistinguishable from the experimental ones for the entire duplex (Figure 2 and Table 6). The C36 helical profiles deviate from the experimental one mostly because of the large fraying at the ends (Table 6), but it terminal base pairs are removed from the study the C36 helical profiles are reasonable except for some problems with roll. Globally (Table 6) the CG helical profiles benefit from a better representation of helix termini, but systematic errors in some of the parameters are very clear (Figure 2). Finally, as commented above the C36pol force-field has problems to define properly the helical structure (see Figure 3).

As noted above, it can be claimed that the good ability of recent AMBER force-fields to reproduce DDD helical profiles might be an overtraining artifact. However, both parmbsc1 and OL15 are able to reproduce also very well the global helical properties for SEQ2 and SEQ3 (Tables 3b and 4) as well as the helical profiles (Table 6, Figure 4). Furthermore, the NMR-violations found in unbiased MD simulations using OL15 and parmbsc1 force-fields for SEQ2 and SEQ3 are similar to those found for the DDD (Table 5), and within the range expected from the experimental noise (compare Table 5 and Table 2). In summary, parmbsc1 and OL15 behaves similarly for the DDD and SEQ2/SEQ3 sequences, suggesting that overtraining is not a major source of artefacts. Detailed analysis of individual helical profiles for SEQ2 and SEQ3 (Fig 4) illustrates, however, the existence of some discrepancies between NMR-derived helical profiles and those obtained from unbiased MD simulations. For example, for SEQ2 both parmbsc1 and OL15 suggest a smoother twist profile than that found in NMR-biased calculations, which suggest a very low twist (around 25 degrees) at the central d(CpG) step which leads to a compensatory increase in twist the neighboring steps, with the d(GpA) step sampling twist values above 40 degrees. For SEQ3 the most significant difference between unbiased parmbsc1/OL15 and NMR-restrained simulations are found for tilt, whose profile is quite flat in unbiased parmbsc1/OL15 simulations, while it shows sharp and compensatory variations along the sequence in the NMR-biased simulations.

The analysis of DDD for which very accurate experimental structures were available warned us against too sharp profiles in NMR-refined structures (see above). Thus, we compared NMR-biased and unbiased helical values with the distribution of values (for the same step) reported in previous experimental studies. Interestingly, the low twist at the central d(CpG) step in SEQ2 is consistent with 100% population of the in “low twist” state, something that is possible, but not common in experimental structures (see Suppl. Figure S3), which seems to favor a twist balance more similar to that obtained parmbsc1 and OL15 simulations (Figure S3). The low twist at the CpG step trigers to compensatory changes, visible in twist values for d(GpA) above 40 degrees, which is significantly larger than the values typically sampled in other experimental structures (Suppl. Figure S3). The sharp compensatory tilt variations found for SEQ3 in the NMR-restrained simulations

have not impact in the overall duplex structure, but lead to tilt values very uncommon in previous experimental structures (Suppl. Figure S4). Finally, large roll values (which compensate each other) found in NMR-restrained, but not in MD unbiased simulations for SEQ3 are again rather unusual in previously structures (Suppl. Figure S4). It is worth noting that these discrepancies found between unbiased and NMR-biased structures cannot be justified from the force-field used in NMR-refinement, as results obtained with $NMR_{OL15}$ and $NMR_{pambsc1}$ are very similar (Figure 4).

Previous analysis suggest that both parmbsc1 and OL15 are able to reproduce very well sequence dependent properties as determined by NMR data in solution, but also that quite surprisingly, when deviations appear, it is not so simple to assign the differences to force-field artefacts. The question is then whether the "uncommon" helical parameters sampled by NMR-restrained simulations are directly supported by direct experimental data, or are a consequence of the limited amount of restraints, the mathematical procedure used to implement the restraints or the annealing procedure (the mild cooling method has been used for all the discussions here).

**Are MD structures better than extrapolated models.** Most studies of DNA use an average representation of DNA derived from fiber diffraction data by Arnott and coworkers. More elaborated models introduce sequence-dependence by using average helical parameters derived at the base-pair step assuming the near neighbor model and experimental data in PDB. Alternative approaches use average helical parameters at the base-pair level but in the tetramer context. This approach is coupled to MD ensembles as few tetramers are well covered in PDB. The last question in this paper is whether atomistic MD simulations are required or these sequence-independent or sequence-dependent helical parameters are able to provide already accurate duplexes. To investigate this point we use BigNasim to create expected models using average helical parameters from Arnott, average base pair step parameters from PDB, and ABC tetramer parameters.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGMENTS

MO is an ICREA academia researcher. PDD is a PEDECIBA and SNI (ANII, Uruguay) researcher.

## FUNDING

MINECO Severo Ochoa Award of Excellence, Government of Spain (to IRB Barcelona); Spanish Ministry of Science [BIO2012-32868, 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.); Funding for open access charge: European Research Council (ERC SimDNA).

## REFERENCES

[1] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case, and M. Orozco, "Parmbsc1: a refined force field for DNA simulations.," *Nat. Methods*, vol. 13, no. 1, pp. 55–58, 2015.

[2] M. Zgarbová, F. J. Luque, J. Šponer, T. E. Cheatham III, M. Otyepka, and P. Jurečka, “Toward improved description of DNA backbone: revisiting epsilon and zeta torsion force field parameters,” *J. Chem. Theory Comput.*, vol. 9, no. 5, pp. 2339–2354, 2013.

[3] M. Krepl, M. Zgarbová, P. Stadlbauer, M. Otyepka, P. Banáš, J. Koča, T. E. Cheatham III, P. Jurečka, and J. Šponer, “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.*, vol. 8, no. 7, pp. 2506–2520, 2012.

[4] M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo, and P. Jurečka, “Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA,” *J. Chem. Theory Comput.*, vol. 11, no. 12, pp. 5723–5736, 2015.

[5] A. A. Chen and A. E. García, “High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 42, pp. 16820–16825, 2013.

[6] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, “Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers,” *Biophys. J.*, vol. 92, no. 11, pp. 3817–3829, 2007.

[7] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, and A. D. MacKerell Jr, “Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium,” *J. Chem. Theory Comput.*, vol. 8, no. 1, pp. 348–362, 2011.

[8] A. Savelyev and A. D. MacKerell, “All-atom polarizable force field for DNA based on the classical drude oscillator model,” *J. Comput. Chem.*, vol. 35, no. 16, pp. 1219–1239, 2014.

[9] D. A. Case, V. Babin, J. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham Iii, T. A. Darden, R. E. Duke, and H. Gohlke, “Amber 14,” 2014.

[10] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, “GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation,” *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, 2008.

[11] S. Jo, T. Kim, V. G. Iyer, and W. Im, “CHARMM-GUI: a web-based graphical user interface for CHARMM.,” *J. Comput. Chem.*, vol. 29, no. 11, pp. 1859–65, Aug. 2008.

[12] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[13] D. E. Smith and L. X. Dang, “Computer simulations of NaCl association in polarizable water,” *J. Chem. Phys.*, vol. 100, no. 5, pp. 3757–3766, 1994.

[14] H. Yu, T. W. Whitfield, E. Harder, G. Lamoureux, I. Vorobyov, V. M. Anisimov, A. D. MacKerell Jr, and B. Roux, “Simulating monovalent and divalent ions in aqueous solution using a Drude polarizable force field,” *J. Chem. Theory Comput.*, vol. 6, no. 3, pp. 774–786, 2010.

[15] W. Jiang, D. J. Hardy, J. C. Phillips, A. D. MacKerell, K. Schulten, and B. Roux, “High-performance scalable molecular

dynamics simulations of a polarizable force field based on classical drude oscillators in NAMD," *J. Phys. Chem. Lett.*, vol. 2, no. 2, pp. 87–92, 2011.

[16] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, 1977.

[17] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log (N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[18] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, S. Dixit, B. Jayaram, F. Lankas, and C. Laughton, "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.*, vol. 38, no. 1, pp. 299–313, 2010.

[19] A. Hospital, P. Andrio, C. Fenollosa, D. Cicin-Sain, M. Orozco, and J. L. Gelpí, "MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations," *Bioinformatics*, vol. 28, no. 9, pp. 1278–1279, 2012.

[20] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+," *Nucleic Acids Res.*, vol. 37, no. 17, pp. 5917–5929, 2009.

[21] M. Orozco, A. Pérez, A. Noy, and F. J. Luque, "Theoretical methods for the simulation of nucleic acids," *Chem. Soc. Rev.*, vol. 32, no. 6, pp. 350–364, 2003.

[22] A. Pérez, J. R. Blas, M. Rueda, J. M. López-Bes, X. de la Cruz, and M. Orozco, "Exploring the essential dynamics of B-DNA," *J. Chem. Theory Comput.*, vol. 1, no. 5, pp. 790–800, 2005.

[23] A. Amadei, A. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins Struct. Funct. Bioinforma.*, vol. 17, no. 4, pp. 412–425, 1993.

[24] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein–DNA crystal complexes," *Proc. Natl. Acad. Sci.*, vol. 95, no. 19, pp. 11163–11168, 1998.

[25] F. Lankaš, J. Šponer, P. Hobza, and J. Langowski, "Sequence-dependent elastic properties of DNA," *J. Mol. Biol.*, vol. 299, no. 3, pp. 695–709, 2000.

[26] A. Noy, A. Perez, F. Lankas, F. J. Luque, and M. Orozco, "Relative flexibility of DNA and RNA: a molecular dynamics study," *J. Mol. Biol.*, vol. 343, no. 3, pp. 627–638, 2004.

[27] I. Andricioaei and M. Karplus, "On the calculation of entropy from covariance matrices of the atomic fluctuations," *J. Chem. Phys.*, vol. 115, no. 14, pp. 6289–6292, 2001.

[28] J. Schlitter, "Estimation of absolute and relative entropies of macromolecules using the covariance matrix," *Chem. Phys. Lett.*, vol. 215, no. 6, pp. 617–621, 1993.

[29] B. Hess, "Similarities between principal components of protein dynamics and random diffusion," *Phys. Rev. E*, vol. 62, no. 6, p. 8438, 2000.

[30] T.D. Goddard, and D.G. Kneller. University of California, San Francisco 2006.

[31] B.A. Borgias, and T.L. James, MARDIGRAS-A procedure for matrix analysis of relaxation for discerning

geometry of an aqueous structure. J. Magn. Reson., 87, p. 475-487, 1990.

[32] G. Rossetti, P.D. Dans, I. Gomez-Pinto, I. Ivani, C. Gonzalez, M. Orozco. The structural impact of DNA mismatches. Nucl. Acids Res., 43, p. 4309-4321, 2015.

[33] P.D. Dans, L. Danilāne, I. Ivani, T. Dršata, F. Lankaš, J. Walther, R. Illa Pujagut, F. Battistini, J. Ll. Gelpí, R. Lavery, M. Orozco. Long-timescale dynamics of the Drew–Dickerson dodecamer. Nucl. Acids Res., 44, p. 4052-4066, 2016.

[34] H.T. Nguyen, S.A. Pabit, S.P. Meisburger, L. Pollack, and D.A. Case. Accurate small and wide angle x-ray scattering profiles from atomic models of proteins and nucleic acids. J. Chem. Phys., 141, p. 22D508. 2014.

[35] X. Zuo, and D.M. Tiede. Resolving conflicting crystallographic and NMR models for solution-state DNA with solution X-ray diffraction. J. Am. Chem. Soc., 127, p. 16–7, 2005.

[36] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2013.

[37] H. Wickham. ggplot2. Springer New York, New York. 2009.

[38] MATLAB and Statistics Toolbox Release 2014a, The MathWorks, Inc., Natick, Massachusetts, United States.

[39] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. J. Mol. Graph., 14, p. 33–8, 1996.

[40] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem., 25, p. 1605–12. 2004.

[41] Soliva, R., Monaco, V., Gomez-Pinto, I., Meeuwenoord, N.J., Marel, G.A., Boom, J.H., Gonzalez, C. and Orozco, M. (2001) Solution structure of a DNA duplex with a chiral alkyl phosphonate moiety. Nucl. Acids Res., 29, 2973-2985

## TABLES

**Table 1** | Comparison of average RMSd values (in Å) of the NOEs-restrained MD simulations calculated in reference to NMR or X-RAY structures of the DDD sequence.

| Structure | Standard[c] | BSC1-NOE[d] | BSC0-NOE[d] | $BSC0_{OL15}$-NOE[d] |
|---|---|---|---|---|
| **1NAJ**[a] | 1.32 | 1.07 | 1.20 | 1.22 |
| **1GIP**[a] | 1.09 | 1.14 | 1.15 | 1.23 |
| **XRAY**[b] | 1.72 | 1.39 | 1.43 | 1.49 |
| **1JGR** | 1.64 | 1.35 | 1.35 | 1.46 |
| **4C64** | 1.67 | 1.37 | 1.39 | 1.48 |

[a] The RMSd calculations were done against an average structure obtained from NMR conformations with PDB code 1NAJ and 1GIP. [b] The averages were obtained combining the X-Ray structures with PDB codes: 1BNA, 2BNA, 7BNA and 9BNA. Note that the capping base-pairs were not considered. [c] NMR structures obtained by using the standard refinement process with annealing and optimization using the default BSC0 force-field (see Methods). [d] NMR structures obtained by using the mild annealing procedure described in the Methods Section using 3 force-fields: BSC1, BSC0, and $BSC0_{OL15}$.

**Table 2** | Summary of NOE distances violation and energy penalties for the DDD sequence using the NMR data obtained in-house (see Methods).[a]

| Structure | Nº of violations | Energy penalty[b] (kcal/mol) | Average violation (Å) | Largest violation (Å) |
|---|---|---|---|---|
| **BSC1-NOE** | 8 | 20.4 | 0.35 | 0.46 |
| **BSC0-NOE** | 9 | 22.7 | 0.35 | 0.43 |
| **$BSC0_{OL15}$-NOE** | 14 | 35.0 | 0.35 | 0.47 |
| **Standard NMR** | 5 | 11.3 | 0.33 | 0.40 |
| **1NAJ**[c] | 35 | 159.2 | 0.47 | 0.63 |
| **1GIP**[c] | 50 | 265.4 | 0.51 | 1.06 |
| **X-RAY**[c] | 46 | 332.1 | 0.60 | 1.58 |

[a] Taking T as 298.15 K, the kT constant has a value of 0.5924812 kcal $mol^{-1}$. We considered an experimental restraint violated when its average penalty energy was above 3·kT. Given the force constant used to apply the distance restraints ($k_{res}$ = 20 kcal/$Å^2$), 3·kT is equivalent to set a tolerance of ±0.3 Å on the experimental range to consider that a specific distance has been violated. [b] For each distance the energy penalty ($E_{pen}$) was computed as: $E_{pen} = k_{res}(dist_{calc}-dist_{obs})^2$. Note that we simply reported the sum of each individual $E_{pen}$. [c] A single-point calculation *in vacuo* was performed on the average experimental structure applying our NMR restraints.

**Table 3a** | Comparison of average RMSd values (in Å) calculated in reference to NMR or X-RAY structures of the DDD sequence.

| Force-field | 1NAJ[a] | NMR[b] | X-RAY[c] | 1JGR | 4C64 |
|---|---|---|---|---|---|
| **BSC1** | 1.39 | 1.61 \| 1.81 | 1.68 | 1.63 | 1.69 |
| **BSC0** | 1.77 | 1.87 \| 2.04 | 2.78 | 2.06 | 2.17 |
| **$BSC0_{OL1}$** | 1.65 | 1.72 \| 1.87 | 1.90 | 1.83 | 1.91 |
| **$BSC0_{OL1+OL4}$** | 1.85 | 1.74 \| 2.00 | 2.06 | 1.94 | 2.03 |
| **$BSC0_{OL15}$** | 1.46 | 1.67 \| 1.83 | 1.66 | 1.65 | 1.70 |
| **Chen-García** | 4.12 | 3.50 \| 3.88 | 4.32 | 4.15 | 4.22 |
| **C36** | 3.29 | 3.27 \| 3.40 | 3.40 | 3.37 | 3.40 |
| **C36pol** | 10.36 | 10.27 \| 10.28 | 10.01 | 10.10 | 10.03 |

[a] The RMSd calculations were done against an average structure obtained from NMR conformations with PDB code 1NAJ. [b] *de novo* NMR data for the DDD sequence were obtained in our labs (see Methods). First row of numbers correspond to the NMR ensemble refined with BSC1, and the second row with $BSC0_{OL15}$.[c] As in (a), the averages were obtained combining the X-Ray structures with PDB codes: 1BNA, 2BNA, 7BNA and 9BNA. Note that the capping base-pairs were not considered in RMSd calculations.

**Table 3b** | Comparison of average RMSd values (in Å) calculated in reference to *de novo* NMR data collected in our lab (refined using parmBSC1 (top value in the cell) or OL15 (bottom value in cell) force fields) for SEQ2 and SEQ3.

| | BSC1 | BSC0 | OL1 | OL1.OL4 | OL15 | Chen-Garcia | C36 | C36dip |
|---|---|---|---|---|---|---|---|---|
| **SEQ2** | 1.69 | 2.10 | 1.69 | 1.80 | 1.72 | 3.36 | 3.41 | 7.39 |
| | 1.68 | 1.93 | 1.63 | 1.70 | 1.71 | 3.27 | 3.42 | 7.23 |
| **SEQ3** | 1.85 | 2.45 | 2.04 | 2.09 | 1.88 | 3.57 | 4.95 | 4.14 |
| | 1.79 | 2.35 | 1.96 | 2.02 | 1.86 | 3.34 | 4.92 | 4.07 |

Note that the capping base-pairs were not considered in RMSd calculations.

**Table 4** | Comparison of global twist and roll values (in degrees) and average canonical WC hydrogen bond count (HB%) with (all) or without (no ends) terminal base pairs.

| | | DDD | | | SEQ2 | | | SEQ3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Twist** | **Roll** | **HB %** | **Twist** | **Roll** | **HB %** | **Twist** | **Roll** | **HB %** |
| **BSC1** | All | 35.23 | 2.66 | 96.2 | 34.06 | 3.28 | 99.1 | 33.89 | 2.53 | 99.2 |
| | No ends | 34.39 | 1.47 | 99.7 | 34.65 | 2.13 | 99.2 | 34.09 | 2.05 | 99.4 |
| **BSC0** | All | 32.99 | 16.65 | 83.4 | 29.85 | 10.44 | 89.2 | 30.09 | 1.87 | 89.4 |
| | No ends | 32.81 | 2.41 | 99.6 | 32.34 | 3.22 | 98.7 | 31.54 | 3.78 | 98.1 |
| **OL1** | All | 34.16 | 15.85 | 84.8 | 32.75 | 3.88 | 95.5 | 31.52 | 3.71 | 90.6 |
| | No ends | 33.59 | 2.26 | 99.6 | 33.67 | 2.89 | 99.3 | 33.29 | 2.81 | 97.8 |
| **OL1.OL4** | All | 33.45 | 7 | 93.7 | 31.8 | 5.8 | 93.5 | 31.67 | 12.25 | 90.1 |
| | No ends | 33.12 | 2.71 | 99.5 | 32.94 | 3.8 | 99.1 | 32.64 | 4.04 | 98.4 |
| **OL15** | All | 35.01 | 2.97 | 98.7 | 34.62 | 2.3 | 99.1 | 34.27 | 2.9 | 97.7 |
| | No ends | 34.49 | 2.11 | 99.6 | 34.84 | 2.46 | 99.4 | 34.47 | 2.74 | 99.1 |
| **Chen-Garcia** | All | 28.05 | 5.42 | 87.2 | 29.62 | 3.12 | 99.9 | 29.2 | 3.39 | 97.6 |
| | No ends | 29.87 | 3.49 | 92.6 | 29.13 | 3.24 | 99.9 | 28.98 | 3.16 | 99.9 |
| **C36** | All | 30.06 | 19 | 85.2 | 30.56 | 13.14 | 79.2 | 30.57 | 9.26 | 78.4 |
| | No ends | 33.61 | 5.36 | 95.4 | 35.06 | 4.33 | 91.2 | 33.71 | 5.92 | 92.3 |
| **C36dip** | All | 30.72 | 1.47 | 49.4 | 29.11 | 0.63 | 52.8 | 18.46 | 3.43 | 68.4 |
| | No ends | 31.01 | 3.37 | 57.2 | 26.27 | 1.27 | 53.1 | 15.09 | 3.87 | 81.6 |
| **NMR-standard** [a] | All | 34.46 | 4.71 | | 34.38 | 3.36 | | 34.51 | 5.65 | |
| | No ends | 34.85 | 2.29 | | 34.1 | 2.59 | | 34.36 | 3.66 | |
| **NMR** [b] | All | 34.10 | 4.22 | | 34.89 | 4.24 | | 34.79 | 5.38 | |
| | No ends | 34.85 | 2.37 | | 35.14 | 4.12 | | 34.63 | 4.68 | |
| **NMR** [c] | All | 34.27 | 4.57 | | 33.25 | 4.29 | | 34.33 | 4.16 | |
| | No ends | 34.69 | 2.29 | | 33.72 | 3.91 | | 33.97 | 3.58 | |
| **1NAJ** | All | 35.71 | 3,27 | | | | | | | |
| | No ends | 36.07 | 2.12 | | | | | | | |
| **X-ray** | All | 35.69 | -0.31 | | | | | | | |
| | No ends | 35.24 | -0.73 | | | | | | | |
| **1JGR** | All | 35.30 | 0.95 | | | | | | | |
| | No ends | 35.36 | -0.63 | | | | | | | |
| **4C65** | All | 35.37 | 0.56 | | | | | | | |
| | No ends | 35.44 | -0.68 | | | | | | | |

[a] NMR values are averages derived from 10 NMR structures per sequence obtained in the group using standard refinement protocol. [b] NMR values correspond to the NMR ensemble refined with BSC1. [c] NMR values correspond to the NMR ensemble refined with $BSC0_{OL15}$.

**Table 5** | Summary of NOE distance violations from unrestrained MD simulations with the reference force-fields (BSC1, BSC0, $BSC0_{OL15}$) using the NMR data obtained in-house (see Methods) and 1NAJ (only for DDD).[a]

| Force-field | Nº of violations | Largest violation (Å) | Average violation (Å) |
|---|---|---|---|
| | *DDD* | | |
| **BSC1** | 40 \| *2*[b] | 0.94 \| *0.76* | 0.51 \| *0.76* |
| **BSC0** | 37 \| *6* | 2.11 \| *1.35* | 0.66 \| *0.83* |
| **BSC0 $_{OL15}$** | 46 \| *4* | 0.94 \| *0.79* | 0.48 \| *0.77* |
| | *SEQ2* | | |
| **BSC1** | 45 | 1.25 | 0.54 |
| **BSC0** | 53 | 4.15 | 0.83 |
| **BSC0 $_{OL15}$** | 46 | 1.37 | 0.54 |
| | *SEQ3* | | |
| **BSC1** | 51 | 1.19 | 0.59 |
| **BSC0** | 51 | 2.18 | 0.78 |
| **BSC0 $_{OL15}$** | 56 | 1.30 | 0.55 |

[a] We considered an experimental restraint violated when its average penalty energy was above 3·kT. See the footnote comment to Table 2 and the Method section for additional details. [b] Number reported in italic were computed using the NMR restraints from PDB code 1NAJ.

**Table 6**| (A) Global accumulated root mean square deviations (first row in each cell; in degrees) and mean signed error (MSE; second row in each cell, in degrees) between twist, roll and tilt profiles determined from MD simulations and those obtained from the same sequences using NMR-retrained ensembles (values considered here are the average of NMR-OL15 and NMR-parmbsc1 simulations). (B) Metrics as before but considering the six inter base-pair parameters (translations and rotations) mixed using the normalization procedure by Lankas and Maddocks.

| (A) | Parmbsc1 | Parmbsc0 | OL1 | OL1+OL4 | OL15 | CG | C36 |
|---|---|---|---|---|---|---|---|
| DDD | 3.66 | 15.98 | 14.65 | 6.12 | 3.91 | 6.04 | 18.57 |
| | -0.58 | 2.85 | 2.87 | -0.39 | -0.96 | -2.46 | 2.04 |
| SEQ2 | 3.81 | 11.48 | 4.10 | 5.60 | 3.59 | 4.49 | 11.57 |
| | -0.12 | 0.26 | -0.99 | -0.36 | -0.75 | -1.97 | 0.17 |
| SEQ3 | 3.10 | 5.12 | 4.43 | 9.25 | 3.72 | 4.37 | 5.84 |
| | -0.80 | -2.97 | -2.14 | 0.36 | -1.11 | -2.09 | -1.31 |
| (B) | Parmbsc1 | Parmbsc0 | OL1 | OL1+OL4 | OL15 | CG | C36 |
| DDD | 0.34 | 1.07 | 0.99 | 0.52 | 0.42 | 0.65 | 1.29 |
| | 0.00 | 0.26 | 0.26 | 0.06 | 0.06 | -0.24 | 0.08 |
| SEQ2 | 0.37 | 0.80 | 0.49 | 0.55 | 0.44 | 0.68 | 1.05 |
| | -0.02 | 0.00 | -0.09 | -0.05 | -0.01 | -0.31 | -0.13 |
| SEQ3 | 0.39 | 0.84 | 0.65 | 1.02 | 0.57 | 0.71 | 1.11 |
| | -0.02 | -0.05 | -0.09 | 0.06 | 0.04 | -0.24 | -0.17 |

## FIGURES

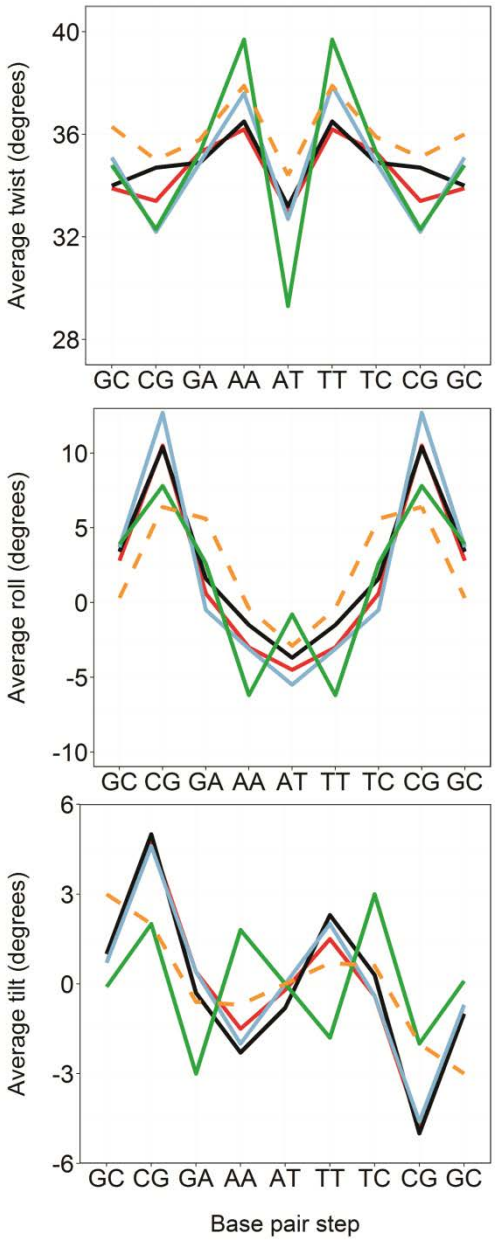



**Figure 1** | Averages values of the average 3 rotational base-pair step helical parameters of DDD computed from NMR-biased MD simulations. BSC0-NOE (red), BSC1-NOE (black), and $BSC0_{OL15}$-NOE (blue). The Standard procedure for refinement, based on fast annealing and optimization with the default BSC0 force-field is shown in green. The rest of profiles were obtained by implementing the mild annealing procedure described in the Methods Section. Profiles are compared with the highest quality NMR structure deposited in the PDB: 1NAJ (orange dashed line). Note that capping base-pair steps were excluded from the analysis.

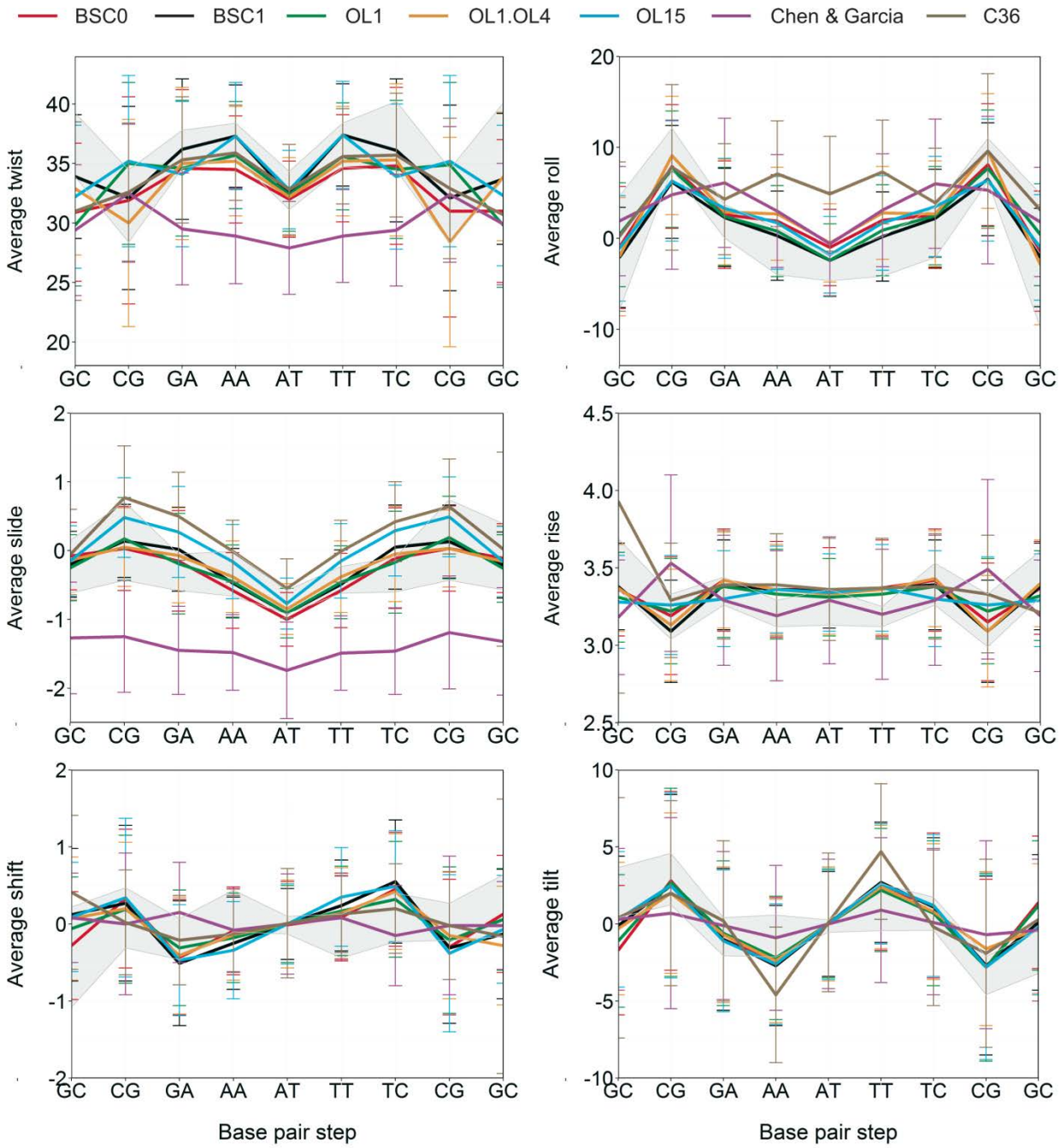



**Figure 2** | Averages and standard deviations of the 6 base-pair step helical parameters of DDD. All the tested force-field are compared with an experimental range (grey zone defined by the average ± standard deviation) obtained by taking the NMR structures 1NAJ, NMR II (in-house data), and the X-ray structures with PDB id 1BNA, 2BNA, 7BNA, 9BNA, 1JGR and 4C64. The polarizable C36 force-field leads to corruption of the helix for μs-scale simulation and results are not shown. See Figure 3 for structural models.

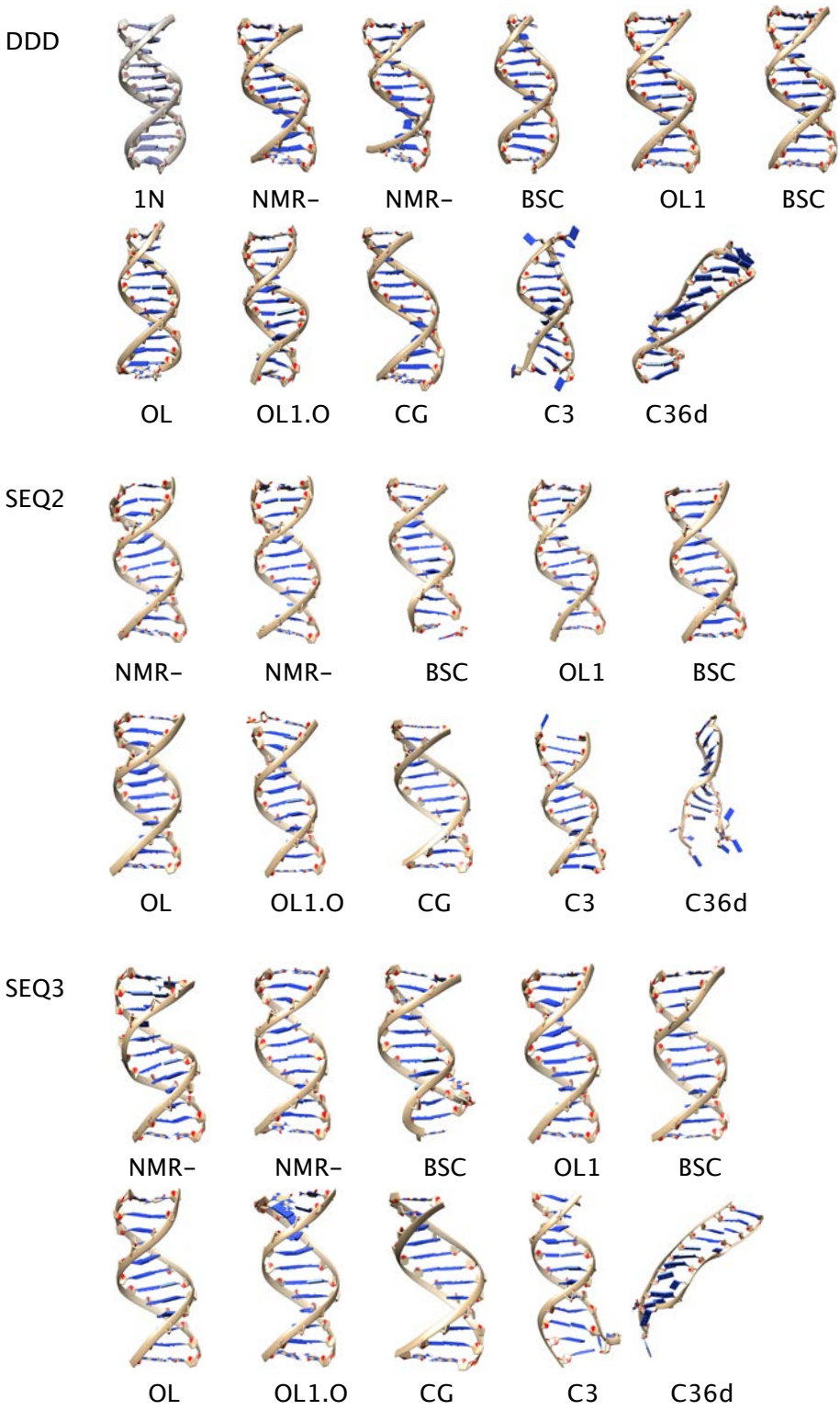



**Figure 3 |** Comparison of MD simulated (colored) structures with NMR obtained (top left; greyish) structures of the three sequences. The MD structures illustrated are average conformations taken from last 20 ns of the trajectory.

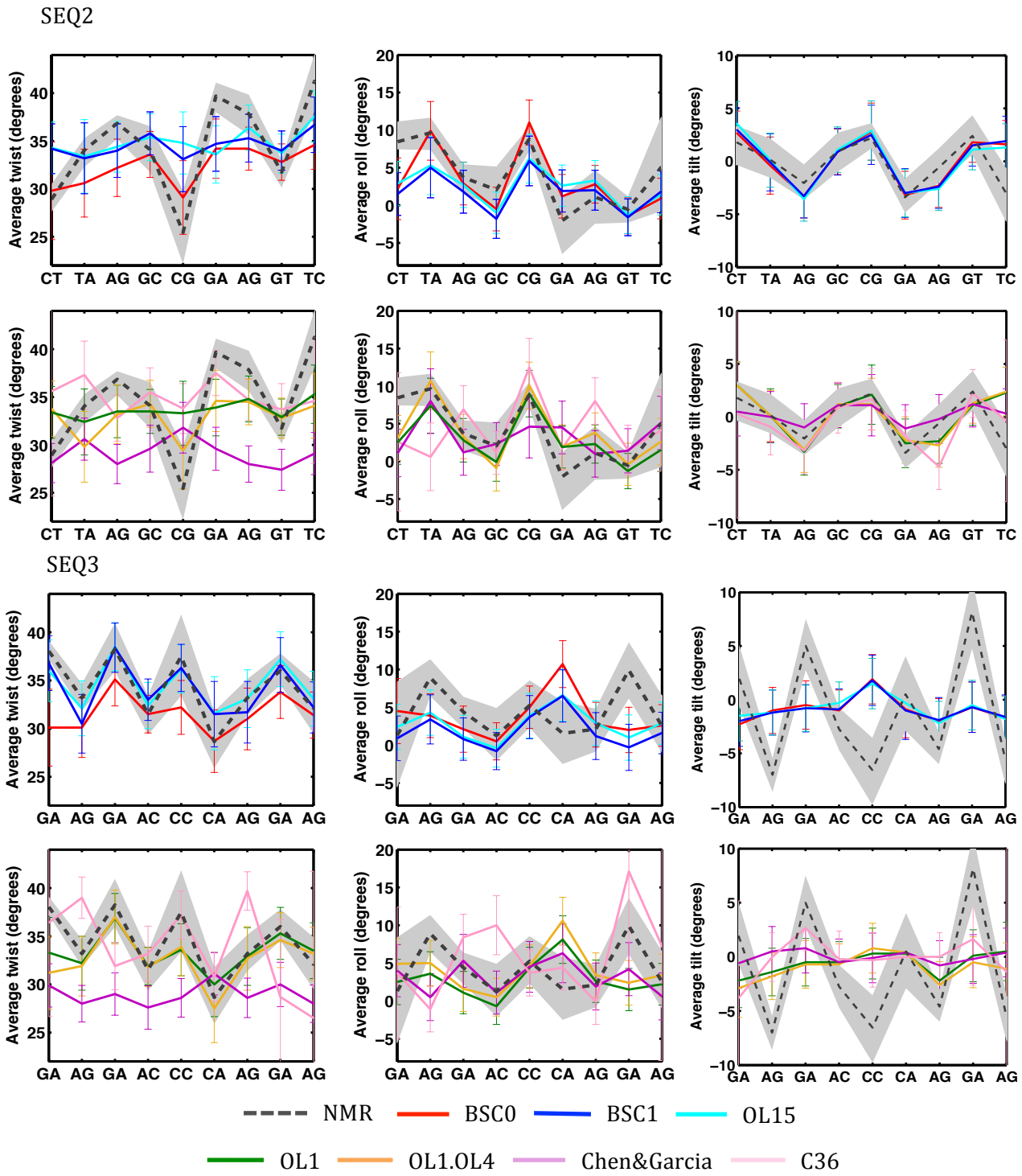



**Figure 4** | Averages and standard deviations of twist, roll and slide of SEQ2 and SEQ3. All the tested force-field are compared with the range of NMR structures refined using parmbsc1 OL15 (average in black dotted line) force-fields. The polarizable C36 force-field leads to complete corruption of the helix in the μs-scale simulation and results are not shown. See Figure 3 for structural models.

# Supporting Information

# HOW ACCURATE ARE ACCURATE FORCE-FIELDS FOR DNA?

Pablo D. Dans[1,2,&], Ivan Ivani[1,2,&], Guillem Portella[1,2,3], Adam Hospital[1,2], Carlos González[4,*] and Modesto Orozco[1,2,5,*]

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

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

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

[4] Instituto Química Física Rocasolano. Consejo Superior de Investigaciones Científicas. Madrid, Spain.

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

[&] Equally contributing authors.

[*] Correspondence to: Prof. Modesto Orozco (modesto.orozco@irbbarcelona.org) or Prof. Carlos González (cgonzalez@iqfr.csic.es).

**Table S1**. $^{1}$H-NMR assignments of **SEQ2:** d(GCTAGCGAGTCC)d(GGACTCGCTAGC) (25 mM sodium phosphate, 125 mM NaCl, pH 7, T = 25ºC).[+]

| SEQ2 | H1' | H2'/H2" | H3' | H4' | H5/Met/H2 | H6/H8 | Imino/amino |
|---|---|---|---|---|---|---|---|
| **G1** | 6.02 | 2.67/2.79 | 4.85 | 4.26 | --- | 7.99 | 12.76 |
| **C2** | 6.07 | 2.13/2.52 | 4.83 | 4.26 | 5.38 | 7.54 | 6.75/8.40 |
| **T3** | 5.57 | 2.15/2.43 | 4.88 | n.a. | 1.67 | 7.41 | 13.93 |
| **A4** | 6.04 | 2.75/2.90 | 5.05 | 4.40 | 7.40 | 8.22 | n.o. |
| **G5** | 5.67 | 2.47/2.59 | 4.95 | 4.36 | --- | 7.66 | 12.88 |
| **C6** | 5.58 | 1.78/2.26 | 4.77 | 4.09 | 5.18 | 7.17 | 6.36/8.22 |
| **G7** | 5.49 | 2.64/2.75 | 4.97 | 4.29 | --- | 7.82 | 12.75 |
| **A8** | 6.06 | 2.71/2.87 | 5.02 | 4.43 | 7.58 | 8.06 | n.o. |
| **G9** | 5.83 | 2.43/2.67 | 4.84 | 4.37 | --- | 7.50 | 12.88 |
| **T10** | 6.02 | 2.13/2.52 | 4.84 | 4.22 | 1.22 | 7.23 | 13.75 |
| **C11** | 6.07 | 2.22/2.49 | 4.84 | 4.17 | 5.69 | 7.60 | 7.01/8.50 |
| **C12** | 6.25 | 2.28 | 4.57 | 4.05 | 5.80 | 7.68 | 7.18/8.38 |
| **G13** | 5.65 | 2.48/2.66 | 4.82 | 4.17 | --- | 7.84 | n.o. |
| **G14** | 5.58 | 2.69/2.78 | 5.02 | 4.37 | --- | 7.84 | 12.91 |
| **A15** | 6.26 | 2.74/2.91 | 5.06 | 4.50 | 7.91 | 8.21 | n.o. |
| **C16** | 5.80 | 1.96/2.49 | 4.66 | n.a. | 5.19 | 7.27 | 6.76/7.97 |
| **T17** | 6.03 | 2.51 | n.a. | 4.18 | 1.50 | 7.37 | 13.88. |
| **C18** | 5.60 | 2.08/2.40 | 4.86 | 4.12 | 5.60 | 7.45 | 6.96/8.47 |
| **G19** | 5.88 | 2.66/2.71 | 4.98 | 4.37 | --- | 7.89 | 12.88 |
| **C20** | 5.86 | 1.98/2.44 | 4.72 | 4.17 | 5.30 | 7.36 | 6.66/8.18 |
| **T21** | 5.53 | 2.10/2.39 | 4.85 | n.a. | 1.64 | 7.38 | 13.88 |
| **A22** | 6.02 | 2.74/2.87 | 5.03 | n.a. | 7.41 | 8.21 | --- |
| **G23** | 5.80 | 2.47/2.64 | 4.94 | n.a. | --- | 7.68 | 12.92 |
| **C24** | 6.12 | 2.12/2.18 | 4.46 | 4.04 | 5.37 | 7.40xx | 6.63/8.22 |

[+] n.a. Not assigned. n.o. Not observed. No purine amino protons could be identified. Exchangeable protons resonance are given at T= 5ºC.

**Table S2**. $^{1}$H-NMR assignments of **SEQ3:** d(GGAGACCAGAGG)d(CCTCTGGTCTCC) (25 mM sodium phosphate, 125 mM NaCl, pH 7, T = 25ºC).[+]

| SEQ3 | H1' | H2'/H2" | H3' | H4' | H5/Met/ H2 | H6/H8 | Imino/amino |
|---|---|---|---|---|---|---|---|
| **G1** | 5.56 | 2.33/2.44 | 4.74 | 4.10 | --- | 7.725 | n.o. |
| **G2** | 5.37 | 2.71 | 4.96 | 4.30 | --- | 7.850 | 12.84 |
| **A3** | 5.98 | 2.67/2.85 | 5.05 | 4.41 | 7.50 | 8.114 | --- |
| **G4** | 5.52 | 2.56/2.69 | 4.99 | 4.36 | --- | 7.693 | 12.71 |
| **A5** | 6.19 | 2.61/2.88 | 4.99 | 4.46 | 7.80 | 8.084 | --- |
| **C6** | 5.78 | 1.95/2.39 | 4.76 | 4.14 | 5.15 | 7.197 | 8.05/6.57 |
| **C7** | 5.41 | 1.93/2.29 | 4.78 | 4.04 | 5.48 | 7.382 | 8.48/6.86 |
| **A8** | 5.88 | 2.68/2.81 | 5.01 | 4.34 | 7.70 | 8.152 | --- |
| **G9** | 5.37 | 2.51/2.63 | 4.96 | n.a. | --- | 7.707 | 12.70 |
| **A10** | 5.94 | 2.54/2.77 | 4.98 | 4.36 | 7.70 | 8.009 | --- |
| **G11** | 5.63 | 2.46/2.61 | 4.92 | 4.30 | --- | 7.595 | 12.70 |
| **G12** | 6.10 | 2.32/2.46 | 4.60 | n.a. | --- | 7.696 | 13.13 |
| **C13** | 6.01 | 2.24/2.57 | 4.78 | n.a. | 7.85 | 6.005 | 8.53/8.09[#] |
| **C14** | 6.05 | 2.19/2.57 | 4.77 | 4.24 | 5.73 | 7.727 | 7.91/6.40 |
| **T15** | 6.11 | 2.28/2.57 | 4.90 | 4.25 | 1.67 | 7.511 | 13.94[*] |
| **C16** | 6.00 | 2.10/2.49 | 4.80 | n.a. | 5.66 | 7.620 | 8.51/7.16 |
| **T17** | 5.67 | 2.00/2.38 | 4.84 | 4.09 | 1.65 | 7.284 | 13.89[*] |
| **G18** | 5.70 | 2.69/2.73 | 4.96 | 4.33 | --- | 7.873 | 12.81 |
| **G19** | 5.92 | 2.49/2.73 | 4.83 | 4.37 | --- | 7.621 | 12.85 |
| **T20** | 5.60 | 2.18/2.51 | 4.80 | n.a. | 1.26 | 7.261 | 13.753 |
| **C21** | 5.99 | 2.15/2.53 | 4.76 | n.a. | 5.58 | 7.593 | 8.37/7.06 |
| **T22** | 6.02 | 2.18/2.50 | 4.87 | 4.17 | 1.66 | 7.474 | 13.954 |
| **C23** | 6.13 | 2.31/2.48 | 4.86 | n.a. | 5.82 | 7.668 | 8.60/7.26 |
| **C24** | 6.23 | 2.36/2.31 | 4.54 | 4.10 | 5.99 | 7.850 | 8.53/8.09[#] |

[+] n.a. Not assigned. n.o. Not observed. No purine amino protons could be identified. Exchangeable protons resonance are given at T= 5ºC.

**Table S3** | Peak positions inferred from experimental and computational solution scattering profiles for the DDD sequence.[a]

| Peak / Structure | P1 [b] | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| **Exp**[c] | 0.456 | 0.750 | 1.127 | 1.513 | 1.834 |
| **1BNA** | --- | 0.750 | 1.145 | 1.520 | 1.900 |
| **1GIP** | 0.460 | 0.770 | 1.150 | 1.530 | 1.850 |
| **1NAJ** | --- | 0.726 | 1.084 | 1.420 | 1.738 |
| **Standard** | --- | 0.691 | 1.060 | 1.386 | 1.943 |
| **$NMR_{BSC0}$** | --- | 0.690 | 1.061 | 1.385 | 1.942 |
| **$NMR_{BSC0\text{-}OL15}$** | --- | 0.688 | 1.057 | 1.404 | 1.890 |
| **$NMR_{BSC1}$** | --- | 0.700 | 1.051 | 1.461 | 1.907 |
| **BSC0** | --- | 0.716 | 1.101 | 1.457 | 1.830 |
| **BSC1** | 0.410 | 0.720 | 1.110 | 1.510 | 1.910 |
| **BSC1 MD**[d] | 0.448 ± 0.014 | 0.725± 0.042 | 1.083 ± 0.055 | 1.506 ± 0.017 | 1.864 ± 0.013 |
| **$BSC0_{OL1}$** | --- | 0.723 | 1.114 | 1.457 | 1.817 |
| **$BSC0_{OL1.OL4}$** | --- | --- | 1.115 | 1.470 | --- |
| **$BSC0_{OL15}$** | 0.440 | 0.703 | 1.112 | 1.501 | 1.968 |
| **Cheng-Garcia** | --- | 0.818 | --- | 1.594 | --- |
| **C36** | 0.474 | --- | 1.118 | 1.499 | 1.798 |
| **C36 MD**[e] | 0.442 | 0.800 | 1.101 | 1.478 | 1.829 |

[a]Values are reported in $Å^{-1}$. [b]Peak positions were determined from zero crossing points in the first derivative [Savelyev A., MacKerell Jr. D. *J. Phys. Chem. Letter***2015**, 6, 212]. [c] The data to produce the experimental curve was a courtesy of Prof. David Tiede[Zuo,X., Cui,G., Merz,K.M., Zhang,L., Lewis,F.D. and Tiede,D.M. (2006) X-ray diffraction "fingerprinting" of DNA structure in solution for quantitative evaluation of molecular dynamics simulation. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 3534–9.]. [d] Taken from an independent larger ensemble by Dans et al, NAR 2016. [e] Taken from Savelyev A., MacKerell Jr. D. *J. Phys. Chem. Letter***2015**, 6, 212.

**Table S4** | Summary of NOE distance violations from NOE-restrained MD simulations with the reference force-fields (BSC1, BSC0, $BSC0_{OL15}$) using the NMR data obtained in-house (see Methods).[a]

| Force-field | Nº of violations | Largest violation (Å) | Average violation (Å) |
|---|---|---|---|
| | | SEQ2 | |
| **$NMR_{BSC1}$** | 6 | 0.58 | 0.41 |
| **$NMR_{BSC0}$** | 6 | 0.56 | 0.41 |
| **$NMR_{BSC0\text{-}OL15}$** | 6 | 0.57 | 0.39 |
| | | SEQ3 | |
| **$NMR_{BSC1}$** | 6 | 0.58 | 0.43 |
| **$NMR_{BSC0}$** | 7 | 0.58 | 0.42 |
| **$NMR_{BSC0\text{-}OL15}$** | 6 | 0.58 | 0.44 |

[a] We considered an experimental restraint violated when its average penalty energy was above 3·kT. See the footnote comment to Table 2 and the Method section for additional details.

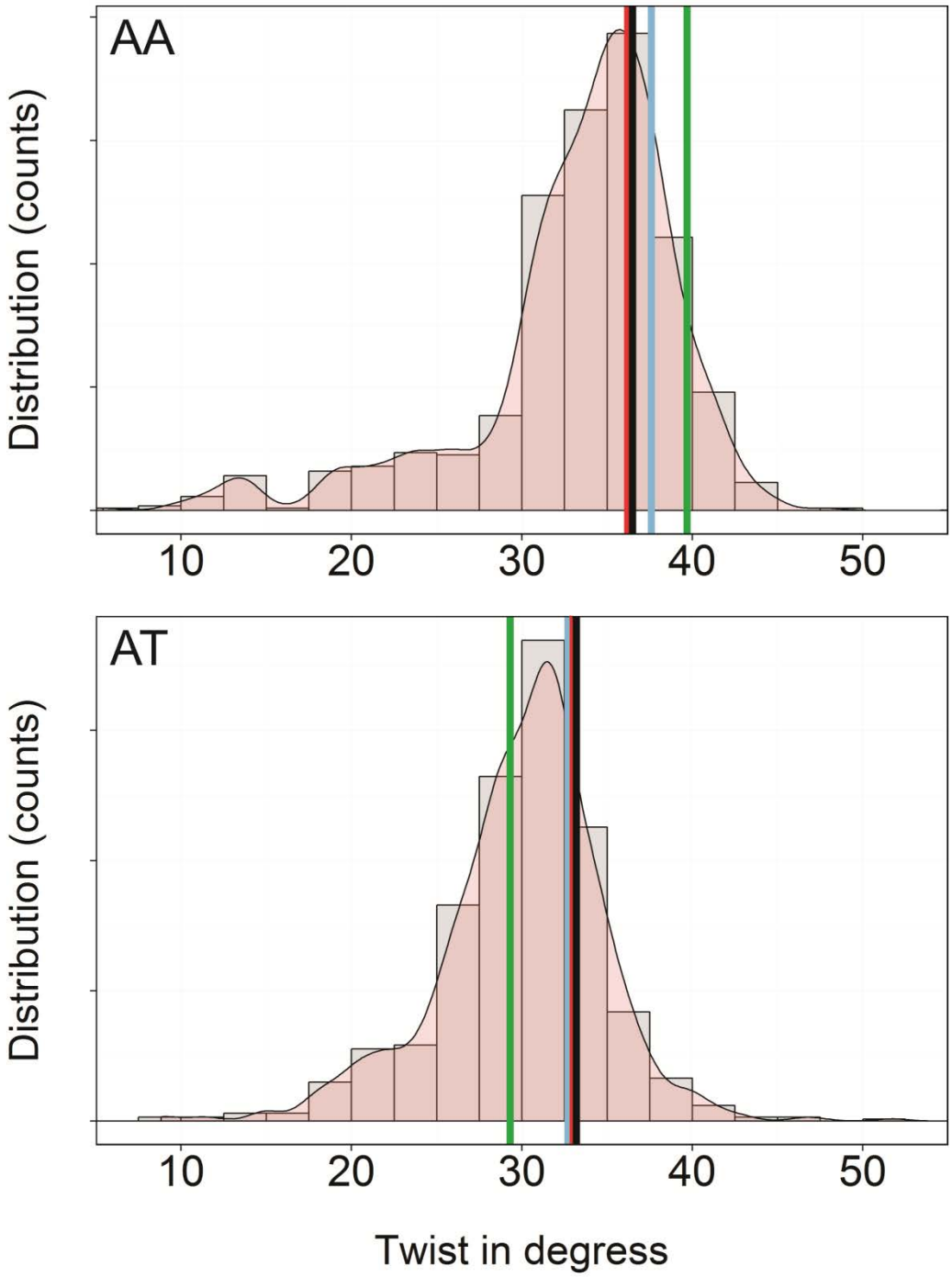



**Figure S1** | Twist distribution for AA (TOP) and AT (BOTTOM) dinucleotide steps found in X-ray structures of naked-DNA and Protein-DNA complexes (Dans, NAR 2012). Vertical lines represent the average values computed from NMR-biased MD simulations of the DDD sequence: BSC0-NOE (red), BSC1-NOE (black), and $BSC0_{OL15}$-NOE (blue). The Standard procedure with the default BSC0 force-field is shown in green.

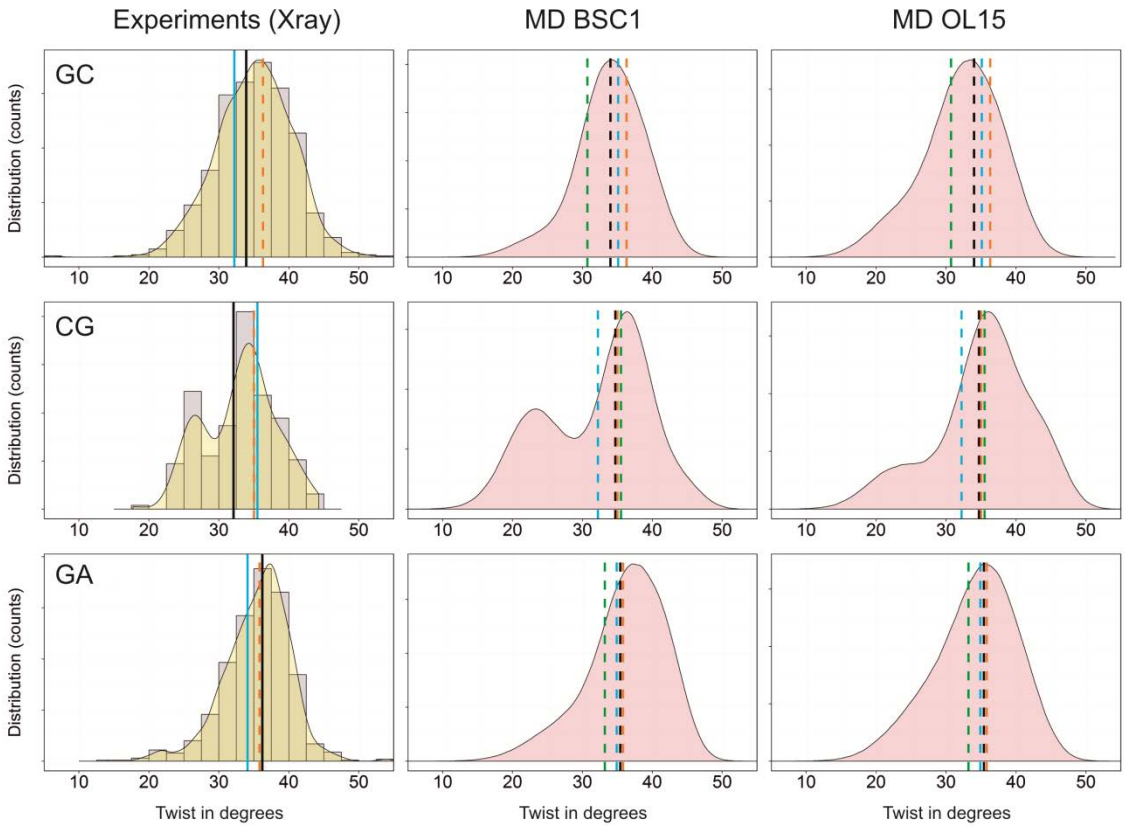



**Figure S2 |** Twist distributions of selected bps from X-ray experiments and MD simulations for DDD. FIRST COLUMN: Grey bars (and the smoothed density in yellow) represent the X-ray structures of naked-DNA ($C_3G_4$ bps)[1], naked-DNA plus DNA-protein complexes ($G_2C_3$ and $G_4A_5$ bps)[1], and the NMR structure with PDB code 1NAJ (dashed line in orange). The average values from MD simulations are depicted in black (BSC1), and light-blue (OL15). SECOND COLUMN: Distributions of MD simulations obtained with BSC1. Dashed vertical lines represent the experimental (NMR) values: NMR-BSC1 (black), NMR-OL15 (light-blue), 1NAJ (orange), and 1GIP (green). THIRD COLUMN: Distributions of MD simulations obtained with OL15.
[1] Dans *et al*, NAR 2012

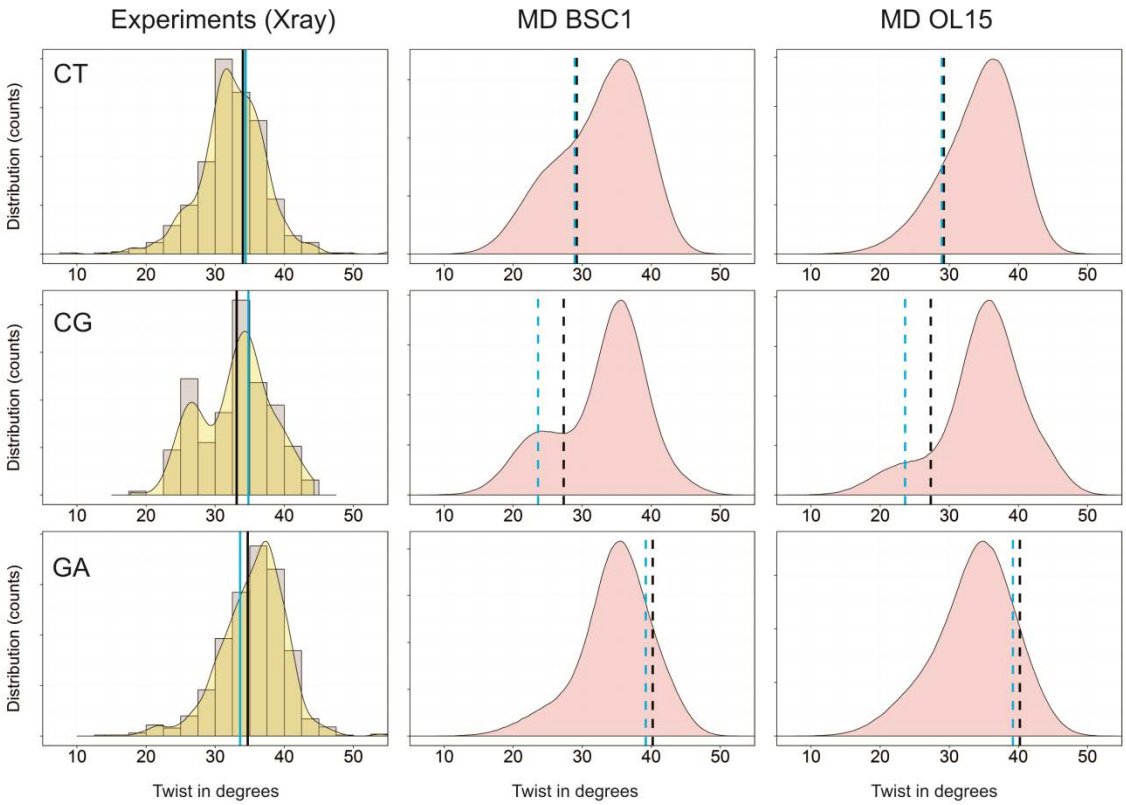



**Figure S3 |** Twist distributions of selected bps from X-ray experiments and MD simulations for SEQ2. FIRST COLUMN: Grey bars (and the smoothed density in yellow) represent the X-ray structures of naked-DNA ($C_6G_7$ bps)[1], and naked-DNA plus DNA-protein complexes ($C_2T_3$ and $G_7A_8$ bps)[1]. The average values from MD simulations are depicted in black (BSC1), and light-blue (OL15). SECOND COLUMN: Distributions of MD simulations obtained with BSC1. Dashed vertical lines represent the experimental (NMR) values: NMR-BSC1 (black), and NMR-OL15 (light-blue). THIRD COLUMN: Distributions of MD simulations obtained with OL15.

[1] Dans *et al*, NAR 2012

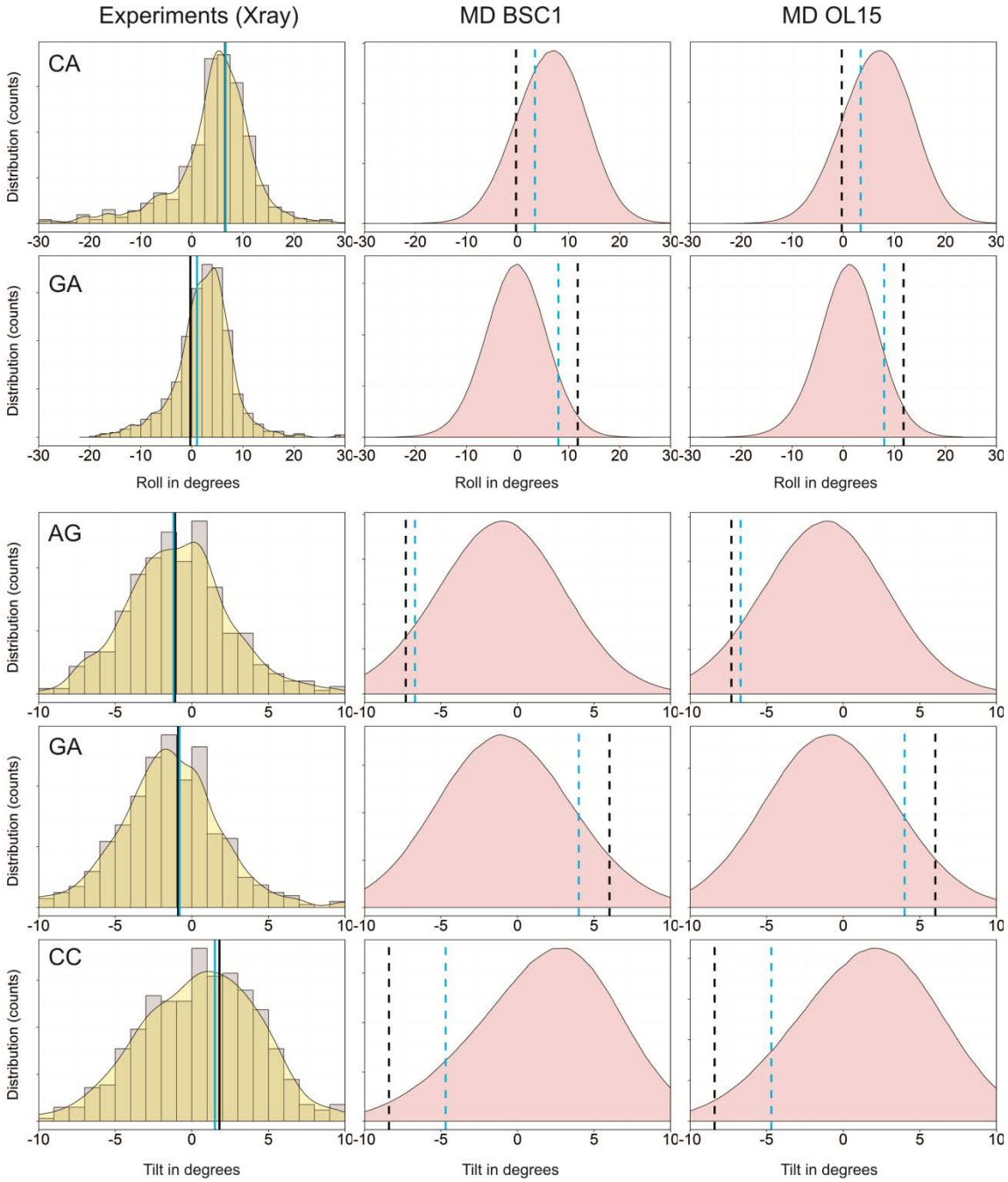



**Figure S4 |** Roll and Tilt distributions of selected bps from X-ray experiments and MD simulations for SEQ3. FIRST COLUMN: Grey bars (and the smoothed density in yellow) represent the X-ray structures of naked-DNA plus DNA-protein complexes ($C_7A_8$, $G_9A_{10}$, $A_3G_4$, $G_4A_5$, and $C_6C_7$ bps)[1]. The average values from MD simulations are depicted in black (BSC1), and light-blue (OL15). SECOND COLUMN: Distributions of MD simulations obtained with BSC1. Dashed vertical lines represent the experimental (NMR) values: NMR-BSC1 (black), and NMR-OL15 (light-blue). THIRD COLUMN: Distributions of MD simulations obtained with OL15.
[1] Dans *et al*, NAR 2012

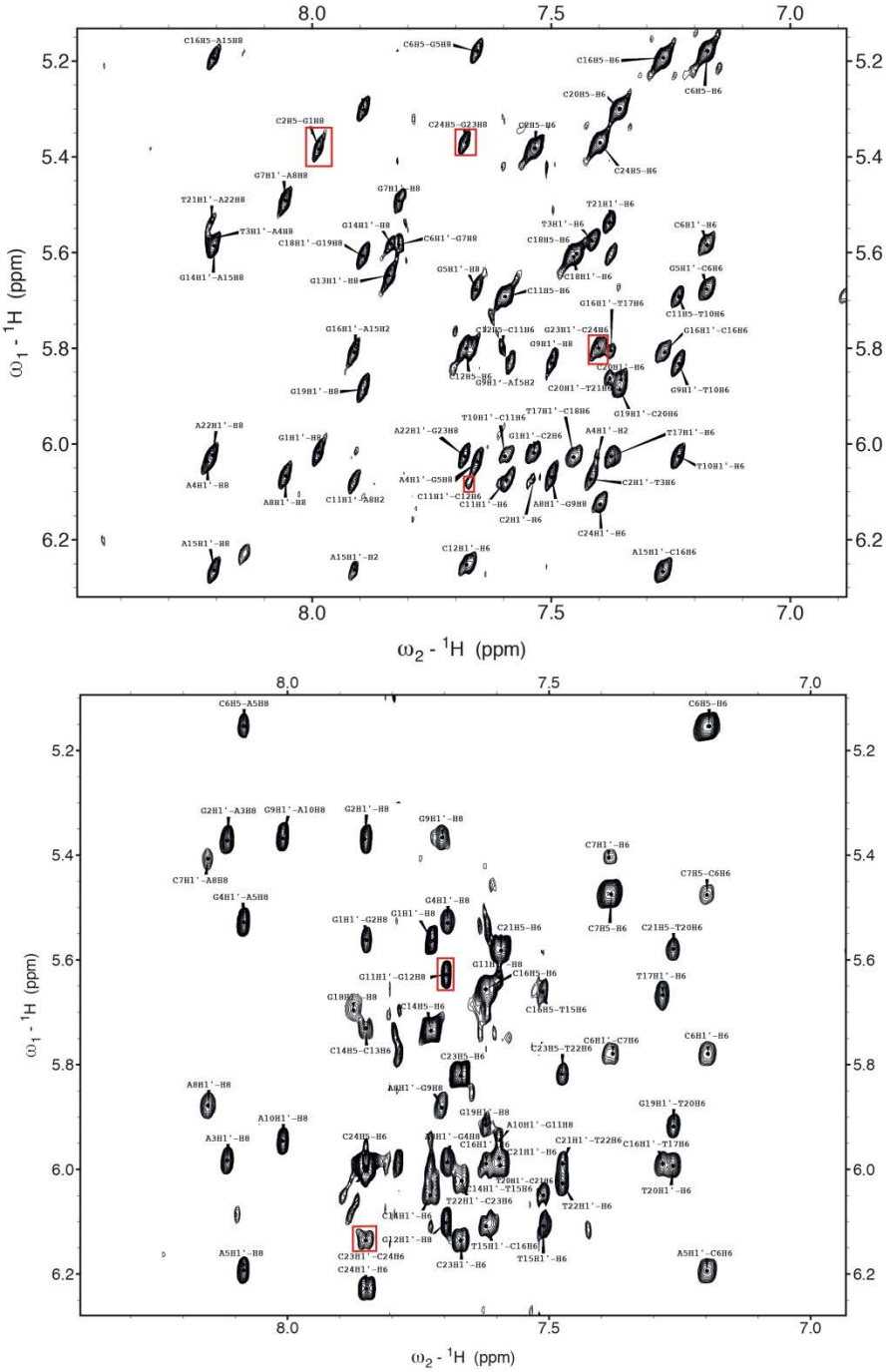



**Figure S5 |** H1´-aromatic region of the NOESY spectra of SEQ3 (top) and SEQ3 (bottom) (mixing time 200 ms, buffer conditions 125 mM NaCl, 25 mM sodium phosphate, pH 7, T = 25ºC). Some relevant cross-peaks involving terminal residues are labelled in red colour. For similar analysis on DDD.

# Bibliography to Chapter 4

Beveridge, D.L. et al., 2004. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophysical journal*, 87(6), pp.3799–13.

Cheatham, T.E.I.I.I. et al., 1995. Molecular dynamics simulations on solvated biomolecular systems: the particle mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *Journal of the American Chemical Society*, 117(14), pp.4193–4194.

Cheatham III, T.E., Cieplak, P. & Kollman, P.A., 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 16(4), pp.845–862.

Chen, A.A. & García, A.E., 2013. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 110(42), pp.16820–16825.

Dans, P.D. et al., 2016. Multiscale simulation of DNA. *Current Opinion in Structural Biology*, 37, pp.29–45.

Drsťata, T. et al., 2012. Structure, stiffness and substates of the Dickerson-Drew dodecamer. *Journal of chemical theory and computation*, 9(1), pp.707–721.

Fadrná, E. et al., 2009. Single stranded loops of quadruplex DNA as key benchmark for testing nucleic acids force fields. *Journal of Chemical Theory and Computation*, 5(9), pp.2514–2530.

Hart, K. et al., 2011. Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *Journal of chemical theory and computation*, 8(1), pp.348–362.

Heddi, B. et al., 2008. Importance of accurate DNA structures in solution: the Jun–Fos model. *Journal of molecular biology*, 382(4), pp.956–970.

Heddi, B. et al., 2006. Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *Journal of the American Chemical Society*, 128(28), pp.9170–9177.

Krepl, M. et al., 2012. Reference simulations of noncanonical nucleic acids with different χ variants of the amber force field: Quadruplex dna, quadruplex rna, and z-dna. *Journal of chemical theory and computation*, 8(7), pp.2506–2520.

Levitt, M., 1983. Computer simulation of DNA double-helix dynamics. In *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, pp. 251–262.

Orozco, M. et al., 2003. Theoretical methods for the simulation of nucleic acids. *Chemical Society Reviews*, 32(6), pp.350–364.

Perez, A. et al., 2008. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic acids research*, 36(7), pp.2379–2394.

Pérez, A. et al., 2005. Exploring the essential dynamics of B-DNA. *Journal of Chemical Theory and Computation*, 1(5), pp.790–800.

Pérez, A., Marchán, I., et al., 2007. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophysical journal*, 92(11), pp.3817–29.

Pérez, A., Luque, F.J. & Orozco, M., 2007. Dynamics of B-DNA on the microsecond time scale. *Journal of the American Chemical Society*, 129(47), pp.14739–14745.

Pérez, A., Luque, F.J. & Orozco, M., 2012. Frontiers in molecular dynamics

simulations of DNA. *Accounts of chemical research*, 45(2), pp.196–5.

Savelyev, A. & MacKerell, A.D., 2014. All-atom polarizable force field for DNA based on the classical drude oscillator model. *Journal of computational chemistry*, 35(16), pp.1219–1239. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4075971&tool=pmcentrez&rendertype=abstract [Accessed February 11, 2016].

Várnai, P. & Zakrzewska, K., 2004. DNA and its counterions: a molecular dynamics study. *Nucleic acids research*, 32(14), pp.4269–4280.

Yang, C., Kim, E. & Pak, Y., 2015. Free energy landscape and transition pathways from Watson-Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Research*, 43(16), pp.7769–7778.

Zgarbová, M. et al., 2013. Toward improved description of DNA backbone: revisiting epsilon and zeta torsion force field parameters. *Journal of chemical theory and computation*, 9(5), pp.2339–2354.

Zgarbová, M. et al., 2011. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation*, 7(9), pp.2886–2902.

Zgarbová, M. et al., 2015. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *Journal of Chemical Theory and Computation*, 11(12), pp.5723–36.

*"The merit of all thing lies in their difficulty."*
*Alexandre Dumas*

# 5| RNA WORLD

As we saw in Chapter 1.3, the diversity of RNA structures is much larger than of DNA. Despite small chemical differences between DNA and RNA, nature has completely separated their functional spaces, selecting DNA as the primary carrier of genetic information, while giving RNA a myriad of other functions.

Despite the interest of understanding RNA structural properties, RNA simulations are still in the infancy compared to DNA. The main reason is that, regardless of recent intense efforts of some groups, RNA force fields are still far in accuracy from those of DNA. Recent efforts to improve RNA force fields included Turner's group who used NMR data on small systems to tune χ torsion (Yildirim et al. 2011), Bussi's group (Gil-Ley et al. 2016) followed similar strategies to correct some of the current AMBER torsional terms. Zgarbová and coworkers used high-level QM to refine χ torsion (called OL3 correction) (Zgarbova et al. 2011), while Chen and Garcia, additionally to χ refinement, scaled base stacking to successfully fold RNA hairpins (Chen & García 2013). Recently, Shaw's group has embarked in an aggressive project to implement a variety of corrections into RNA force field (Pianna 2016). None of these force fields provide, however good representations of complex model RNA systems (Bergonzo et al. 2015, unpublished data from the group).

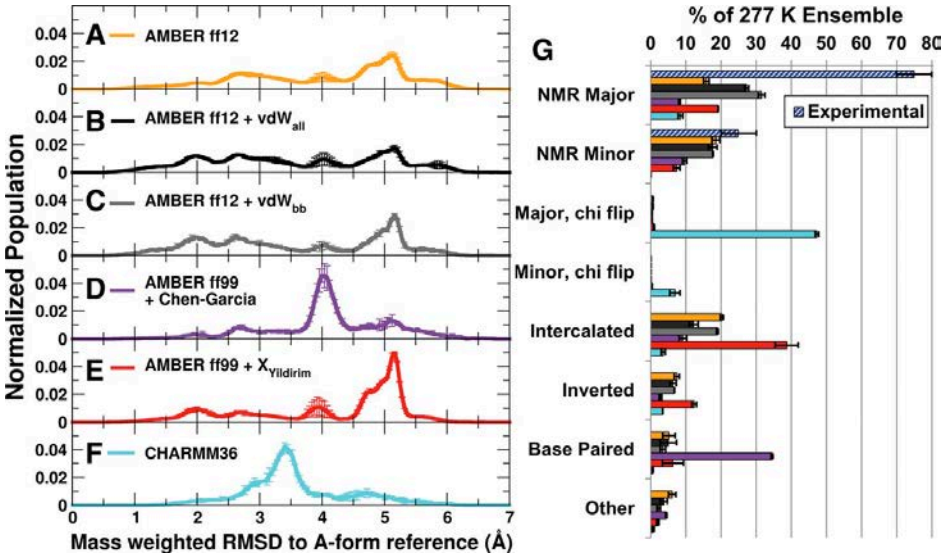



**Figure 5.1. Benchmark of RNA force field on a small tetreanucleotide.** (Left) M-REMD RMSD to A-form Reference for r(GACC). (*A*) ff12, (*B*) ff12 + vdW$_{all}$, (*C*) ff12 + vdW$_{bb}$, (*D*) ff99 + Chen–Garcia, (*E*) ff99 + $\chi_{Yil}$, and (*F*) C36 force fields. The averages between two runs per force field are shown, with error bars shown as the standard deviation. RMSDs corresponding to NMR major and minor structures are ~2.0 Å and 2.6 Å, respectively. (Right) Populations of major conformation types seen in cluster analysis of each M-REMD simulation. Bars indicate the average and standard deviation between two independent runs and are colored by force field to match the RMSD histograms. Experimental values are shown on the table in a blue striped pattern (taken from (Bergonzo et al. 2015)).

The poorer performance of RNA force field compared to that of last-generation DNA force fields might surprise a non-expert, but despite the chemical similarity, DNA and RNA are structurally very different. Thus, while DNA structural universe is dominated by the double helix, the RNA samples much more complex conformations. This means that a force field for RNA should be able to represent a much wider conformational space than that expected for DNA. Another important aspect of complexity is the presence of the 2'-OH group, which dramatically increases the relationship between the different degrees of freedom (P Auffinger & Westhof 1997), adding more difficulties for the parameterization. As shown later, the problem is not only the inclusion of one more degree of freedom, but its complex coupling with all the remaining torsions of the oligonucleotide (Denning et al. 2011). We combined a variety of computational techniques (database analysis, atomistic MD simulations, high-level QM and hybrid QM/MM calculations) to explore in detail the conformational preference of the $C_{2'}$-OH bond and its impact in RNA conformation. We found that 2'-OH orientation in big extent determines the RNA conformation and most probably serves as the molecular switch to modulate protein induced-fit mechanisms (see **Chapter 5.1**). Furthermore, we unrevealed the complex coupling of this torsion with the other degrees of freedom of RNA, and unexpected transferability

problems that force-us to move out of the standard combinatorial rules of van der Waals interactions.

In a more practical study, we design a degradation resistant synthetic siRNA that could be used in gene regulation based on a RISC-dependent mechanism. Previous results obtained in the group suggested that a new class of 3'-exonuclease-resistant modification of the siRNA duplex structures, called dumbbell, showed a resistance to nuclease digestion (Terrazas et al. 2013). The approach consisted in replacing the 3'-terminal natural dinucleotide overhangs with dimeric N-ethyl-N bridged nucleosides (called BCn dimer). The aim of this work was to computationally design the length of the linker, predict its flexibility and ability to mimic the wild-type siRNA duplex. This study showed promising results where one dumbbell structure (with BC6-linker) showing higher biostability than other synthetic siRNA described in the literature, which could be used in breast cancer therapy.

We are now in the validation-stage of the RNA version of the parmbsc1 force field.

## 5.1 $C_{2'}$-OH study (Publication 4)

We saw in Chapter 1.3, small chemical differences between DNA and RNA lead to a big differences in functionality of the two molecules. The consensus idea is that the presence of the 2'-OH has an impact on the sugar, driving its conformational change from South to North and global change from the B- to the A- form (Soliva et al. 1999). However, our knowledge of the rotational states of $C_{2'}$-$O_{2'}$ and its connection to the local and global conformations is still rather limited. Generally, one defines $C_{2'}$-$O_{2'}$ in term of its orientation towards $H_{2'}$ also known as $\varkappa = (H_{2'} - C_{2'} - O_{2'} - H_{2''})$. Previous studies of the 2'-OH rotation suggested that there are three preferred orientations pointing towards: $O_{3'}$ atom ($\varkappa$ in *gauche*+), $O_{4'}$ atom ($\varkappa$ in *trans*) or the nucleobase ($\varkappa$ in *gauche*-). NMR and QM data point that orientations towards $O_{3'}$ and nucleobase are the most frequent (Fohrer et al. 2006; Mládek et al. 2014) and dependent on the orientation specific hydration. A-form is stabilized by the water interaction with 2'-OH oriented toward the base, while non-canonical conformations benefit from water interaction with the 2'-OH pointing towards $O_{3'}$ atom (Denning & MacKerell Jr 2012).

To better understand the mechanism of how 2'-OH impacts the RNA conformation, we combined database analysis, high-level QM and hybrid QM/MM calculations, and atomistic MD, providing evidence that 2'-OH orientation is a main determinant of the local and global structure of RNA. We performed a QM scan of the pseudorotational angle of the ribose for the three orientation of $\varkappa$, which suggested that in North puckering, $O_{3'}$ orientation is the most favorable orientation with the base orientation being close in energy (~ 0.5 kcal/mol), while in South puckering, $O_{4'}$ orientation is the most stable orientation. Moreover, $O_{4'}$ orientation in South puckering is overall the most stable sugar puckering state, suggesting that 2'-OH group can induce changes in sugar puckering, which can later lead to global changes in overall structure of the RNA.

This work is in the process of publication and here we present the manuscript.

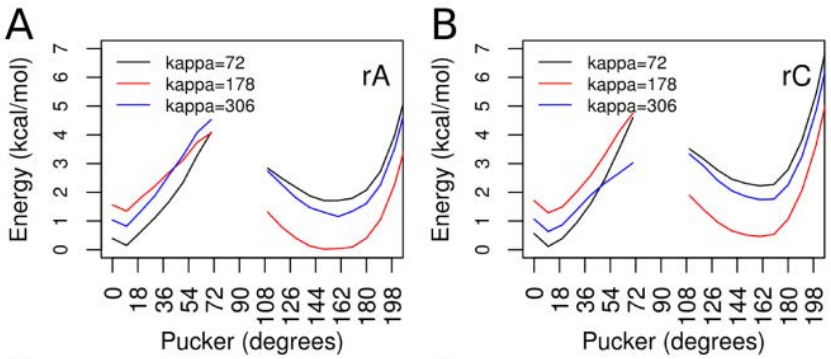



**Figure 5.2. QM scan of the pseudorotational angle in three orientation of 2'-OH for adenosine (A) and cytosine (B).** The dependence of $\chi$ torsion on the sugar puckering was taken into account by fixing $\chi$=190º for North and North-East puckers, and $\chi$=230º for South and South-East puckers.

# *Small Details Matter: the Importance of the "Innocuous" 2'Hydroxyl*

Leonardo Darré[1,2], Ivan Ivani[1,2], Pablo D. Dans[1,2], Hansel Gómez[1,2], Adam Hospital[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. Barcelona, Spain

[3]Department of Biochemistry, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain

*KEYWORDS: RNA, 2'OH, pucker, QM/MM, data mining, molecular simulations.*

**ABSTRACT:** Despite the few chemical differences existing between DNA and RNA, the two polymers have different structures and play completely different roles in the cell. While DNA forms very long and regular double helices which carry the genetic information, RNA can form very complicated and conserved 3D structures displaying a large variety of functions, others that being an intermediate in expressing genetic information. Despite decades of work, the origins of the structural and functional differences between DNA and RNA are still obscure, and general belief is that differences emerge exclusively from the different sugar puckering of the ribose and the 2'deoxyribose. By combining database analysis with extensive molecular dynamics, quantum mechanics, and hybrid QM/MM simulations, we provide direct evidence on the dramatic role of the 2'OH group as a tunable conformational switch for RNA, as one of the main determinant of differential DNA/RNA properties, and as a key element in modulating RNA-protein recognition.

## *INTRODUCTION*

There is general consensus that life originated in an RNA-world, as this oligonucleotide is a very versatile entity that is able to self-replicate, transmitting information to descendants, and at the same time adopt complex three dimensional structures acting as catalyzers of complex reactions. However, at an early point of evolution, DNA was selected as the primary carrier of genetic information, while RNA maintained a myriad of other functions, the most important ones related to translating DNA information into protein sequence. While the DNA has a very simple conformational landscape, dominated by a right-handed double helix, the RNA can display very complex three-dimensional structures, some of them, such as the transfer or ribosomal RNAs, exquisitely refined by evolution. (Caetano-Anollés & Caetano-Anollés 2015; Petrov et al. 2015; Petrov et al. 2014; Saint-Leger et al. 2016; Zhang & Ferré-D'Amaré 2016)

Despite their coexistence in some cellular organelles, nature has completely separated DNA and RNA functional spaces, something quite surprising considering the minuscule chemical differences between them: the presence/absence of one methyl group at position five of uridine, and the presence/absence of a hydroxyl at position 2' of the sugar. The consensus idea is that the presence of the 2'OH drives the puckering preferences of the sugar from South (S, C2'endo) to North (N, C3'endo) conformation, which is known to drive a global conformational change from the B- to the A- form.(Soliva et al. 1999) However, our understanding of the connection between the rotational state of the C2'-O2' bond and the local and global conformation of the RNA is still rather limited. In the A-form (sugar in North), the 2'OH could adopt three preferred orientations pointing toward: the O3' atom (*gauche+* measured from H2'), the nucleobase (*gauche-*), or the O4' atom (*trans*).(Pascal Auffinger & Westhof 1997) The first two being the most frequent ones according to NMR(Fohrer et al. 2006) and QM calculations,(Mládek et al. 2014) and subject to orientation specific hydration. When water interacts with the 2'OH in the base orientation the A-form is stabilized,(Egli et al. 1996; Fohrer et al. 2006) while non-canonical conformations gain stabilization from water molecules interacting with the 2'OH in the O3' orientation.(Denning & MacKerell Jr 2012) The less frequent South sugar conformation has received less attention but is believed to favor a 2'OH oriented mainly toward the O3' atom but with a C2'-O2' torsion shifted to *trans* orientation.(Pascal Auffinger & Westhof 1997; Mládek et al. 2014)

In the present work, a combination of database analysis, atomistic molecular dynamics (MD), high level quantum mechanical (QM), and hybrid quantum

mechanics/molecular mechanics (QM/MM) calculations was used to explore in detail the conformational preferences of the C2'OH bond and its impact in RNA conformation. We found that, the 2'OH rotation is a major determinant of RNA conformation, and a molecular switch, which can be tuned by proteins and other effectors to control the entire RNA structure.

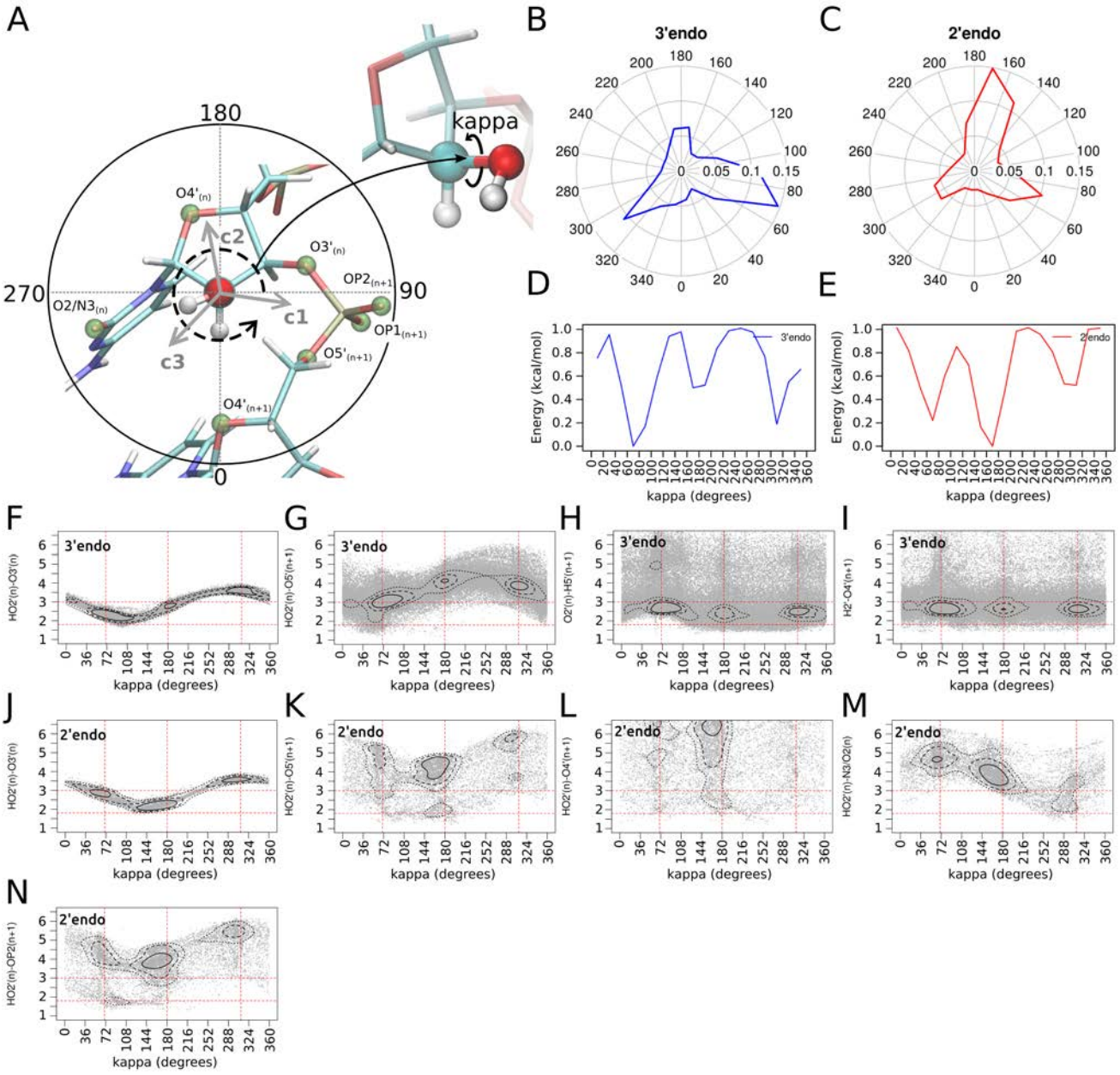



Figure 1. Kappa torsion preferred orientations from the Protein Data Bank. (A) Nucleotide in a RNA strand indicating the possible orientations of the 2'OH, and the local of hydrogen bond acceptors/donors. (B) Probability distribution of the torsion angle between the atoms H2'-C2'-O2'-HO2' for all the 3'endo ribonucleotides of the RNA dataset obtained from the current state of the PDB. (C) Same as in (A) but for 2'endo ribonucleotides. (D,E) Empirical free energy calculated from the experimental kappa distributions in (B) and (C), respectively. Scatter plots of kappa torsion vs distance between HO2' and local acceptors/donors of hydrogen bonds, are shown for nucleotides with pucker phase in 3'endo (F-I), and 2'endo (J-N). Red dotted lines indicate optimal and maximum hydrogen bond distances (horizontal), and kappa rotation minimum energy positions (vertical). Contour lines correspond to points with density values equal to the average density plus 1, 2 and 4 standard deviations.

## METHODS

**Database mining**. All NMR-solved RNA structures deposited in PDB were analyzed (see Supplementary Methods 1), accounting for a total of 174,511 2'OH groups, 115,513 with the ribose in *North* puckering (0<=pucker phase<=36) and 11,626 with the ribose in the unusual *South* puckering (144<=pucker phase<=180). The orientation around the C2'-O2' torsion (herein called

kappa, κ) was defined using the atoms H2'-C2'-O2'-HO2', following Auffinger and Westhof ideas(Pascal Auffinger & Westhof 1997) (see Figure 1A). In order to analyze the potential role of the 2'OH group in modulating protein-RNA interactions and the connection with the RNA local conformation we performed additional analysis using only RNA-protein complexes solved again by NMR. Additionally we explored heavy atom contacts involving the 2'OH group considering not only NMR, but also X-ray (resolution <=2.5 Å) protein-RNA complex structures, which means exploration of 26,760 2'OH groups from 500 PDB entries). Additional details of the database analysis can be found in Supplementary Methods 1. To double-check the observations made from the data sets mentioned above, the same analysis was repeated using a non-redundant database(Leontis & Westhof 2012) containing both NMR and X-ray (resolution<=2.5 Å) solved structures (see Supplementary Tables 1-3 for details).

**Quantum simulations**. The pseudo-rotational profile of ribose was first explored along the North <-> East <-> South transition path. To avoid discontinuities in the energy profiles, geometry optimizations at each point were performed keeping β, γ, ε and χ torsions at their standard values in RNAs. In the case of χ, the dependence on the sugar puckering was taken into account setting $\chi_N$=190 degrees and $\chi_S$=230 degrees. To explore whether or not pseudo-rotation was dependent on the orientation of the C2'O2' bond, profiles were calculated fixing the κ angle at three typical values (72, 178, 306 degrees; the most populated values found in our database analysis). Energy profiles were obtained at the B3LYP/6-31++G(d,p) level and selected points were refined at the MP2/aug-cc-pVDZ level. All profiles were obtained in water as simulated by the IEFPCM continuum method.(Marenich et al. 2009)

Analysis of electron distribution using Bader's atoms in molecules (AIM) theory(Bader 1998; Bader 1991; Bader 1994) was performed on reduced clusters representative of the most prevalent orientations of the k angle (three replicas per relevant κ orientation). Single point calculations at MP2(FC)/6-31G(d,p) level were performed at the dinucleotide level, removing the base at 3' and completing the valence of the C1', O5' and O3' atoms with H atoms. This analysis allowed us to explore the potential formation and intensity of canonical O-H-X (for X=O) or non-canonical O-H-X (for X=C) hydrogen bonds by searching for bond critical points connecting such atoms and quantifying the associated electron density. The AIM-UC package(D & D 2014) was used for the AIM analysis.

**Classical simulations**. A large variety of MD simulations were performed to analyze the connection between RNA and C2'O2' conformation. They include i) standard simulations in hairpin and kissing loop RNA motives ; ii) potentials of mean force of the κ rotation at the nucleoside (rC) and dinucleotide (rCpC) levels using Umbrella Sampling (US) with a 18 degrees interval grid of the κ torsional or pseudo-rotation space (500 ps equilibration and 2.5 ns of averaging per window); and iii) hamiltonian-replica exchange molecular dynamics (H-REMD) to evaluate the conformational landscape of two small RNA tetra-nucleotides (rGACC and rCCCC) for which experimental structural data in solution is available.(Tubbs et al. 2013; Yildirim et al. 2011) All calculations were performed using the parm99 force field(Cheatham III et al. 1999; Cornell et al. 1995) supplemented with the bsco(Pérez et al. 2007) and chiOL3(Zgarbová et al. 2011; Banáš et al. 2010) modifications for RNA; some control simulations were performed with a local experimental RNA-version of the parmbsc1 force-field.(Ivani et al. 2015) Electro-neutrality was achieved by adding K+ and extra K+Cl- to generate a 150 mM concentration (taking Dang's parameters(Dang & Kollman 1995; Dang 1995; Smith & Dang 1994) to represent ions). Additional details of classical simulations can be found in Supplementary Methods 2.

**QM/MM simulations**. Hybrid QM/MM simulations were used extensively to analyze the free energy profile of the C2'O2' rotation for an isolated rC nucleoside and a r(CpC) dinucleotide in aqueous solution. BLYP/6-31G(d) functional was used to represent the nucleic acid, while the solvent was represented at the classical level. Umbrella sampling free energy profiles were computed by scanning in 18 degrees intervals the κ torsional space (5 ps equilibration and 40 or 25 ps of averaging per window for the nucleoside rC or the dinucleotide rCpC, respectively). Extended descriptions of QM/MM simulations can be found in Supplementary Methods 3.

## RESULTS AND DISCUSSION

**C2'O2' torsion Experimental distribution.** The 2'OH group of a ribose in the major **North** conformation (ratio N:S is around 10:1 in the database) samples three rotational states in NMR-PDB (Figure 1B): i) the κ region between 40 and 140 degrees (peak at ~70 degrees; *conformer 1*), ii) the κ region between 140 and 240 degrees (peak at ~180 degrees; *conformer 2*), and iii) the κ region between 240 and 40 degrees (peak at ~310 degrees; *conformer 3*). Transforming populations into conformational free energies (Figure 1D) points to a nearly barrier-less rotation, with three minima of free energies of 0 (*conformer 1*), ~0.5 (*conformer 2*) and ~0.2 kcal/mol (*conformer 3*). Interestingly, no significant differences are found in the κ torsional distribution for the four ribonucleotides (Supplementary Figure 1), suggesting that base-sugar contacts are not crucial to determine the 2'OH group orientation (see below). Contact analysis reveals some interactions that appear in all the conformations of this set of structures, such as a non-canonical $C5'_{(n+1)}$-$H5'_{(n+1)}$···O2' hydrogen bond and the non-canonical C2'-H2'···$O4'_{(n+1)}$ hydrogen bond previously reported by Auffinger and Westhof,(Pascal Auffinger & Westhof 1997) while others like the strong O2'-HO2'···O3' hydrogen bond appear only in *conformer 1* (Figure 1F-I). Close contacts between the 2'OH group and the nucleobase, or the OP1/2 groups are uncommon in experimental structures of **North** riboses (Supplementary Figure 2). *Conformer 2*, which was the least populated orientation for **North** puckering, becomes dominant for **South** riboses, probably due to the formation of O2'-HO2'···O3' hydrogen bonds (Figure 1J-N). *Conformer 1*

instead becomes the second most populated orientation and conformer 3 the least populated one (see Figure 1C,E). This is reflected on the relative free energy difference between the conformers 1, 2 and 3 (~0.2, 0 and ~0.5 kcal/mol respectively; see Figure 1E). Some variability (~0.1 kcal/mol) is observed in the relative energy values depending on partitioning of the k coordinate, in particular for the **South**-puckering profile (Supplementary Figure 4C,D), however the overall trend remains consistent. Further-

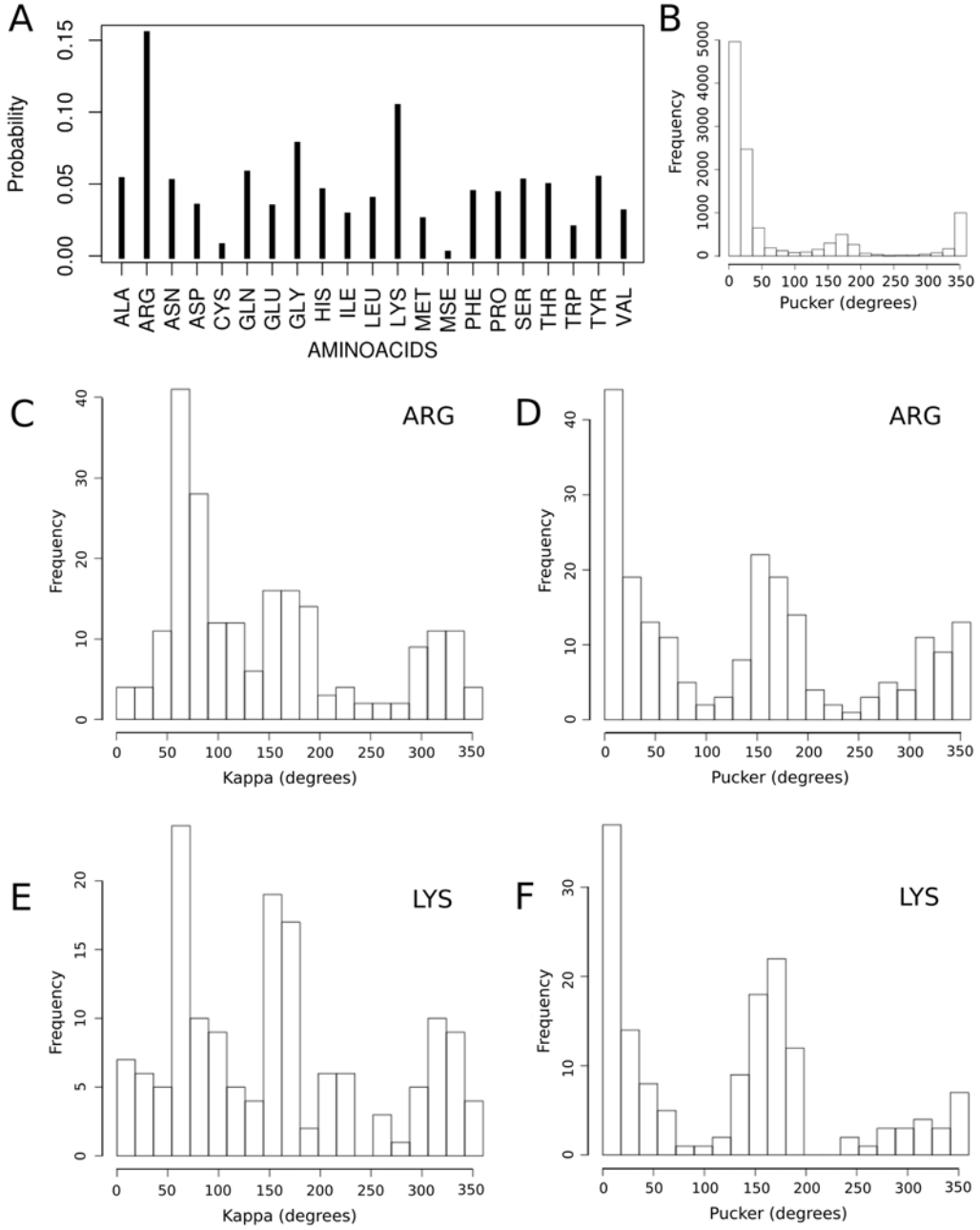



Figure 2. Protein-RNA contacts. (A) Probability of contact between a given aminoacid and the HO2' given a protein-RNA contact occur, calculated from counting all contacts (distance<=4 Å) between any protein atom and the oxygen of 2'OH, and splitting the counts per aminoacid identity. Multiple atoms of a given aminoacid within the distance cutoff were counted as one contact. X-ray and NMR chains and models specified in the Non-Redundant Dataset were used, see Supplementary Methods 1 for details. (B) Frequency of pucker phase values for all RNA nucleotides obtained from NMR structures in the Non-Redundant Dataset. (C) Frequency of kappa values for RNA nucleotides in contact with ARG atoms (distance <=4 Å) obtained from NMR structures in the Non-Redundant Dataset. (D) Frequency of pucker phase values for RNA nucleotides in contact (distance <=4 Å) with ARG atoms obtained from NMR structures in the Non-Redundant Dataset . (E,F) Same as (C) and (D), respectively, but for LYS aminoacid.

more, when the calculation is repeated for the Non-Redundant Dataset, equivalent results are obtained (Supplementary Figure 4A,B).

To gain additional information, we focus our study in those 2'OH interacting with protein residues. As for the κ distribution analysis, both the Full and the Non-Redundant Datasets were originally used. However,

although qualitatively similar trends are observed, differences between both datasets points toward some bias in the Full Dataset, that led us to discuss below only results on the Non-Redundant Dataset (see Figure 2; results from the Full dataset can be found in Supplementary Figure 5). In most cases 2'OH acts as a hydrogen bond acceptor, Lys and Arg being the preferred interacting partners (Figure 2A). Interaction of 2'OH with these protein side-chains leads to a stabilization of *conformers 1 and 2* and a parallel enrichment in **South** puckering (Figure 2B-E). Altogether analysis of experimental databases strongly suggest that sugar puckering and C2'OH rotational states are coupled, and that proteins interacting with the C2'OH can modulate the sugar puckering by biasing κ torsional preferences, which can lead to global structural changes in RNA. These findings suggest that the "innocuous" 2'OH group can be, in reality, a main player in determining protein-RNA recognition and the overall RNA structure.

**Puckering and C2'O2' torsions are coupled in ribonucleosides**. Database analysis above can be subjected to criticism, since the orientation of the C2'-O2' bond is not directly observed in the spectra, but inferred from indirect restraints. Thus, to support our database analysis we first performed QM studies of the pseudo-rotation profile of ribose for the three C2'O2' rotational states in dilute aqueous solution (see *Methods*). For both adenosine and cytosine, in the **North** state, *conformer 1* is the most stable orientation, *conformer 3* is close in energy (~0.5 kcal/mol), while *conformer 2* is disfavored by ~1.2 kcal/mol (Figure 3A,B). However, as suggested from database analysis, *conformer 2* (poorly populated in the **North** state) becomes the most stable orientation when the sugar samples the **South** state. Very exciting, if *conformer 2* is forced, **North** and **South** relative energies invert, with the latter becoming the most stable sugar puckering state (Figure 3A,B). This suggests that already at the nucleoside level the orientation of the 2'OH group can induce changes in sugar puckering. Very encouraging, similar results are obtained when flexibility and explicit solvent are considered in QM/MM PMFs of the C2'-O2' rotation (see Methods and Supplementary Methods 3), with restraints in sugar puckering (see Figure 3C,D). In summary, QM/SCRF and QM/MM calculations provide a picture of the κ torsional space of the nucleoside which qualitatively agrees with the database analysis of RNA motives. Furthermore, it reinforces the idea that C2'O2' torsion and puckering are coupled and that biasing of the κ torsion can lead to changes in puckering, which in turn dramatically affects the RNA conformation.

**C2'O2' torsions in RNA oligomers**. State-of-the-art simulations discussed above present a major caveat: the neglect of the polynucleotide environment, which can force the approach of different interactors to the 2'OH group, modifying the intrinsic properties of nucleosides described in the previous section. To solve this potential caveat, we computed the QM/MM PMF of the κ rotation in the rCpC dinucleotide in explicit solvent (see Methods and Supplementary Methods 3). Very encouraging, results in Figure 4A qualitatively agree with the observed preferred orientations for North-puckering riboses in the PDB analysis (Figure 1B,D) with *conformer 1* being the global minimum followed closely by *conformer 3*, ~0.3 kcal/mol higher energy, and *conformer 2* the least stable, ~1 kcal/mol above *conformer 1*. Conformational transitions between *conformers 1* and *3* happen through a ~1.8 kcal/mol free energy barrier localized at the eclipsed κ~ 0

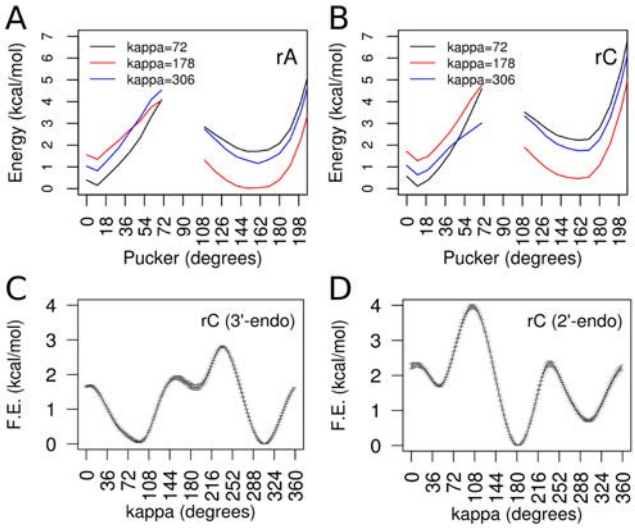



Figure 3. 2'OH kappa torsion and sugar pucker phase preferred conformers at the nucleoside level. (A,B) QM potential energy scans of the sugar pucker phase (for adenosine and cytosine, respectively) restraining the kappa torsion at the main observed minima in the kappa free energy profile (72, 178 and 306 degrees). The dependence of χ on the sugar puckering was taken into account by fixing χ=190 degrees for pucker values in the range 0-70 degrees and χ=230 degrees for pucker values in the range 110-216 degrees. (C,D) US QM/MM free energy profiles for the kappa torsion of a rC nucleoside with restraints on the sugar pucker phase at the 3'endo and 2'endo conformations, respectively. The continuous line and error bars correspond to the average and standard deviation of the free energy, respectively, calculated from the energy profiles obtained after 31, 32, 33, 34, 35, 36, 37, 38, 39 and 40 ps of US simulation.

value. These values are also consistent with high level QM calculations in solution for an isolated nucleoside, and in astonishing agreement with database analysis. In addition, the 2'OH contacts that were frequent in NMR-refined structures are also frequent in our QM/MM trajectories. Bader's analysis of electron densities in QM/MM snapshots (see Figure 4B-D and *Methods*) confirms the formation of hydrogen bond interactions both canonical (2'OH···3'OH: ρ~0.020 au) and non-canonical (2'OH···H5': ρ~0.009 au and H2'···O4': ρ~0.010 au). These electron density values confirm that "non-canonical" O···H···C hydrogen bonds are quite stable (~2-3 kcal/mol as estimated from the linear relationship between the interaction energy and bond critical point density reported in Cubero et at.(Cubero et al. 1999)), not far from a medium-strength canonical H-bond. This confirms previous claims on the stabilizing role of ribose

aliphatic hydrogens as "non-canonical" H-bond donors in modified oligonucleotides.(Martin-Pintado et al. 2013; Martín-Pintado et al. 2013)

**The impact of the C2'O2' torsion on the global RNA structure.** We performed MD simulations of a variety of standard RNA motives (see Supplementary Methods 2) to see whether or not the most accurate RNA force-field is able to capture the κ distribution found in database analysis and QM/MM calculations. Results in Figure 5A clear-

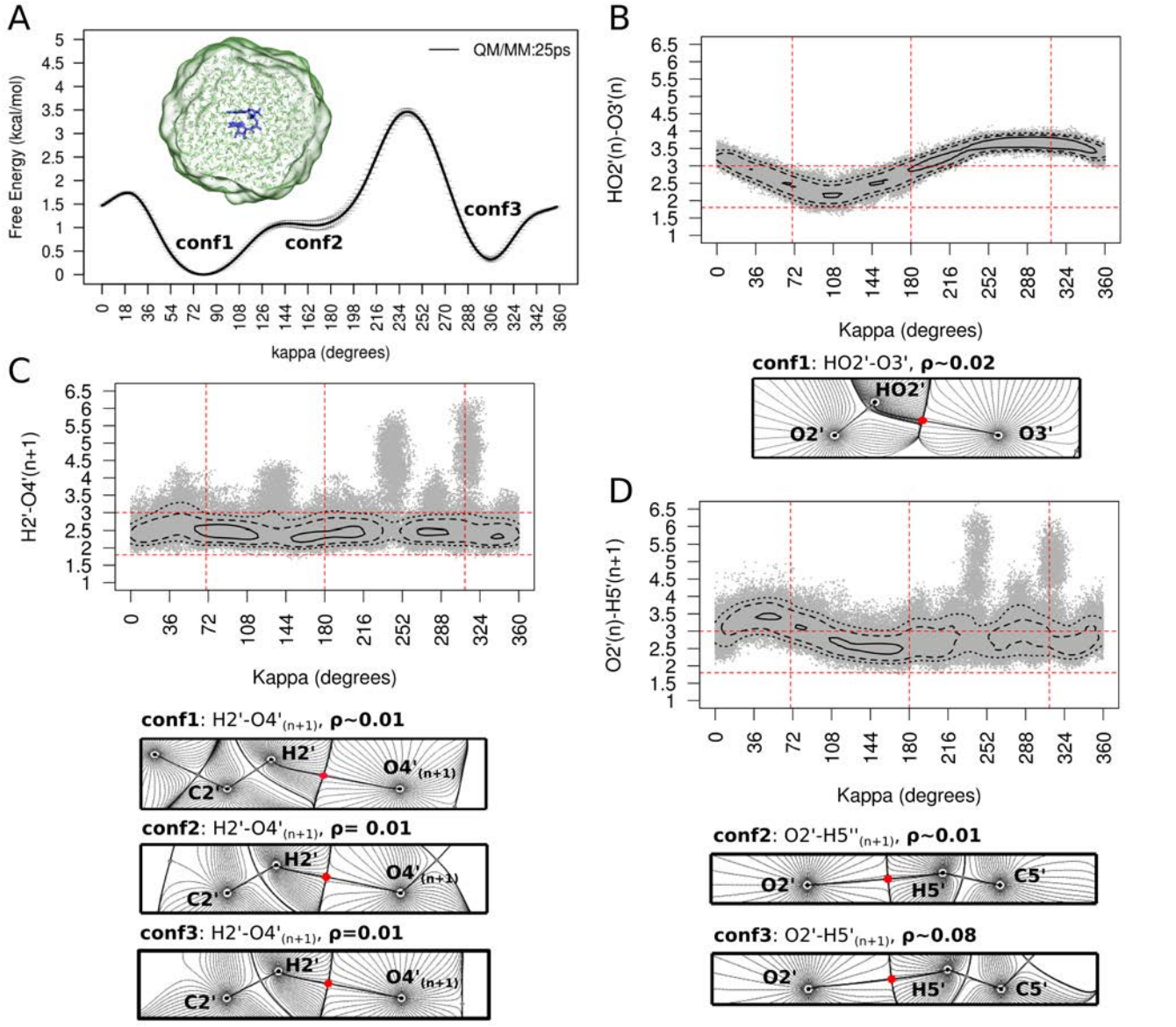



Figure 4. 2'OH kappa torsion preferred orientations at the dinucleotide level. (A) US QM/MM free energy profile for the kappa torsion of a rCC dinucleotide indicating three main orientations ("conf1", "conf2" and "conf3"). The continuous line and error bars correspond to the average and standard deviation of the free energy, respectively, calculated from the energy profiles obtained after 20, 21, 22, 23, 24 and 25 ps of US simulation. A snapshot of the simulated system is shown indicating in blue the QM region (rCC dinucleotide) and in green the MM region (water and $K^+$ ion). (B) Kappa vs HO2'-O3' distance scatter plot obtained form the US QM/MM simulation. Red dotted lines indicate optimal and maximum hydrogen bond distances (horizontal), and kappa rotation minimum energy positions (vertical). Contour lines correspond to points with density values equal to the average density plus 1, 2 and 4 standard deviations. In addition, AIM projection on the O2'-HO2'···O3' plane is also shown for a simulation snapshot corresponding to "conf1". The position of the bond critical point, the atomic nuclei involved in the interaction and gradient field lines are indicated with red and black dots, and grey lines, respectively. The density at the bond critical point (average over three simulation snapshots taken from "conf1") is also shown. (C) Same as (B) but for the interaction between C2'-H2'···O4'$_{(n+1)}$. In this case, the AIM analysis is shown for conformers 1-3. (D) Same as in (B) but for the interaction between C5'$_{(n+1)}$-H5'$_{(n+1)}$···O2'. In this case, the AIM analysis is shown for conformers 2-3.

ly indicate major errors in the κ distribution, which is dramatically biased towards *conformer 1*. This artifact is clearly related to a poor description of the C2'O2' torsion as highlighted by MM PMF calculations of the κ torsion (Figure 5B), which compared with the QM/MM reference (Figure 4A) shows a serious unbalance in the *conformer 1* vs *conformer 3* ratio. This behavior is not corrected if a local RNA adaptation of the DNA parmbsc1 force-field is

used (data not shown), and only slightly improved (Figure 5B) if a correction in the Lennard-Jones specific interaction between the O2' and the phosphate oxygens is introduced (see Supplementary Methods 4 and 5). Thus, the error in the κ distribution is related to an incorrect representation of the C2'O2' torsion in current state-of-the-art

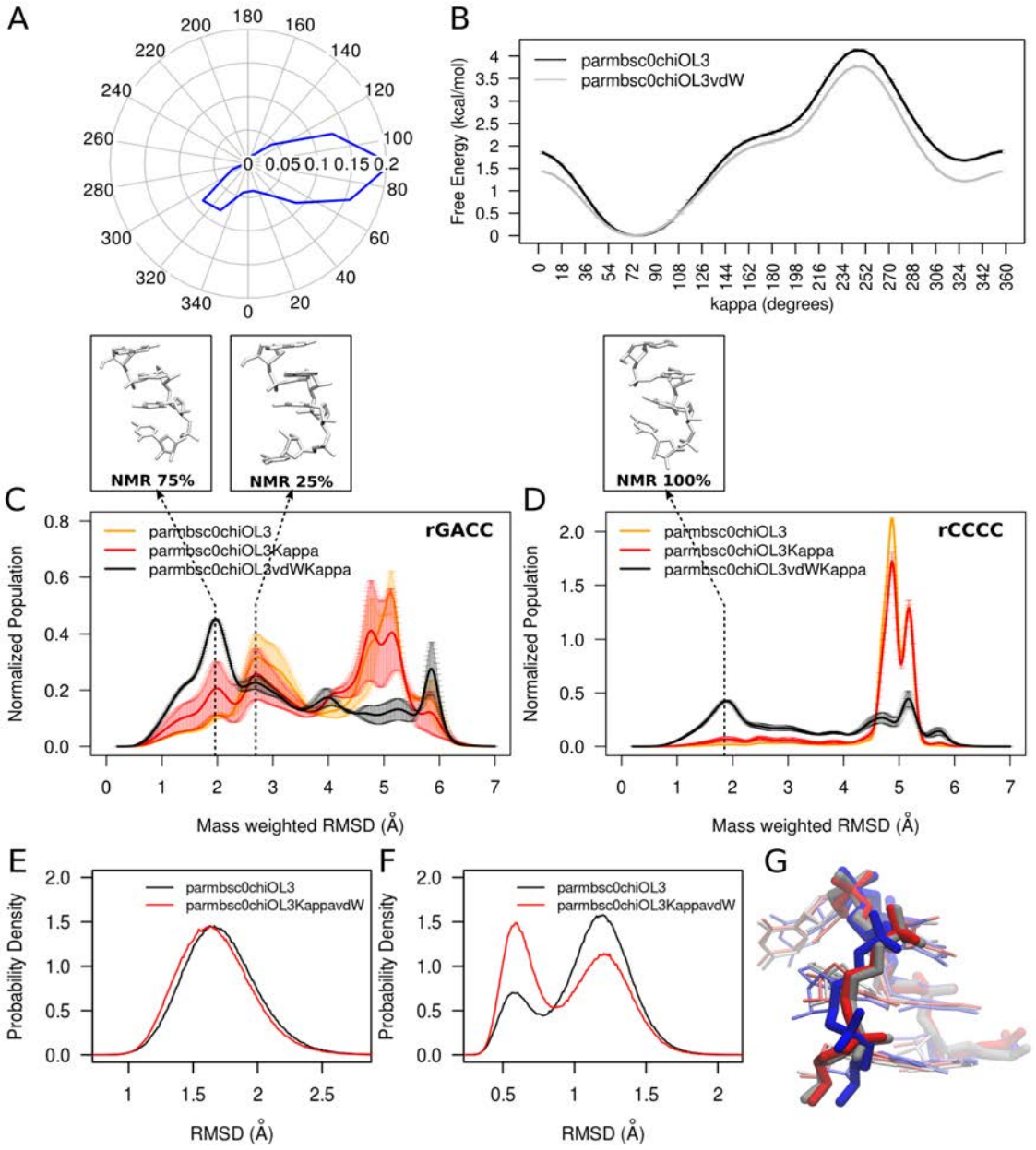



Figure 5. Kappa behaviour in RNA MD simulations using parmbscochiOL3. (A) Probability distribution of the Kappa torsion from unbiased MD simulations of three hairpins and three kissing-loops (see Supplementary Methods 2). (B) US MM free energy profile for the kappa torsion of a rCC dinucleotide with (gray line) and without (black line) a specific correction in the Lennard-Jones potential between the phosphate oxygen atoms and the hydroxyl oxygen atoms (see Supplementary Methods 4). The profile and error bars shown correspond to the average and standard deviation from five energy profiles obtained between 2 and 2.5 ns every 100 ps. (C) RMSD distribution (calculated using all atoms) from HREMD simulations of the rGACC tetra-nucleotide using parmbscochiOL3 (orange line), parmbscochiOL3Kappa (red line) and parmbscochiOL3KappavdW (black line). Error bars correspond to the standard deviation from the average (continuous line) obtained from two duplicates of the HREMD simulations (see Supplementary Methods 2). The reference structure used for the alignment (prior to RMSD calculation) corresponds to an A-form portion of the H. marismortui ribosome crystal structure (PDBID: 3G6E, residues 2623-2626). Representative structures of the first two peaks are indicated with dotted arrows, and correspond to NMR major and NMR minor structures,19 respectively. (D) Same as in (C) but for the tetra-nucleotide rCCCC. In this case, the reference structure used for the alignment corresponds to a canonical A-form generated using NAB. A representative structure of the first peak is indicated with a dotted arrow, and corresponds to the unique conformation observed in NMR.18 (E) RMSD distribution calculated using the backbone atoms of all residues in a RNA hairpin (PDBID: 2KOC) from two unbiased 1ms long MD simulations using parmbscochiOL3 (black line) or parmbscochiOL3KappavdW (red line). (F) Same as in (E) but for the region containing the loop plus the first stem base pair. (G) Three-dimensional representation of the region considered in (D) taken

from the experimental structure (gray) and the centroids of the clusters corresponding to the peaks at ~0.6 Å (red) and 1.2 Å (blue) in the RMSD distribution shown in (D).

force-fields. The impact of this inaccuracy is maximized in RNAs showing low level of secondary structure, as is the case of the tetra-nucleotides r(GACC) and r(CCCC), where the incorrect sampling of the κ torsion contributes to the formation of artefactual contacts stabilizing incorrect structures for the oligo in HREMD simulations (see Figure 5C,D). Correction of the C2'O2' torsion to reproduce the QM/MM κ profiles (see Supplementary Figure 6) improves the results (Figure 5C,D), but there is a problem of transferability of the parameters between nucleotides in the middle and termini of the stand, as the presence/absence of neighboring phosphates generate different environments. Adding the specific Lennard-Jones tuning (see before) improves the fitting and guarantees transferability (see Supplementary Methods 4 and 5; Supplementary Figure 6), and yields a much better representation of the tetra-nucleotide conformational space (see Figure 5 C,D).

Very encouraging, the improvement is also visible in a longer system (the 14 mer r(GGCACUUCGGUGCC) hairpin 2KOC containing the UUCG tetra-loop), where the loop region (r(CUUCGG)) is poorly described by the standard force-field (Figure 5F), while it is well represented with the modified parameters (Figure 5 E,F). Inspection of the three-dimensional structure (Figure 5G) shows that the peak at ~1.2 Å is consequence of a torsional shift affecting the phosphate group linking the loop with the stem in 3' (blue representation in Figure 5G). Such an effect is not observed in the structure corresponding to the peak at 0.6 Å (red representation), which appears well overlapped with the experimental structure (gray representation). Inspection of the backbone torsion angles indicate a shift of ε from *trans* to *gauche(-)* and a change in β from the canonical *trans* to an artificial 90 degrees conformation. These results highlight the importance of the suggested modifications, especially in regions of linkage between single and double stranded regions, which in fact, are of the most relevant ones for defining the RNA secondary structure.

## CONCLUSIONS

By combining a variety of complementary techniques (database analysis, high level QM calculations, QM/MM and classical simulations) we provided convincing evidence that the C2'O2' torsion has a key role in differentiating between DNA and RNA, being a main determinant of the local and global structure of RNA. The C2'O2' torsion is not an isolated degree of freedom, but correlates with a myriad of non-bonded contacts and it is strongly coupled with sugar puckering. Some C2'O2' torsional states favor the transition to unusual puckerings, whose presence is required in several protein-DNA contacts. Our results demonstrate that protein binding can bias the C2'O2' torsional state by forming specific hydrogen bonding with 2'OH group, leading to a **North**<->**South** transition required for functional RNA-protein binding and that can introduce global changes in the overall structure of the RNA. A variety of different techniques agree in the C2'O2' torsion acting as the trigger for a general novel induced fit mechanism of protein-RNA recognition. Finally, our results rises concerns on the current state-of-the-art RNA force fields, but also suggest that simple recalibration of the C2'O2' torsion, which will be implemented in our new force-field for RNA, can lead to a much improved description of unusual RNA conformations.

## ASSOCIATED CONTENT

**Supporting Information**

**Supplementary Methods:** **1**-Database Analysis, **2**-MD Additional Details, **3**-QM/MM Additional Details, **4**-Kappa Parametrization, **5**-Parmed.py commands for the Lennard-Jones specific interactions modification; **Supplementary Tables:** **S1**-Kappa Torsion Angle Analysis, **S2**-Kappa and Pucker Analysis for Ribonucleotides with 2'OH in Contact with ARG or LYS, **S3**-Protein-RNA Contacts Analysis; **Supplementary Figures:** **S1**-Preferred Orientations of the Kappa Torsion per Base Type, **S2**-Possible hydrogen bonds nearby the 2'OH group, **S3**-Preferred Orientations of the Kappa Torsion Angle from a Non-Redundant Database, **S4**-Kappa Energy Profile for Different Window Sizes, **S5**-Protein-RNA Contacts from a Non-Redundant Database, **S6**-Kappa fitting to reproduce QM/MM potential of mean force. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

* modesto.orozco@irbbarcelona.org

## ACKNOWLEDGMENT

This work has been supported by the Spanish Ministry of Science (BFU2014-61670-EXP), the Catalan SGR, the Instituto Nacional de Bioinformática, and the European Research Council (ERC SimDNA), the European Union's Horizon 2020 research and innovation programme under grant agreement No 676556, the Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) cofunded by the Fondo Europeo de Desarrollo Regional (FEDER), and the MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). M. O. is an ICREA academia researcher. L. D. is a SNI (Sistema Nacional de Investigadores; ANII, Uruguay) researcher. P. D. D. is a SNI (Sistema Nacional de Investigadores; ANII, Uruguay) and PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) researcher. The authors also acknoledge the Barcelona Supercomputing Center for CPU and GPU time on MareNostrum and MinoTauro. Federica Battistini, Fernando Romeo and Adria Ferandez are acknowledged for providing parmbscochiOL3 MD trajectories of hairpin and kissing-hairpin systems and Diego Gallego for contributing to the R-scripting used in the experimental database analysis.

## REFERENCES

(1) Caetano-Anollés, G.; Caetano-Anollés, D. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 427.

(2) Petrov, A. S.; Gulen, B.; Norris, A. M.; Kovacs, N. A.; Bernier, C. R.; Lanier, K. A.; Fox, G. E.; Harvey, S. C.; Wartell, R.

M.; Hud, N. V.; Williams, L. D. *Proc. Natl. Acad. Sci.* **2015**, *112* (50), 15396.
(3) Petrov, A. S.; Bernier, C. R.; Hsiao, C.; Norris, A. M.; Kovacs, N. A.; Waterbury, C. C.; Stepanov, V. G.; Harvey, S. C.; Fox, G. E.; Wartell, R. M.; Hud, N. V.; Williams, L. D. *Proc. Natl. Acad. Sci.* **2014**, *111* (28), 10251.
(4) Saint-Leger, A.; Bello, C.; Dans, P. D.; Torres, A. G.; Novoa, E. M.; Camacho, N.; Orozco, M.; Kondrashov, F. A.; Ribas de Pouplana, L. *Sci. Adv.* **2016**, *2* (4), e1501860.
(5) Zhang, J.; Ferré-D'Amaré, A. *Life* **2016**, *6* (1), 3.
(6) Soliva, R.; Luque, F. J.; Alhambra, C.; Orozco, M. *J. Biomol. Struct. Dyn.* **1999**, *17* (1), 89.
(7) Auffinger, P.; Westhof, E. *J. Mol. Biol.* **1997**, *274* (1), 54.
(8) Fohrer, J.; Hennig, M.; Carlomagno, T. *J. Mol. Biol.* **2006**, *356* (2), 280.
(9) Mládek, A.; Banáš, P.; Jurečka, P.; Otyepka, M.; Zgarbová, M.; Šponer, J. *J. Chem. Theory Comput.* **2014**, *10* (1), 463.
(10) Egli, M.; Portmann, S.; Usman, N. *Biochemistry (Mosc.)* **1996**, *35* (26), 8489.
(11) Denning, E. J.; MacKerell, A. D. *J. Am. Chem. Soc.* **2012**, *134* (5), 2800.
(12) *RNA 3D structure analysis and prediction*; Leontis, N. B., Westhof, E., Eds.; Nucleic acids and molecular biology; Springer: Heidelberg ; New York, 2012.
(13) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113* (18), 6378.
(14) Bader, R. F. W. *J. Phys. Chem. A* **1998**, *102* (37), 7314.
(15) Bader, R. F. W. *Chem. Rev.* **1991**, *91* (5), 893.
(16) Bader, R. F. W. *Atoms in molecules: a quantum theory*; The International series of monographs on chemistry; Clarendon Press ; Oxford University Press: Oxford [England] : New York, 1994.
(17) D, V.; D, A. *J. Comput. Methods Sci. Eng.* **2014**, No. 1–3, 131.
(18) Tubbs, J. D.; Condon, D. E.; Kennedy, S. D.; Hauser, M.; Bevilacqua, P. C.; Turner, D. H. *Biochemistry (Mosc.)* **2013**, *52* (6), 996.
(19) Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H. *J. Phys. Chem. B* **2011**, *115* (29), 9261.
(20) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16* (4), 845.
(21) 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. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179.
(22) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92* (11), 3817.
(23) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. *J. Chem. Theory Comput.* **2011**, *7* (9), 2886.
(24) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6* (12), 3836.
(25) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. *Nat. Methods* **2015**.
(26) Dang, L. X.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99* (1), 55.
(27) Dang, L. X. *J. Am. Chem. Soc.* **1995**, *117* (26), 6954.
(28) Smith, D. E.; Dang, L. X. *J. Chem. Phys.* **1994**, *100* (5), 3757.
(29) Cubero, E.; Orozco, M.; Hobza, P.; Luque, F. J. *J. Phys. Chem. A* **1999**, *103* (32), 6394.
(30) Martin-Pintado, N.; Deleavey, G. F.; Portella, G.; Campos-Olivas, R.; Orozco, M.; Damha, M. J.; González, C. *Angew. Chem. Int. Ed.* **2013**, *52* (46), 12065.
(31) Martín-Pintado, N.; Yahyaee-Anzahaee, M.; Deleavey, G. F.; Portella, G.; Orozco, M.; Damha, M. J.; González, C. *J. Am. Chem. Soc.* **2013**, *135* (14), 5344.

# Supplementary Information

# SMALL DETAILS MATTER: THE IMPORTANCE OF THE "INNOCUOUS" 2'HYDROXYL

Leonardo Darré[1,2], Ivan Ivani[1,2], Pablo D. Dans[1,2], Hansel Gómez[1,2], Adam Hospital[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. Barcelona, Spain

[3]Department of Biochemistry, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain

* Send correspondence to M.Orozco: modesto.orozco@irbbarcelona.org

## Contents

**Supplementary Methods 1. Database Analysis.**

All the analysis of NMR or X-ray structures was done using local R scripts using the bio3D[1] libraries.

**Kappa Torsion Distribution.** Two datasets were used to build the kappa torsion empirical distribution: i- the "Full Dataset" which contains the current state of the PDB up to June 2016 for NMR-solved structures containing RNA (610 entries), and ii- the "Non-Redundat Dataset" which contains NMR-solved RNA structures (476 entries) proposed by Leontis et al.[2] to avoid structural redundancy available from the BGSU Structural Bioinformatics Group web page (http://rna.bgsu.edu/rna3dhub/nrlist/), see Supplementary Table 1 for further details. For the Full Dataset, all NMR models in every PDB entry were split into RNA continuous segments (two or more residues), and the kappa torsion angle was measured for every ribonucleotide within a given segment. The canonical hydrogen bond local interactions of the 2'OH group were analyzed by measuring the distance between the 2'OH hydrogen atom and the atoms: O3', O4' and O2 (pyrimidines)/N3 (purines) from the same ribonucleotide, or O5', OP1, OP2 and O4' of the ribonucleotide in 3'. In addition, non-canonical hydrogen bonds were assessed by measuring the distance between the H2' or O2' of a given ribonucleotide and O4' or H5'/H5'' of the ribonucleotide in 3', respectively. To capture the effect of the sugar conformation on the kappa torsion angle, the pucker phase was also measured using Westhof & Sundaralingam definition[3] and obtaining kappa/pucker phase pairs for each analyzed ribonucleotide. Kappa probability distributions were calculated using angle windows of 20 degrees and plotted for 3'endo and 2'endo pucker phases separately, for all bases together or split by base type. The correlation between the kappa torsion angle and the distances to local hydrogen bond acceptors/donors are shown by means of scatter plots and three density contours corresponding to point in the distance-kappa space with density equal to the average density plus one, two or four standard deviations. Finally, kappa distributions were converted to empirical free energies from the relative populations of kappa values between 0 and 360 degrees, considering windows of 20, 15, 10 and 5 degrees, using the relation: $\Delta G_{i/0}=R*T*Ln(P_i/P_0)$, where $P_i$ and $P_0$ are the population of kappa values for windows i and 0, respectively. The windows [0,20], [0,15], [0,10], and [0,5] were used as reference (window 0) for each of the four striding options mentioned above. The measurement of kappa and pucker and the calculation of the empirical free energy was repeated for the Non-Redundant

Dataset although in this case only specific chains and NMR models were used as suggested in the BGSU Structural Bioinformatics Group web page.

**Kappa Torsion and Pucker Phase Distributions in 2'OH-ARG/LYS Contacts.** Both dataset mentioned in the previous section were filtered keeping only PDB entries corresponding to protein-RNA complexes (see Supplementary Table 2). Kappa and pucker distributions were obtained for ribonucleotides with the 2'OH group in contact with the aminoacids ARG and LYS (distance between any ARG or LYS atom and the oxygen atom of the 2'OH moiety lower or equal to 4 Å). When multiple atoms from the same ARG or LYS residue were in contact with a given ribonucleotide 2'OH, the corresponding kappa/pucker pair was counted only once.

**Probability of Contacts Between a Given Aminoacid and the 2'OH Group.** The Full and the Non-Redundant Datasets filtered to keep only protein-RNA complexes, which contain only NMR-solved structures, were supplemented with X-ray solved protein-RNA complexes obtained from the current state of the PDB (up to June 2016) or the Leontis et al. non-redundant database, respectively, for resolutions below 2.5 Å (see Supplementary Methods 3). For both NMR/X-ray datasets, the number of contacts (distance <= 4 Å) between any aminoacid atom and the 2'OH oxygen atom was counted. When multiple atoms from the same aminoacid were in contact with a a given ribonucleotide 2'OH, the contact was counted only once to eliminate repeated counts per aminoacid. The contacts frequency per aminoacid was divided by the total number of observed contacts, thus obtaining the aminoacid-2'OH interaction probability given that a contact exists.

**Supplementary Methods 2. MD Additional Details.**

All classical MD simulations were run using AMBER-14 suite. TLEAP code was used for systems preparation, CPPTAJ for post-processing and analyzing trajectories and ParmEd to modify and check topologies when needed (e.g. scale torsion angles force constants for HREMD calculations). Restraints were imposed using native AMBER algorithms or by means of the PLUMED 2.2 patch to AMBER-14. Generation of free energy profiles from umbrella sampling simulations was achieved using vFEP.[4]

**Unbiased Molecular Dynamics Simulations.** Microsecond long MD simulations of six RNA structures corresponding to three hairpins (PDBIDs: 1JJ2, 1Q9A and 2KOC ) and three kissing loops (PDBIDs: 1BAU, 2BJ2 and 2RN1) were run using parm99 forcefield[5,6] supplemented with the bsc0[7] and chiOL3[8,9] corrections (here in called “parmbsc0chiOL3”) to model the RNA. To take into account solvent model effects, two

of the most widely used water models were employed, TIP3P[10] for the hairpin structures and SPC/E[11] for the kissing loops structures. In all cases a 150mM ionic environment was represented using Dang parameters[12–14] for K+ and Cl-. MD simulations were performed in the NPT ensemble using Berendsen thermostat[15] with a time constant of 5 $ps^{-1}$ and the Berendsen barostat with a time constant of 5 $ps^{-1}$. Equations of motion were integrated using a time step of 2fs with the pmemd.cuda code[16] was used. Each system was subject to 2000 steps of energy minimization with position restraints in the solute of 25 kcal/mol, followed by 1 ns of position restrained (5 kcal/mol) thermalization in the NVT ensemble and 10 ns unrestrained equilibration in the NPT ensemble. Production MD simulations were run for 1 μs. Non-bonded direct cut-off was set to 9 Å and particle mesh Ewald[17] was used for reciprocal space calculations. All bonds involving hydrogen atoms were constrained by means of SHAKE algorithm.[18]

**Hamiltonian Replica Exchange Molecular Dynamics Simulations.** The conformational landscape of two tetranucleotides, rGACC and rCCCC, were explored enhancing the sampling by allowing coordinates exchange between eight replicas where all torsion angle force constants are scaled by: 1, 0.9 , 0.8, 0.7, 0.6, 0.5, 0.4, and 0.3, achieving an exchange acceptance in the range of 25-60%. rGACC initial structure was taken from an A-form portion of the *H. marismortui* ribosome crystal structure (PDBID: 3G6E, residues 2623-2626), following the same approach as Henriksen et al.[19] rCCCC initial structure was generated in a random conformation using NAB. The RNA molecule in each system was modelled using parmbsc0chiOL3, solvated using the TIP3P model[10] and neutralized with three K+ ions using Dang parameters.[12–14] Preparation of both systems for the first set of HREMD involved 2000 steps of position restrained (25 kcal/mol) minimization, and heated during 2 ns of MD from 10-150 K (NVT and 25 kcal/mol position restraints) and from 150-300 K (NPT and 5 kcal/mol position restraints), using a time step of 1 fs. System density at 300 K and 1 Bar was relaxed in 5 ns of 2 fs time step MD in the NPT ensemble with soft position restraints (0.5 kcal/mol) further extended by 500 ps of unrestrained equilibration in NVT. Production HREMD simulations were run in the NVT ensemble at 300 K using the Langevin thermostat with a collision frequency of 2 $ps^{-1}$ and resetting the random seed at each restart to avoid synchronization effects. A 2 fs time step was used with an exchange attempt every 1 ps. Non-bonded direct cut-off was set to 8 Å and particle mesh Ewald[17] was used for reciprocal space calculations. All bonds involving hydrogen

atoms were constrained by means of SHAKE algorithm.[18] The independent second run of HREMD simulations were started from the restart structures of the first run after 500 ns, assigning new velocities and equilibrating for 1 ns in the NVT ensemble. Total simulated time for both independent runs was 1.2 μs per replica. Equations of motion were integrated using the pmemd.cuda.MPI code.

**Umbrella Sampling Molecular Dynamics Simulations.** Classical mechanics umbrella sampling simulations were run for the rC nucleoside and the rCpC dinucleotide to obtain the kappa torsion potential of mean force in order to compare with the corresponding profiles at QM/MM level. For both systems, the solute was modelled using parmbsc0chiOL3 forcefield, solvated using TIP3P water model[10] and neutralized (rCpC) with one K+ ions using Dang parameters[12–14]. The rotation of the kappa torsion was sampled in twenty windows of 18 degrees applying a restraining potential on kappa of 35 kcal/mol. Each window initial configuration was extracted from an exploratory well tempered metadynamics[20] simulation (50 ns; initial Gaussian high of 1.2 kJ/mol; deposition period of 1ps; sigma=0.35 radians; BIASFACTOR=4, T=300 K) of the rCpC dinucleotide, and further equilibrated for 500 ps in the NPT ensemble at 300K and 1 Barr. Production data was collected for 2.5 ns of NPT molecular, dynamics for each window. Restraints on beta and gamma backbone torsions, as well as on the sugar pucker were used as in the QM/MM simulations detailed below.

**Supplementary Methods 3. QM/MM Additional Details.**

All QM/MM dynamics simulation were run using the interface between TERACHEM[23–26] and AMBER (MM) as implemented in AMBER-14, with a time step for the integration of the equations of motion of 1 fs. Potential energy walls (when required) and/or restraints were enforced by means of PLUMED 2.2[22] patch to AMBER-14. Calculation of the free energy profile from the umbrella sampling trajectories was achieved using vFEP.[4]

**Kappa Torsion Potential of Mean Force.** Umbrella sampling QM/MM simulations were run to obtain the free energy profile of the C2'O2' (kappa) torsion rotation for a cytosine nucleoside (rC) and for a cytosine dinucleotide (rCpC) in aqueous solution. The system setup was the same as per the classical umbrella sampling calculations (see previous section). In both cases the nucleic acid was treated at the quantum level BLYP/6-31G(d) while the aqueous environment (water or water plus one $K^+$ ion) was treated at the classical level (TIP3P[10] and Dang parameters[12–14] for ions). The rotation of the kappa torsion was sampled in twenty windows of 18 degrees applying a

restraining potential on kappa of 35 kcal/mol. Each window was first equilibrated fully classically ("parmbsc0chiOL3") for 500 ps in the NPT ensemble (300 K and 1 Barr). The restart classical configurations were relaxed at the QM/MM level for 5 ps and production simulations were carried out for 40 and 25 ps for rC and rCpC, respectively. Wavefunction SCF calculations were done in mixed precision including DFTD3 dispersion corrections.[27] In the case of the rC nucleoside, sugar pucker transitions were frequently observed affecting the sampling of the kappa rotation. Consequently, a potential energy wall as implemented in PLUMED 2.2[22] was applied to one of the Zx Cartesian coordinates of the ring puckering[21] (a lower wall at Zx=0.3 to maintain the 3'endo conformation or an upper wall at Zx=-0.3 to maintain the 2'endo conformation). The dinucleotide simulation maintained the 3'endo initial pucker, thus the use of walls was not required (that was not the case for the MM simulations where pucker phase restraints were needed). For both rC and rCpC, 5kcal/mol restraints on the beta and gamma backbone torsions were applied to avoid interactions with the phosphate oxygen atoms. For rC additional restraints (5kcal/mol) were also applied on epsilon backbone torsion to keep it at the standard value.

**Supplementary Methods 4. Kappa Parametrization.** In parm99bsc0chiOL3 the C2'O2' torsion rotation is controlled by three dihedral angles: C1'-C2'-O2'-HO2' (dihedral type: CT-CT-OH-HO), C3'-C2'-O2'-HO2' (dihedral type: CT-CT-OH-HO) and H2'-C2'-O2'-HO2' (dihedral type: H1-CT-OH-HO). To avoid affecting non-RNA OH moieties described using the current AMBER forcefield distributions, a new atom type for the O2' atom was introduced (OK) for refitting the Kappa torsion angle. The dihedral type H1-CT-OH-HO was substituted by H1-CT-OK-HO with a new set of parameters, while the dihedral type CT-CT-OH-HO was renamed CT-CT-OK-HO but keeping the original set of parameters. As in the parmbsc0 and parmbsc1 parametrization procedure, a flexible Metropolis Monte Carlo algorithm was used to fit a truncated third order Fourier series to the difference between: i- QM/MM pmf of the Kappa rotation for the rCpC dinucleotide, and ii- the corresponding pmf obtained at MM level (parmbsc0chiOL3$_{\text{H1-CT-OK-HO=0}}$). Both QM/MM and MM potentials of mean force were obtained from umbrella sampling calculations for the sugar in North conformation as described in Supplementary Methods 2 and 3 (see Supplementary Figure 6A). The obtained new parameters (see Supplementary Table 4) were tested on two tetranucleotide systems (rGACC and rCCCC) exhaustively exploring their conformational landscapes by means of Hamiltonian Replica Exchange simulations

(see Supplementary Methods 2 for simulation details). In addition to the previous parametrization, a second fitting was performed considering a specific modification of the Lennard-Jones potential (increase in the sigma parameter) between the phosphate oxygen atoms and : i- the ribose O2', O3' atoms, and ii- the amine nitrogen of the base (N6 in A, N2 in G and N4 in C), herein called “parmbsc0chiOL3vdW”. This correction to the Lennard Jones potential is based on the AMBER parameters revision for organic phosphates proposed by Steinbrecher et al,[28] which was recently shown to improve the description of RNA tetranucleotides.[29] In the present work, instead of including a general Lennard-Jones correction affecting the interaction between the phosphate oxygen atoms and all other atoms in the system, the specific terms affecting only the atoms mentioned above were corrected (see Supplementary Methods 5 section for the parmed.py script). The Kappa torsion parameters were fitted as before but using the parmbsc0chiOL3vdW$_{H1\text{-}CT\text{-}OK\text{-}HO=0}$ pmf for the MM level (see Supplementary Figure 6). reference The obtained parameters (see Supplementary Table 5) were tested again on the tetranucleotide systems (rGACC and rCCCC) and on microsecond-long unbiased MD simulations of three RNA hairpins and three kissing-loop systems (see Supplementary Methods 2 for simulation details).

**Supplementary Methods 5. Parmed.py commands for the Lennard-Jones specific interactions modification.**

```
changeLJPair @%OS @%N2 3.5958 0.17
changeLJPair @%O2 @%N2 3.5733 0.188944436
changeLJPair @%O2 @%OH 3.4703 0.210199905
changeLJPair @%OS @%OH 3.4928 0.189124298
changeLJPair @%O2 @%OK 3.4703 0.210199905
changeLJPair @%OS @%OK 3.4928 0.189124298
addLJType @O4' radius 1.6837 epsilon 0.1700
parmout OUTFILE
go
```

**Supplementary Table 1. Kappa Analysis (only NMR structures).[a]**

| | Full Dataset | Non-redundant Dataset |
|---|---|---|
| Number of entries | **610 (*7518*)[b]** | **476 (*531*)[b]** |
| Number of analysed entries | **584 (*7256*)[b]** | **459 (*503*)[b]** |
| Number of analysed nucleotides | **174511** | **11212** |

[a]All available NMR models were used in the PDB (10/06/2016) set analysis, while only specific models were used for the non-redundant dataset (see Supplementary Methods 1).

[b] Number of NMR models for the given set of PDB entries.

**Supplementary Table 2. Kappa and pucker analysis for ribonucleotides with 2'OH in contact (distance <=4 Å) with ARG or LYS (only NMR structures).**

| | | Full Dataset | Non-redundant Dataset |
|---|---|---|---|
| Number of protein-RNA PDB entries available | | **107 (*1709*)[a]** | **89 (*135*)[a]** |
| Number of protein-RNA PDB entries analysed | | **107 (*1709*)[a]** | **89 (*135*)[a]** |
| Number of analysed nucleotides with the 2'OH in contact with: | ARG | **1756[b]** | **212[b]** |
| | LYS | **1647[b]** | **152[b]** |

[a] Number of NMR models for the given set of PDB entries.

[b] Removing repeated kappa/pucker values due to contacts with different atoms of the same aminoacid in a given contact.

**Supplementary Table 3. Protein-RNA contacts analysis.**

| | Full Dataset | Non-redundant Dataset |
|---|---|---|
| Number of available PDB entries | **514** | **319** |
| Number of available X-ray entries | **407** | **230 (*238*)[b]** |
| Number of available NMR entries | **107 (*1709*)[c]** | **89 (*135*)[c]** |
| Total number of available models (X-RAY+NMR) | **2116** | **373** |
| Number of analysed PDB entries | **500** | **307** |
| Total Number of analysed models (X-RAY+NMR) | **2102** | **361** |
| Number of analysed contacts (distance<=4Å) | **26760[a]** | **5309[a]** |

[a] Removing repeated counts from different atoms of the same aminoacid in a given contact.

[b] Number of X-RAY models for the given set of PDB entries.

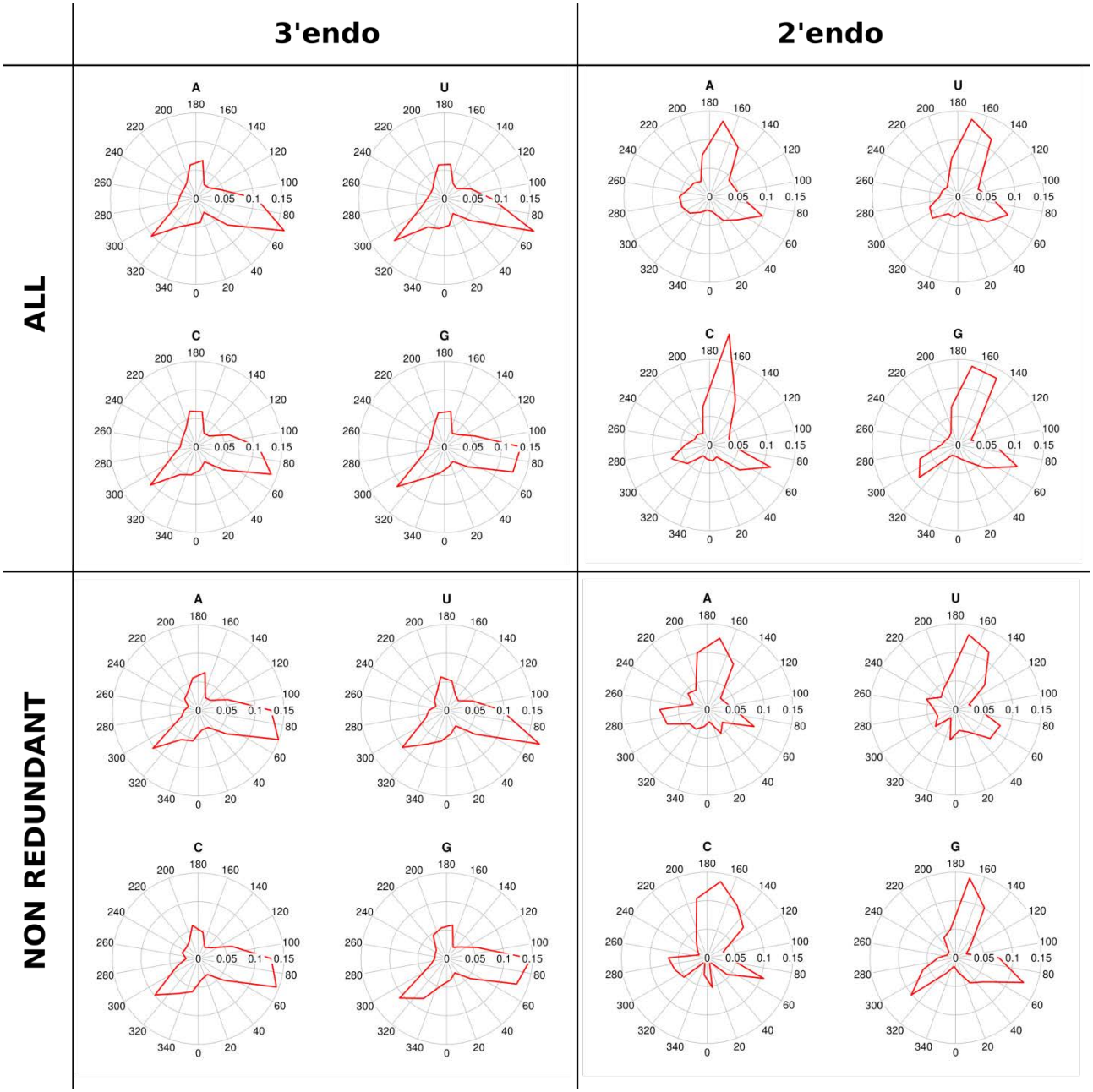
[c] Number of NMR models for the given set of PDB entries.

**Supplementary Table 4. H1-CT-OK-HO parameters.**

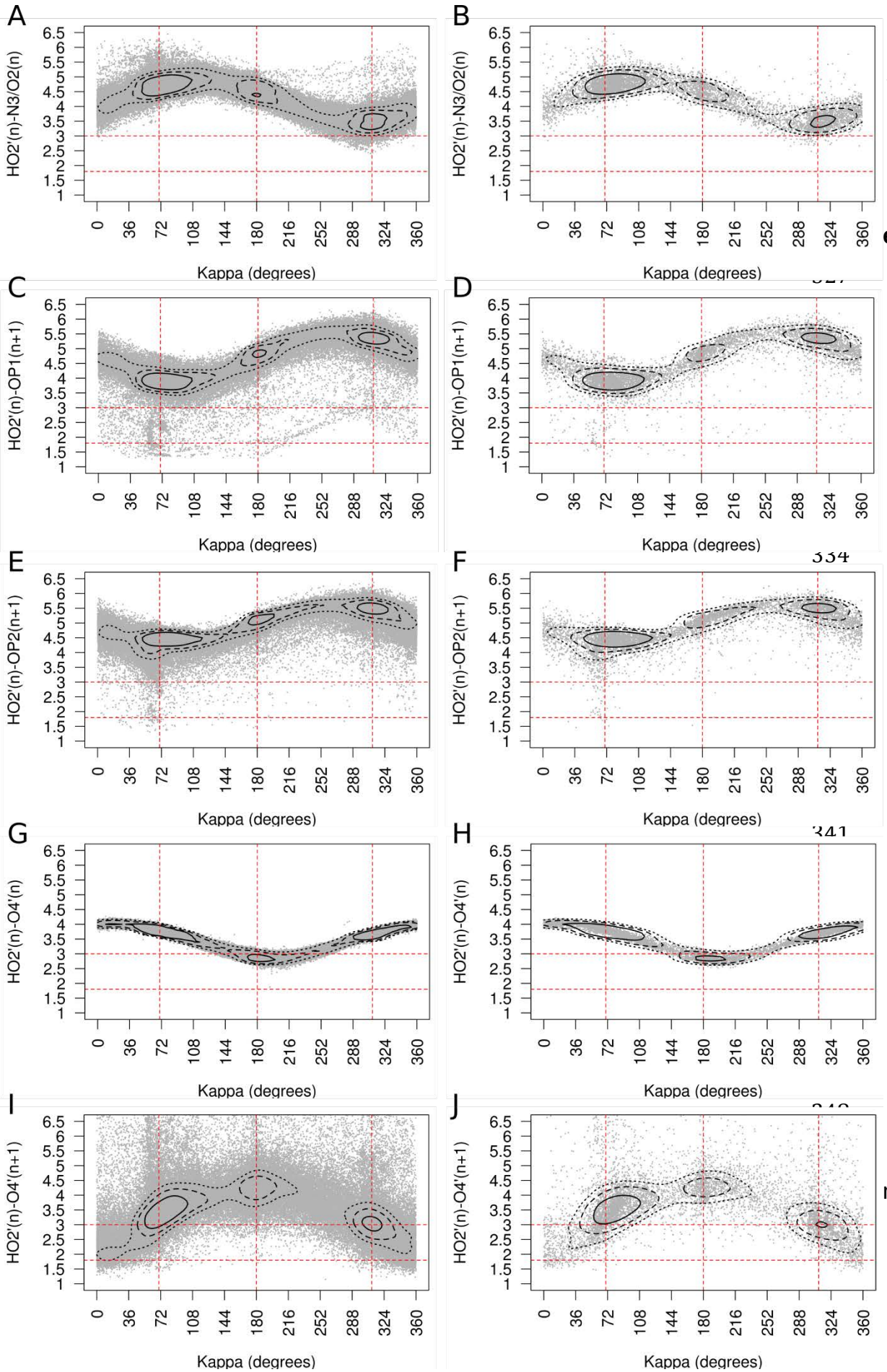
| Torsion | $V_n/2$ | Phase | Periodicity |
|---|---|---|---|
| H1-CT-OK-HO | 0.482 | 18.8 | -3 |
| H1-CT-OK-HO | 0.336 | 59.4 | 2 |
| H1-CT-OK-HO | 0.549 | 96.9 | 1 |

**Supplementary Table 5. H1-CT-OK-HO parameters considering vdW specific corrections.**

| Torsion | $V_n/2$ | Phase | Periodicity |
|---|---|---|---|
| H1-CT-OK-HO | 0.501 | 0.0 | -3 |
| H1-CT-OK-HO | 0.287 | 74.3 | 2 |
| H1-CT-OK-HO | 0.519 | 60.7 | 1 |



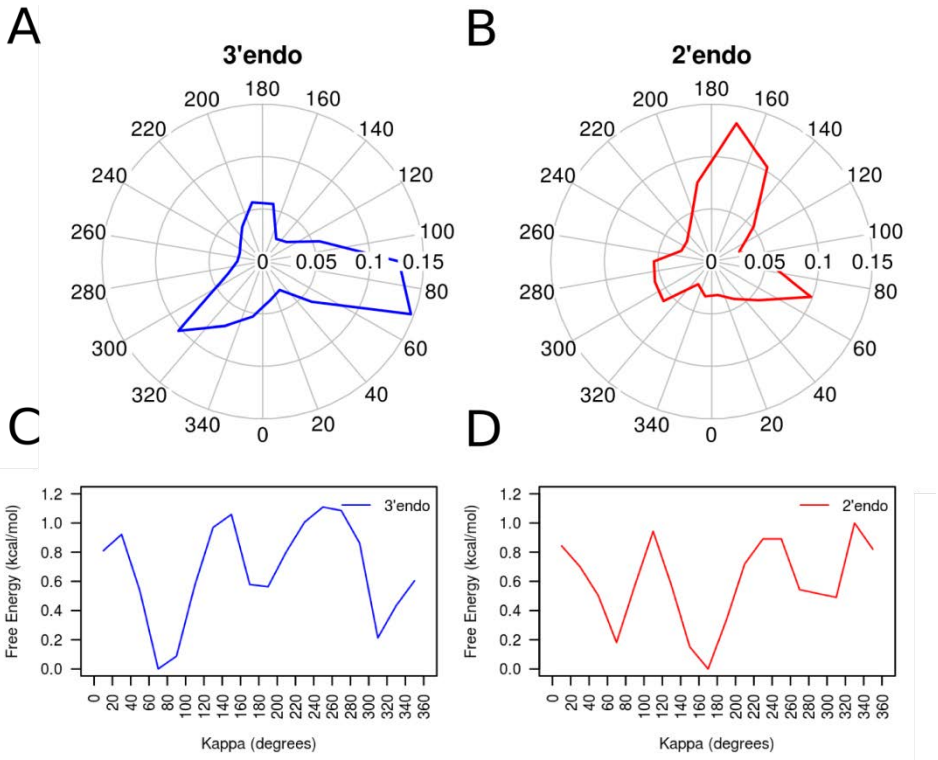

**Supplementary Figure 1. Preferred orientations of the kappa torsion per base type.** The plots show the probability distribution of the torsion angle between the atoms H2'-C2'-O2'-HO2' for 3'endo or 2'endo ribonucleotides and for the current state of the PDB or a non-redundant database (see Supplementary Methods 1), split by base type.



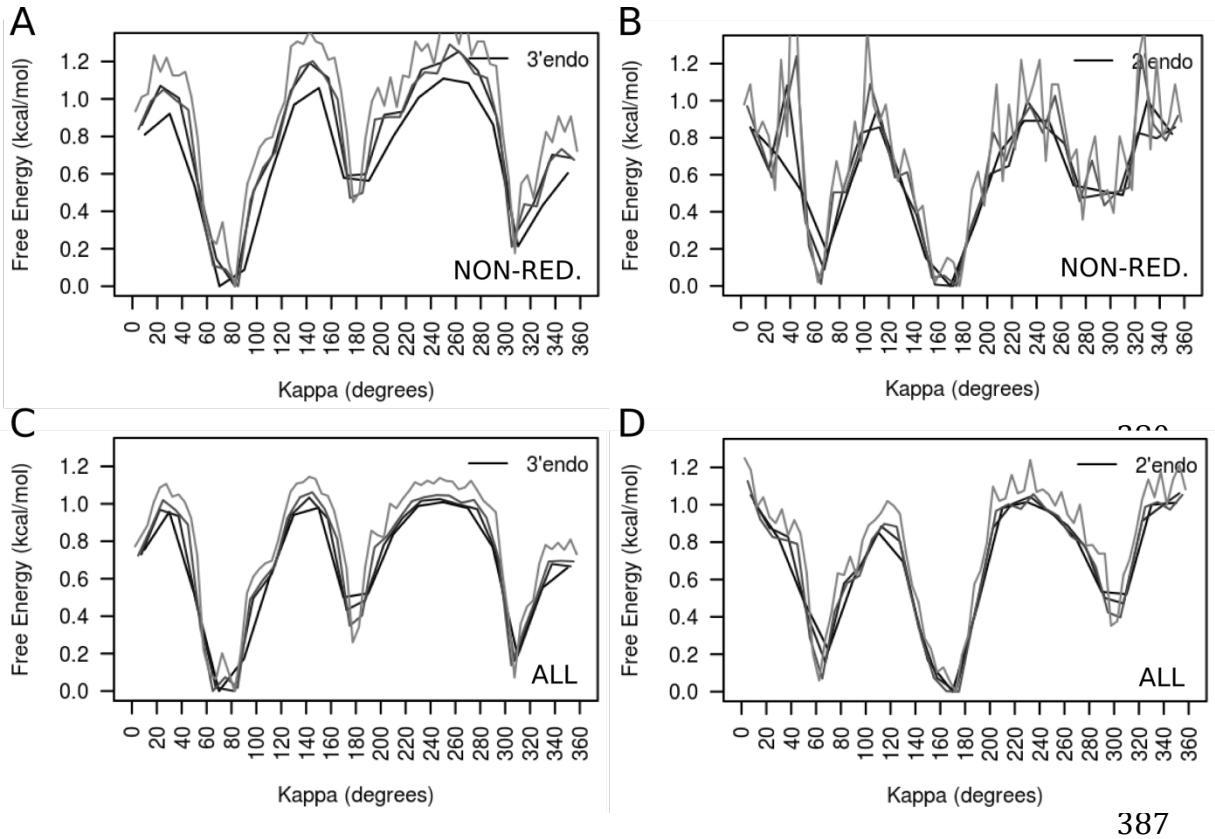

**Supplementary Figure 2. Possible hydrogen bonds acceptor/donors near by the 2'OH group.** Scatter plots of kappa torsion vs distance between

355 HO2' and local acceptors/donors of hydrogen bonds, are shown for nucleotides with

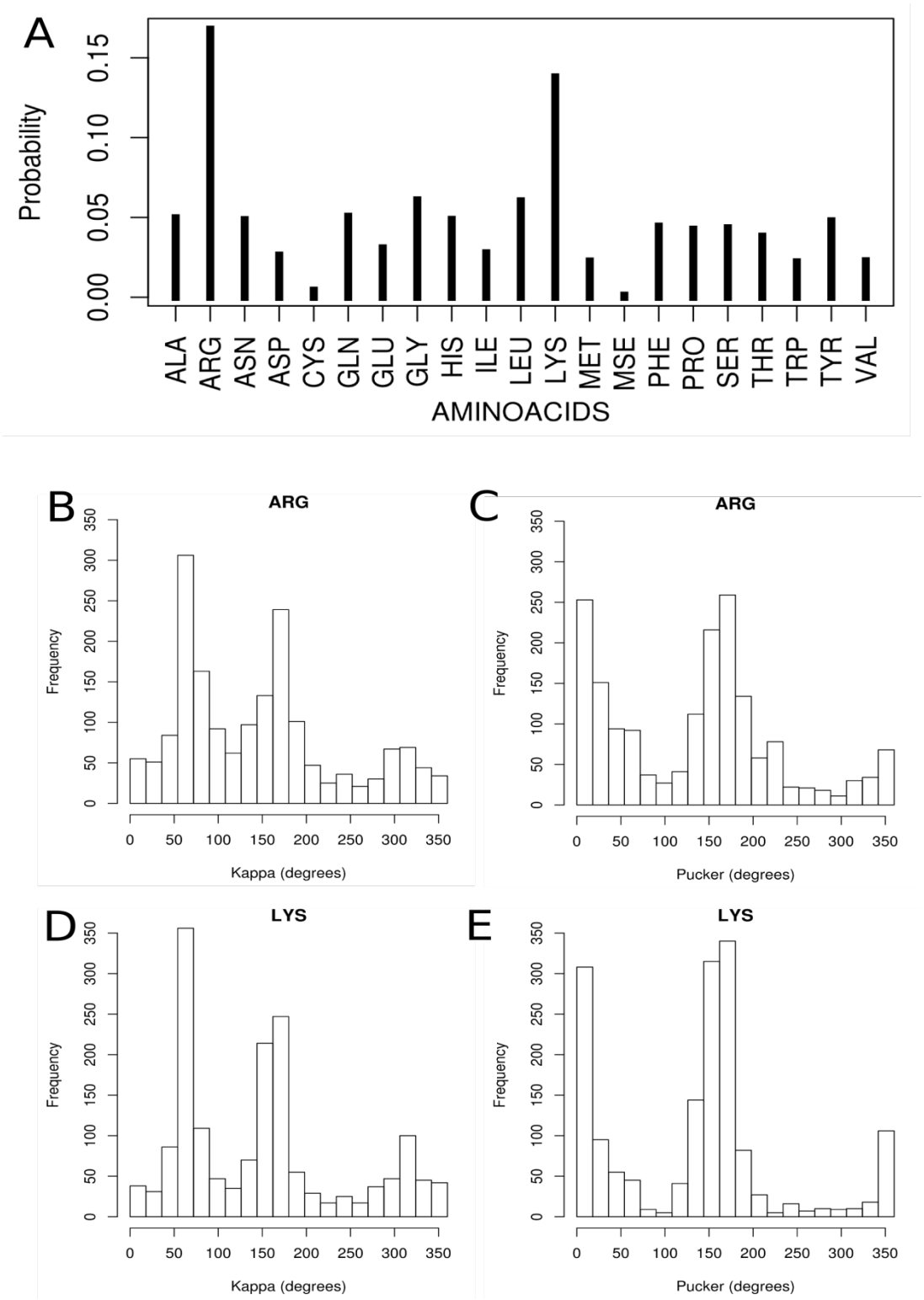
pucker phase in North for both Full Dataset (A, C, E, G and I) and Non-Redundant Dataset (B, D, F, H and J). Red dotted lines indicate optimal and maximum hydrogen bond distances (horizontal), and kappa rotation minimum energy positions (vertical). Contour lines correspond to points with density values equal to the average density plus 1, 2 and 4 standard deviations. Data for both




**Supplementary Figure 3. Preferred orientations of the kappa torsion from a non-redundant database.** (A) Probability distribution of the torsion angle between the atoms H2'-C2'-O2'-HO2' for all the 3'endo ribonucleotides of the RNA dataset obtained from a non-redundant database (see Methods). (B) Same as in (A) but for 2'endo ribonucleotides. (D,E) Empirical free energy calculated from the experimental kappa distributions in (B) and (C), respectively.



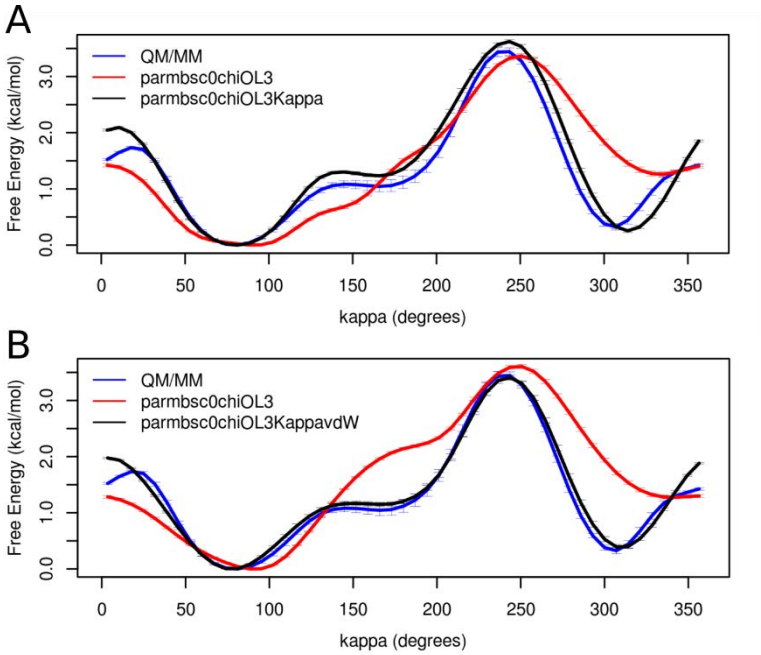

**Supplementary Figure 4. Kappa energy profile for different window sizes.** (A) Empirical free energy calculated from the kappa distribution of 3'endo ribonucleotides of the non-redundant RNA dataset (see Supplementary Methods 1), splitting the data using four different window sizes: 20 degress (black), 15 degrees (dark gray), 10 degrees (gray), and 5 degrees (light gray). (B) Same as in (A) but for 2'endo ribonucleotides. (C) Same as in (A) but using all current RNA entries in the PDB. (D) Same as (C) but for 2'endo ribonucleotides.




**Supplementary Figure 5. Protein-RNA contacts for the Full Dataset.** (A) Probability of contact between a given aminoacid and the HO2' given a protein-RNA contact occur, calculated from counting all contacts (distance<=4 Å) between any protein atom and the oxygen of 2'OH, and splitting the counts per aminoacid identity. Multiple atoms of a given aminoacid within the distance cutoff were counted as one contact. All X-ray and NMR (multiple models) from the Full Dataset (see Methods) were used. B) Frequency of kappa values for RNA nucleotides in contact with ARG atoms (distance <=4 Å)

obtained from NMR (multiple models) structures in the Full dataset. C) Frequency of pucker phase values for RNA nucleotides in contact (distance <=4 Å) with ARG atoms obtained from NMR (multiple models) structures in the Full Dataset. D,E) Same as B and C, respectively, but for LYS aminoacid.

**Supplementary Figure 6:** Kappa fitting to reproduce QM/MM potential for mean force.




(A) US QM/MM (blue), parmbsc0chiOL3$_{H1\text{-}CT\text{-}OK\text{-}HO=0}$ (red) and parmbsc0chiOL3 with the correction on the kappa torsion (parmbsc0chiOL3Kappa, black) free energy profiles for the kappa torsion of a rCC dinucleotide. The profile and error bars correspond to the average and standard deviation from five energy profiles obtained after 20-25ps every 1 ps (QM/MM) and 2-2.5ns every 100ps (parmbsc0chiOL3$_{H1\text{-}CT\text{-}OK\text{-}HO=0}$ and parmbsc0chiOL3Kappa). (B) Same as in (A) but including the Lennard-Jones modification (see Supplementary Methods 4) on parmbsc0chiOL3$_{H1\text{-}CT\text{-}OK\text{-}HO=0}$ (red) and on parmbsc0chiOL3Kappa (black).

## Supplementary References

1. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22,** 2695–2696 (2006).
2. Leontis, N. B. & Zirbel, C. L. in *RNA 3D Structure Analysis and Prediction* (eds.

Leontis, N. & Westhof, E.) **27,** 281–298 (Springer Berlin Heidelberg, 2012).

3. 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.* **105,** 970–976 (1983).
4. Lee, T.-S., Radak, B. K., Pabis, A. & York, D. M. A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *J. Chem. Theory Comput.* **9,** 153–164 (2013).
5. Cheatham, T. E., Cieplak, P. & Kollman, P. A. A Modified Version of the Cornell *et al.* Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J. Biomol. Struct. Dyn.* **16,** 845–862 (1999).
6. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).
7. Pérez, A. *et al.* Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **92,** 3817–3829 (2007).
8. Zgarbová, M. *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.* **7,** 2886–2902 (2011).
9. Banáš, P. *et al.* Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **6,** 3836–3849 (2010).
10. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926 (1983).
11. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91,** 6269–6271 (1987).
12. Dang, L. X. & Kollman, P. A. Free Energy of Association of the K+:18-Crown-6 Complex in Water: A New Molecular Dynamics Study. *J. Phys. Chem.* **99,** 55–58 (1995).

13. Dang, L. X. Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J. Am. Chem. Soc.* **117,** 6954–6960 (1995).
14. Smith, D. E. & Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **100,** 3757 (1994).
15. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684 (1984).
16. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **9,** 3878–3888 (2013).
17. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N  log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089 (1993).
18. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23,** 327–341 (1977).
19. Henriksen, N. M., Roe, D. R. & Cheatham, T. E. Reliable Oligonucleotide Conformational Ensemble Generation in Explicit Solvent for Force Field Assessment Using Reservoir Replica Exchange Molecular Dynamics Simulations. *J. Phys. Chem. B* **117,** 4014–4027 (2013).
20. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100,** (2008).
21. Huang, M., Giese, T. J., Lee, T.-S. & York, D. M. Improvement of DNA and RNA Sugar Pucker Profiles from Semiempirical Quantum Methods. *J. Chem. Theory Comput.* **10,** 1538–1545 (2014).
22. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185,** 604–613 (2014).
23. Ufimtsev, I. S. & Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles

Molecular Dynamics. *J. Chem. Theory Comput.* **5,** 2619–2628 (2009).

24. Titov, A. V., Ufimtsev, I. S., Luehr, N. & Martinez, T. J. Generating Efficient Quantum Chemistry Codes for Novel Architectures. *J. Chem. Theory Comput.* **9,** 213–221 (2013).
25. Götz, A. W., Clark, M. A. & Walker, R. C. An extensible interface for QM/MM molecular dynamics simulations with AMBER. *J. Comput. Chem.* **35,** 95–108 (2014).
26. Isborn, C. M., Götz, A. W., Clark, M. A., Walker, R. C. & Martínez, T. J. Electronic Absorption Spectra from MM and *ab Initio* QM/MM Molecular Dynamics: Environmental Effects on the Absorption Spectrum of Photoactive Yellow Protein. *J. Chem. Theory Comput.* **8,** 5092–5106 (2012).
27. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132,** 154104 (2010).
28. Steinbrecher, T., Latzer, J. & Case, D. A. Revised AMBER Parameters for Bioorganic Phosphates. *J. Chem. Theory Comput.* **8,** 4405–4412 (2012).
29. Bergonzo, C. & Cheatham, T. E. Improved Force Field Parameters Lead to a Better Description of RNA Structure. *J. Chem. Theory Comput.* **11,** 3969–3972 (2015).

## 5.2 RNA dumbbells (Publication 5)

RNA interference is a natural defense mechanism of gene regulation triggered by 21-23 base pairs RNA duplexes with 3'-terminal overhangs, called siRNAs, which, when incorporated into the RISC protein complex, induce degradation of the complementary target mRNAs (Fire et al. 1998). Soon after, it was discovered that synthetic siRNAs produce similar effect, thus having an attractive biomedical potential (Elbashir et al. 2001). Their biggest limitation is their vulnerability to degradation by serum exonucleases. To address this issue, our group developed a new class of 3'-exonuclease-resistant modification of the siRNA molecule (Terrazas et al. 2013). The approach consisted in replacing the 3'-terminal natural dinucleotide overhangs with dimeric N-ethyl-N bridged nucleosides (two 2'-deoxy-5-methylcytidine units linked together by ethyl chain through the exocyclic amino group, also called BCn dimer; *n* being the number of carbon atoms of the alkyl chain). The results of their study showed that these capped duplex structures, called dumbbell, showed an extreme resistance to nuclease digestion, without affecting RISC sensitivity.

Based on these promising results, we decided to computationally investigate the configurational and dynamic behavior of BCn dimers. We performed MD simulations of the BC2-, BC6- and BC8-linker dumbbell structures, and the control linear dsRNA duplex analogue. Linker structures were geometrically optimized on QM level from which point charges were parameterized using standard RESP calculation. We used latest force field for RNA, which has the same approach as parmbsc1 for DNA, but it is still in the development stage up to this day. Simulation results showed stability of all 3 designed dumbbell structures, similar to the wild-type simulation of the RNA duplex (see Figure 2). From helical perspective, there are almost no changes in the distribution of helical parameters between the 4 structures studies (see Supplementary Figure S1). Looking at the dynamics of the linkers, we saw higher flexibility of BC6- and BC8- linkers (see Figure 5.2), with the BC8- linker exploring a much larger configurational space (see Supplementary Figure S2).

Overall, computational study allowed us to design new dumbbell-shaped BC-n RNA loop, whose higher biostability with respect to their linear 3'-BC-modified version and the 7 nt-loop dumbbells has been experimentally verified. The BC6-linker dumbbell could be used for targeting the relevant GRB7 oncogene in SKBR3 breast cancer cells (work under development).

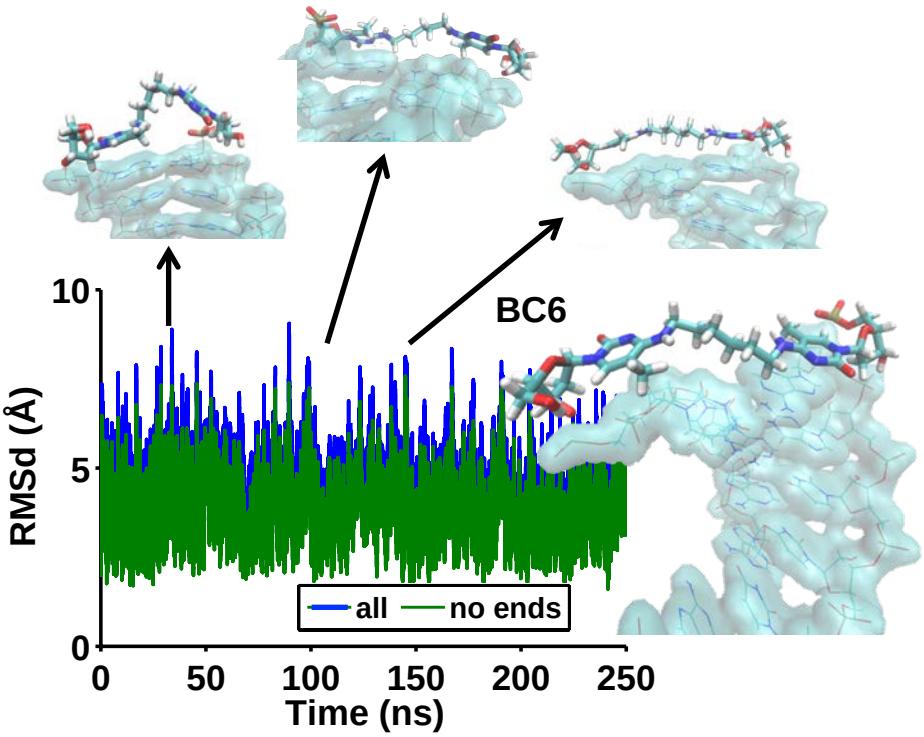



**Figure 5.3. Dynamics of the BC6 dumbbell.** RMSd of the entire dumbbell (all) with representative snapshots, and excluding terminal dimers (no ends).

Nucleic Acids Research Advance Access published March 14, 2016

*Nucleic Acids Research, 2016* **1**
*doi: 10.1093/nar/gkw169*

# Rational design of novel N-alkyl-N capped biostable RNA nanostructures for efficient long-term inhibition of gene expression

**Montserrat Terrazas[1,*], Ivan Ivani[1], Núria Villegas[1,2], Clément Paris[3], Cándida Salvans[1], Isabelle Brun-Heath[1] and Modesto Orozco[1,2,4,*]**

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Joint IRB-BSC Program in Computational Biology, Baldiri Reixac 10-12, 08028 Barcelona, Spain, [2]Barcelona Supercomputing Center, Jordi Girona 29, 08034 Barcelona, Spain, [3]Department of Organic Chemistry and IBUB, University of Barcelona, Martí i Franquès 1-11, 08028 Barcelona, Spain and [4]Department of Biochemistry and Molecular Biology, University of Barcelona, 08028 Barcelona, Spain

Received November 06, 2015; Revised March 01, 2016; Accepted March 03, 2016



## ABSTRACT

**Computational techniques have been used to design a novel class of RNA architecture with expected improved resistance to nuclease degradation, while showing interference RNA activity. The *in silico* designed structure consists of a 24–29 bp duplex RNA region linked on both ends by N-alkyl-N dimeric nucleotides (BCn dimers; n = number of carbon atoms of the alkyl chain). A series of N-alkyl-N capped dumbbell-shaped structures were efficiently synthesized by double ligation of BCn-loop hairpins. The resulting BCn-loop dumbbells displayed experimentally higher biostability than their 3′-N-alkyl-N linear version, and were active against a range of mRNA targets. We studied first the effect of the alkyl chain and stem lengths on RNAi activity in a screen involving two series of dumbbell analogues targeting *Renilla* and Firefly luciferase genes. The best dumbbell design (containing BC6 loops and 29 bp) was successfully used to silence GRB7 expression in HER2+ breast cancer cells for longer periods of time than natural siRNAs and known biostable dumbbells. This BC6-loop dumbbell-shaped structure displayed greater anti-proliferative activity than natural siRNAs.**

## INTRODUCTION

RNA interference (RNAi) is an innate defense mechanism of gene regulation triggered by 21–23 nt RNA duplexes with 3′-terminal dinucleotide overhangs (siRNAs) (1,2) that are generated in the cytoplasm by Dicer cleavage of longer RNAs (3–5). After incorporation into the RISC protein complex, siRNAs induce degradation of the complementary target mRNAs. Shortly after the discovery of RNAi, synthetic siRNAs were found to produce the same effect (6,7). Since then, much effort has been made to exploit the RNAi process experimentally to inhibit the expression of genes of choice for therapeutic purposes (8,9). However, despite the attractive biomedical potential of this approach, siRNAs are not drug-like molecules. One of their most important limitations is their vulnerability to degradation by serum exo- and endonucleases (10,11).

Extensive research has been conducted to increase the biostability of these agents (8,9). These efforts have yielded a wide number of siRNAs containing chemical modifications in the sugar ring or the phosphate backbone (8,9,12–20). Relevant examples are siRNAs that incorporate electronegative substituents at the 2′-position in the sugar ring such as 2′-fluoro (12–16) and 2′-*O*-methyl substitutions (17,18), which are known to increase the biostability and thermal stabilities of siRNAs to a great extent (21–24). Another modification that has been widely studied is the phosphorothioate linkage (PS) (19,20).

Two main approaches are usually considered in order to increase nuclease resistance. The first one is based on the extensive modification of the RNA molecule. For example, Sirna Therapeutics has several products that involve combination of 2′-fluoro pyrimidines, DNA purines, PS linkages and abasic caps (9) in the sense and the guide strands. These combinations have yielded products of increased potency and long serum half-life (48–72 h) that have been successfully used in a Hepatitis B virus mouse model (25). However, the extensive modification of the siRNA molecules often is not-well tolerated by the RNAi machinery and can lead to the reduction of the gene silencing activity (7,17). The

*To whom correspondence should be addressed. Tel: +34 934020228; Fax: +34 034037175; Email: montserrat.terrazas@irbbarcelona.org
Correspondence may also be addressed to Modesto Orozco. Tel: +34 934037155; Fax: +34 034037175; Email: modesto.orozco@irbbarcelona.org

selective modification of the nuclease-sensitive sites represents an alternative approach to the extensive modification of the siRNA. For example, Alnylam Pharmaceuticals has several siRNA products that are selectively modified with 2′-*O*-methyl or 2′-fluoro substitutions at vulnerable sites (9,26) (http://www.alnylam.com/Programs-and-Pipeline/index.php). Therefore, development of minimal modification approaches aimed at improving or optimizing the properties of the siRNA molecule (e.g. by protecting the 3′-terminal ends of the duplex from exonuclease cleavage) are of significance nowadays.

With the aim of developing new and complementary approaches for modifying the siRNA molecule, in a recent study, we developed a new class of 3′-exonuclease-resistant modification (27) that involved minimal and selective alteration of the oligonucleotide. In particular, our strategy involved replacement of the 3′-terminal natural dinucleotide overhangs of siRNAs by dimeric nucleotides composed of two 2′-deoxy-5-methylcytidine units linked together by an ethyl chain through the exocyclic amino group of the nucleobase (N-ethyl-N bridged nucleosides). Based on the promising results of those studies, here we have undertaken a computational investigation of the conformational and dynamic behavior of 3′-terminal N-alkyl-N bridged nucleotides (BCn dimers; n = number of carbon atoms of the alkyl chain) that has allowed us to develop new RNA architectures with even higher nuclease resistance. In particular, dumbbell-shaped structures having a duplex RNA region comprised of 24–29 base pairs linked on both ends by BCn dimers (Figure 1).

There are only a few reports in the literature of dumbbell-shaped dsRNAs acting as siRNA precursors (28–30). Two relevant examples are nuclease-resistant RNA dumbbells with loops composed of seven natural nucleotides linked by standard phosphodiester bonds (28) and dimeric 1,2-bis(maleimido)ethane crosslinkers (29), which have been used to silence the expression of transiently transfected luciferase genes. In particular, the 7 nt-loop dumbbells were found to display significantly longer RNAi effect in cell culture than natural siRNA (28).

In view of the similarity between the BCn and 7 nt-loop dumbbells and the longer-lived RNAi activity of these structures, we decided to use them as controls in our studies. Stability studies in serum and cytosol cell extract confirmed that our BCn-loop design was more stable than the 7 nt-loop dumbbells described in the literature (28). Moreover, the best BCn-loop dumbbell design could be successfully applied to inhibit GRB7 (31) expression in HER2+ breast cancer cells and showed longer inhibitory effect than natural siRNA and their 7 nt-loop analogues (28). To our knowledge, the present work is the first example that provides data on long-term inhibition of endogenous genes induced by dumbbell-shaped RNAs.

## MATERIALS AND METHODS

### RNA synthesis

Oligonucleotide sequences that did not contain modified nucleotides were purchased from Sigma Aldrich. All modified sequences were synthesized at the 1 μmol scale via solid phase synthesis using standard phosphoramidite methods (32). Reagents for oligonucleotide synthesis including 2′-*O*-TBDMS-protected phosphoramidite monomers of $A^{Bz}$, $C^{Ac}$, $G^{dmf}$ and U, solid chemical phosphorylation reagent (5′-phosphate), the 5′-deblocking solution (3% TCA in $CH_2Cl_2$), activator solution (0.3 M 5-benzylthio-1-H-tetrazole in $CH_3CN$), CAP A solution (acetic anhydride/pyridine/THF), CAP B solution (THF/*N*-methylimidazole 84/16) and oxidizing solution (0.02 M iodine in tetrahydro-furan/pyridine/water (7:2:1)) were obtained from commercial sources.

For the synthesis of RNA strands containing BCn loops, commercially available 5′-*O*-DMT-$A^{Bz}$-3′-succinyl-LCAA-CPG, 5′-*O*-DMT-$C^{Ac}$-3′-succinyl-LCAA-CPG, 5′-*O*-DMT-$G^{dmf}$-3′-succinyl-LCAA-CPG and 5′-*O*-DMT-U-3′-succinyl-LCAA-CPG were used as the solid supports. For the synthesis of 3′-BC2-modified RNA strands, CPG functionalized with a BC2 unit (27) was used as the solid support. The coupling time was 15 min. The coupling yields of natural and modified phosphoramidites were around 95%. Incorporation of the dimeric nucleoside modification did not have a negative effect in the yield. Terminal 5′-phosphate group was incorporated by using the commercially available solid chemical phosphorylation reagent. All oligonucleotides were synthesized in DMT-OFF mode.

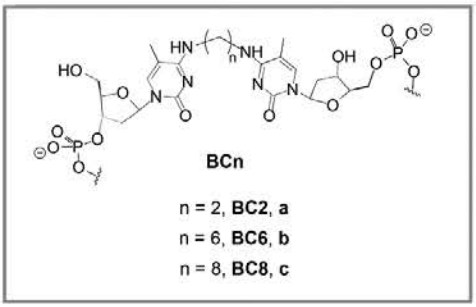
### Deprotection and purification of unmodified and modified RNA oligonucleotides

After the solid-phase synthesis, the solid support was transferred to a screw-cap vial and incubated at 55°C for 2 h with 1.5 ml of $NH_3$ solution (33%) and 0.5 ml of ethanol. The vial was then cooled on ice and the supernatant was transferred into a 2 ml eppendorf tube. The solid support and vial were rinsed with 50% ethanol (2 × 0.25 ml). The combined solutions were evaporated to dryness using an evaporating centrifuge. The residue that was obtained was dissolved in 1 M TBAF in THF (330 μl) and incubated at room temperature for 15 h. Then, 1 M triethylammonium acetate (TEEA) and water were added to the solution (330 μl TEEA and 330 μl water). The oligonucleotides were desalted on NAP-10 columns using water as the eluent and evaporated to dryness. The oligonucleotides were purified by 20% polyacrylamide gel electrophoresis (DMT-OFF). After purification, the RNAs were isolated by the crush and soak method, dialyzed, quantified by absorption at 260 nm and confirmed by Matrix-Assisted Laser Desorption/Ionization (MALDI) mass spectrometry (see Supplementary Table S1).

For annealing of linear siRNAs, 20 μM single strands were incubated in siRNA buffer (100 mM KOAc, 30 mM HEPES-KOH at pH 7.4, 2 mM $MgCl_2$) for 1 min at 90°C followed by 1 h at 37°C.

### Enzymatic synthesis of RNA dumbbells using T4 RNA ligase 2

Final composition of the reaction mixture (250 μl) was as follows: 2 μM 5′-phosphorylated-dsRNA, 22.5 units T4 RNA ligase 2 (New England Biolabs), 50 mM Tris-HCl (pH 7.5), 2 mM $MgCl_2$, 1 mM DTT, 400 μM ATP. After 5′-phosphorylated RNAs had been annealed, T4 RNA




5'-AUCUGAAGAAGGAGAAAAA**BC2**-3'
3'-**BC2**UAGACUUCUUCCUCUUUUU-5' **1**

**BCn**(UCAAAUCUGAAGAAGGAGAAAAAAUGGUU / AGUUUAGACUUCUUCCUCUUUUUUACCAA)**BCn** **2a-c**

**3a-c** **BCn**(UCAAAUCUGAA 3' pGAAGGAGAAAAAAUGGUU 5' / AGUUUAGACUUCUUCCUCp UUUUUUACCAA 3')**BCn** **4a-c**

G A G UCAAAUCUGAAGAAGGAGAAAAAAUGGUU U C A A
A A C U AGUUUAGACUUCUUCCUCUUUUUUACCAA G A G **5**

5'-UCAAAUCUGAAGAAGGAGAAAAAAUGGUU-3'
3'-AGUUUAGACUUCUUCCUCUUUUUUACCAA-5' **LdsRNA29**

5'-AUCUGAAGAAGGAGAAAAATT-3'
3'-TTUAGACUUCUUCCUCUUUUU-5' **siRNA I**

3'-TTUGCACUGUGCAAGCCUCUU-5'
5'-ACGUGACACGUUCGGAGAATT-3' **scr**

**BC2**(AGUGCGCUGC 3' pU 5' / UCACGCGAC G AC CACGGUUGGGA)**BC2** **6**

**7a,b** **BCn**(AAGUGCGCUG 3' pCUGGUGCCAACCCUU 5' / UUCACGCGACGACCAC p GGUUGGGAA 3')**BCn** **8a,b**

**BCn**(AAGUGCGCUGCUGGUGCCAACCCU / UUCACGCGACGACCACGGUUGGGA)**BCn** **9a,b**

G A G AAGUGCGCUGCUGGUGCCAACCCUU C A A
A A C U UUCACGCGACGACCACGGUUGGGA G A G **10**

5'-AAGUGCGCUGCUGGUGCCAACCCU**BC2**-3'
3'-**BC2**UUCACGCGACGACCACGGUUGGGA-5' **11**

5'-AAGUGCGCUGCUGGUGCCAACCCU-3'
3'-UUCACGCGACGACCACGGUUGGGA-5' **LdsRNA24**

5'-GUGCGCUGCUGGUGCCAACTT-3'
3'-TTCACGCGACGACCACGGUUG-5' **siRNA II**

**12** **BC6**(UCAAGCCCAAC 3' pA 5' AGCUUCGAAAUGGCCAC / AGUUCGGGUUGUUCGAAG p CUUUACCGGUG 5' 3')**BC6** **13**

**BC6**(UCAAGCCCAACAAGCUUCGAAAUGGCCAC / AGUUCGGGUUGUUCGAAGCUUUACCGGUG)**BC6** **14**

G A G UCAAGCCCAACAAGCUUCGAAAUGGCCAC U C A A
A A C U AGUUCGGGUUGUUCGAAGCUUUACCGGUG G A G **15**

5'-CCCAACAAGCUUCGAAAUGTT-3'
3'-TTGGGUUGUUCGAAGCUUUAC-5' **siRNA III**

**16** **BC6**(GUGAACGAGUC 3' pC 5' AACGUGUACGUGGUGAC / CACUUGCUCAGGUUGCAC p AUGCACCACUG 5' 3')**BC6** **17**

**BC6**(GUGAACGAGUCCAACGUGUACGUGGUGAC / CACUUGCUCAGGUUGCACAUGCACCACUG)**BC6** **18**

G A G GUGAACGAGUCCAACGUGUACGUGGUGAC U C A A
A A C U CACUUGCUCAGGUUGCACAUGCACCACUG G A G **19**

5'-CGAGUCCAACGUGUACGUGTT-3'
3'-TTGCUCAGGUUGCACAUGCAC-5' **siRNA IV**

**Figure 1.** BCn loops and RNA structures used in this study. The 29 bp RNAs **2a-c**, 24 bp RNAs **9a,b** and 29 bp RNAs **14** and **18** (and the corresponding linear controls **siRNA I**, **siRNA II**, **siRNA III** and **siRNA IV,** and 7 nt-loop dumbbell analogues **5**, **10**, **15** and **19**) target the *Renilla* and Firefly mRNAs and the 1019–1037 and 943–962 sites in the GRB7 mRNA. Top strand depicts the sense strand in the 5′→3′ direction (same as the target sequence). Bottom strand depicts the antisense strand in the 3′→5′ direction (complementary to the target). BCn: N-alkyl-N dimeric nucleoside; n: number of carbon atoms of the alkyl chain; scr: scrambled sequence; **p**: 5′-terminal phosphate group.

ligase 2 (New England Biolabs) was added to the concentrations described above and incubated at 37°C overnight. The RNA was precipitated by the addition of ethanol and sodium acetate (pH 5.2). Ligated products were purified by preparative (1 mm thick) denaturing PAGE (10%PAGE, 25% formamide, 7 M urea in 1X TBE). Bands were visualized by UV shadowing, and crushed and extracted with 0.1 M NaCl. The eluate was desalted by using Slide-A-Lyzer dyalisis columns (ThermoFischer Scientific).

### Dicer cleavage reaction of RNAs

RNAs (0.91 μM) were mixed with Dicer enzyme (0.091 units/μl; Recombinant Human Turbo Dicer Enzyme Kit from Genlantis, USA) in the buffer system supplied. The mixtures were incubated at 37°C and aliquots (2.2 μl) were taken from the mixture after 0, 1, 6 and 20 h. They were analyzed by 15% non-denaturing PAGE. The gels were visualized with SYBR Gold.

### UV-monitored thermal denaturation studies

Absorbance versus temperature curves of duplexes **LdsRNA24**, **9a**, **9b** and **10** were measured at 1 μM strand concentration in 10 mM phosphate buffer (pH 7.0) containing 1 mM EDTA. In the case of duplexes **LdsRNA29**, **2a** and **2b**, 10 mM phosphate, 5 mM EDTA buffer (pH 7.0) was used. Experiments were performed in Teflon-stoppered 1 cm path length quartz cells on a Varian-Cary-100 spectrophotometer equipped with thermoprogrammer. The samples were heated to 100°C, allowed to slowly cool to 20°C, and then warmed during the denaturation experiments at a rate of 0.5°C/min to 100°C, monitoring absorbance at 260 nm. The data were analyzed by the denaturation curve processing program, MeltWin v. 3.0. Melting temperatures ($T_m$) were determined by computerfit of the first derivative of absorbance with respect to 1/T.

### Cell culture

All cell lines (HeLa, HeLa H/P, SKBR3 and MCF7) were mantained at 37°C in a humidified atmosphere with 5% $CO_2$. HeLa and MCF7 cells were cultured in Dulbecco's modified Eagle's medium (DMEM; GIBCO) supplemented with fetal bovine serum (FBS, 10%), penicillin (100 U $ml^{-1}$) and streptomycin (100 μg $ml^{-1}$). SKBR3 cells were cultured in McCoy's modified medium (GIBCO) supplemented with FBS (10%), penicillin (100 U $ml^{-1}$) and streptomycin (100 μg $ml^{-1}$). HeLa H/P cells stably expressing pGL4.14 [luc2/Hygro] (Promega) and pRL-tk-Puro (a kind gift of Dr Alagia) were maintained under hygromycin B (200 μg $ml^{-1}$) and puromycin (2 μg $ml^{-1}$) selection pressure.

### Luciferase siRNA assays

HeLa cells were regularly passaged to mantain exponential growth. The cells were seeded one day prior to the experiment in a 24-well plate at a density of 1.5 x $10^5$ cells/well in complete DMEM containing 10% FBS (500 μl per well). Following overnight culture, the cells were treated with luciferase plasmids and siRNAs. Two luciferase plasmids—*Renilla* luciferase (pRL-TK) and firefly luciferase (pGL3) from Promega—were used as a reporter and control. Cotransfection of plasmids and siRNAs was carried out with Lipofectamine 2000 (Life Technologies) as described by the manufacturer for adherent cell lines; pGL3-control (1.0 μg), pRL-TK (0.1 μg) and siRNA duplex (20 nM) formulated into liposomes were added to each well with a final volume of 600 μl. After a 5-h incubation period, cells were rinsed once with phosphate buffered saline (PBS) and fed with 600 μl of fresh DMEM containing 10% FBS. After a total incubation period time of 22 h, the cells were harvested and lysed with passive lysis buffer (100 μl per well) according to the instructions of the Dual-Luciferase Reporter Assay System (Promega). The luciferase activities of the samples were measured with a MicroLuma*Plus* LB 96V (Berthold Technologies) with a delay time of 2 s and an integration time of 10 s. The following volumes were used: 20 μl of sample and 30 μl of each reagent (Luciferase Assay Reagent II and Stop and Glo Reagent). The inhibitory effects generated by siRNAs were expressed as normalized ratios between the activities of the reporter (*Renilla* or Firefly) luciferase gene and the control (Firefly or *Renilla*, respectively) luciferase gene.

### Statistical analysis

Data were analyzed by using the GraphPad Prism 5 program (GraphPad Software). Where appropriate, the results are expressed as mean ± standard deviation (SD). *P*-values of 0.05 or less were accepted as indicators of statistically significant data. Significant differences were assessed by Student's *t*-tests or by ANOVA to compare three or more groups followed by Bonferroni test. Each experiment was performed in triplicate.

### Analysis of GRB7 protein knockdown by Western blot

SKBR3 cells were seeded 24 h before transfection in 60 mm dishes at a density of 8 x $10^5$ cells/dish in medium containing 10% FBS. Following overnight culture, siRNA duplexes (60 nM per dish) formulated into liposomes were added to each dish with a final volume of 6 ml. Cotransfection of siRNAs was carried out using Lipofectamine 2000. After a 5-h incubation period, the transfection medium was changed to complete medium containing 10% FBS. After a 24-h, 48-h, 72-h and 6-days incubation time, the cells were harvested with PBS and lysed by incubation in RIPA buffer containing protease inhibitors (Roche) at 4°C for 1 h. Cell debris were removed by centrifugation at 8000 xg for 20 min at 4°C, and protein concentration was determined using the BCA assay (Pierce). Thirty micrograms of protein were resolved by SDS electrophoresis and transferred to a poly(vinylidene difluoride) membrane (Immobilon-P, Millipore). The membrane was blocked with 5% skim milk in TBS containing 0.1% Tween for 1 h at r.t. and subsequently probed with anti-GRB7 monoclonal rabbit antibody (Santa Cruz Biotechnology) (diluted 1:1000 in blocking buffer) overnight at 4°C. Anti-rabbit (goat) IgG HRP conjugated secondary antibody (Thermo Scientific, Rockford, IL, USA) was incubated at 1:5000 dilution in the blocking solution for 1 h at r.t. β-Actin was selected as internal control and was detected by incubation with anti-β-actin HRP conjugated antibody (Abcam) (at a dilution

of 1:20 000 in blocking buffer) for 1 h at r.t. The intensities of the bands were analyzed using ImageJ 1.45 software (Rasband, W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, 1997–2011).

### Cell proliferation assay

Interference with *in vitro* growth rate of SKBR3 and MCF7 cells by natural siRNAs and BC6-loop dumbbell was measured using crystal violet. 1.5 x $10^5$ SKBR3 and MCF7 cells were plated in 24-well plates. Twenty-four hours after plating (0 h) cells were transfected with control GRB7 **siRNA III**, BC6-loop dumbbell **13** and non-targeting (anti-*Renilla*) **siRNA I** (60 nM) using Lipofectamine 2000. At different time points (48 or 72 h) cells were fixed with 4% formalin for 10 min, then washed twice with distilled water and stained with 0.1% freshly prepared crystal violet for 30 min. After washing, the stain was dissolved with 10% acetic acid and subsequently quantified by absorbance at 570 nM.

### 3′-exonuclease digestions

Each RNA oligomer (120 pmol) was incubated with Phosphodiesterase I from *Crotalus adamanteus* venom (SNVPD; 340 ng, 10 mU or 680 ng, 20 mU) in a buffer containing 56 mM Tris-HCl (pH 7.9) and 4.4 mM $MgSO_4$ (total volume = 40 μl) at 37°C. At appropriate periods of time, aliquots of the reaction mixture (5 μl) were taken and added to a solution of 0.5 M EDTA, pH 8.0 (15 μl), and the mixtures were immediately frozen. The samples were analyzed by electrophoresis on 15% polyacrylamide gel under non-denaturing conditions. The oligonucleotide bands were visualized with the SYBR Gold reagent.

### Calf intestinal phosphatase-5′-exonuclease digestions

Each RNA oligomer (120 pmol) was incubated with Calf Intestinal Phosphatase (1 mU) in a buffer containing 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate 100 μg/ml BSA, pH 7.4 at 37°C for 30 min. The enzyme was deactivated by heating at 65°C for 10 min and the RNA products were ethanol precipitated. After resuspension with 40 μl of 100 mM sodium acetate pH 6.5 buffer, the RNA samples were treated with Bovine Spleen Phosphodiesterase (10 mU) and incubated at 37°C. At appropriate periods of time, aliquots of the reaction mixture (5 μl) were taken and added to a solution of 0.5 M EDTA pH 8.0 (15 μl), and the mixtures were immediately frozen. The samples were analyzed by electrophoresis on 15% polyacrylamide gel under non-denaturing conditions. The oligonucleotide bands were visualized with the SYBR Gold reagent.

### Stability of RNA and DNA oligonucleotides in PBS containing human serum

Each oligonucleotide (300 pmol) was incubated in PBS containing 50% of human serum (total volume = 75 μl) at 37°C. At appropriate periods of time, aliquots of the reaction mixture (5 μl) were separated and added to a solution of 0.5 M EDTA pH 8.0 (15 μl), and the mixtures were immediately frozen. The samples were run on a 15% polyacrylamide gel under non-denaturing conditions. The oligonucleotide bands were visualized with the SYBR Gold reagent.

### Stability of RNA and DNA oligonucleotides in S100 HeLa cell cytosol extract

Each oligonucleotide (100 pmol) was incubated in 20 mM HEPES-Na pH 7.9, 42 mM ammounium sulfate, 0.2 mM EDTA, 0.5 mM DTT, 20% glycerol buffer containing 10% HeLa cell cytosol extract (S100, human; Jena Bioscience) at 37°C. At appropriate periods of time, aliquots of the reaction mixture (3.7 μl) were separated and added to a solution of 0.5 M EDTA pH 8.0 (6.7 μl), and the mixtures were immediately frozen. The samples were run on a 15% polyacrylamide gel under non-denaturing conditions. The oligonucleotide bands were visualized with the SYBR Gold reagent.

### Molecular dynamics simulations

We simulated four systems, BC2-, BC6- and BC8-dumbbells (**2a**, **2b** and **2c**, respectively), as well as the linear dsRNA duplex analogue (**LdsRNA29**), used as a control system. The approach was similar to that used in previous study (27). Linker structures were first geometrically optimized using Gaussian09 (33) package from which were calculated point charges using RESP calculation (34). Simulations were done using AMBER14 package (35). We used the latest DNA force-field, parmbsc1 (36) with new RNA parameters developed recently in the lab. RNA structures were created using generic NAB module from AMBER14 package. The structures were solvated with TIP3P water molecules (37), neutralized with $Na^+$ ions (38) with an additional 0.15 M of $Na^+Cl^-$ added to the system, in order to approximate experimental conditions. Systems were minimized, annealed and equilibrated following the standard AMBER simulation protocol, with 1 ns of equilibration time. Data were analyzed using Curves+ program (39).

Molecular Dynamics (MD) simulations of the double-stranded RNA containing 3′-terminal 1,2-di(5-methylcytidin-$N^4$-yl)ethane units (**1**; Figure 1) were performed using the same approach.

## RESULTS AND DISCUSSION

### Design of N-alkyl-N capped RNA structures

We started our study by performing MD simulations of a 19 bp siRNA complementary to the 501–519 A-rich site of the *Renilla* luciferase mRNA, where each 3′-terminal dinucleotide overhang (TpT) had been replaced by a BC2 dimer (RNA **1**, Figure 1). Interestingly, starting from an intramolecular stacked conformation, the BC2 dimer progressed to a completely extended form during the course of the simulations, with the final snapshot revealing favorable stacking interactions between the nucleobase of the second subunit of the dimer and the 5′-terminal nucleobase of the complementary strand (Figure 2A), suggesting that we could take advantage of the flexibility of the dimer to create a closed dumbbell-shaped dsRNA with the 3′- and

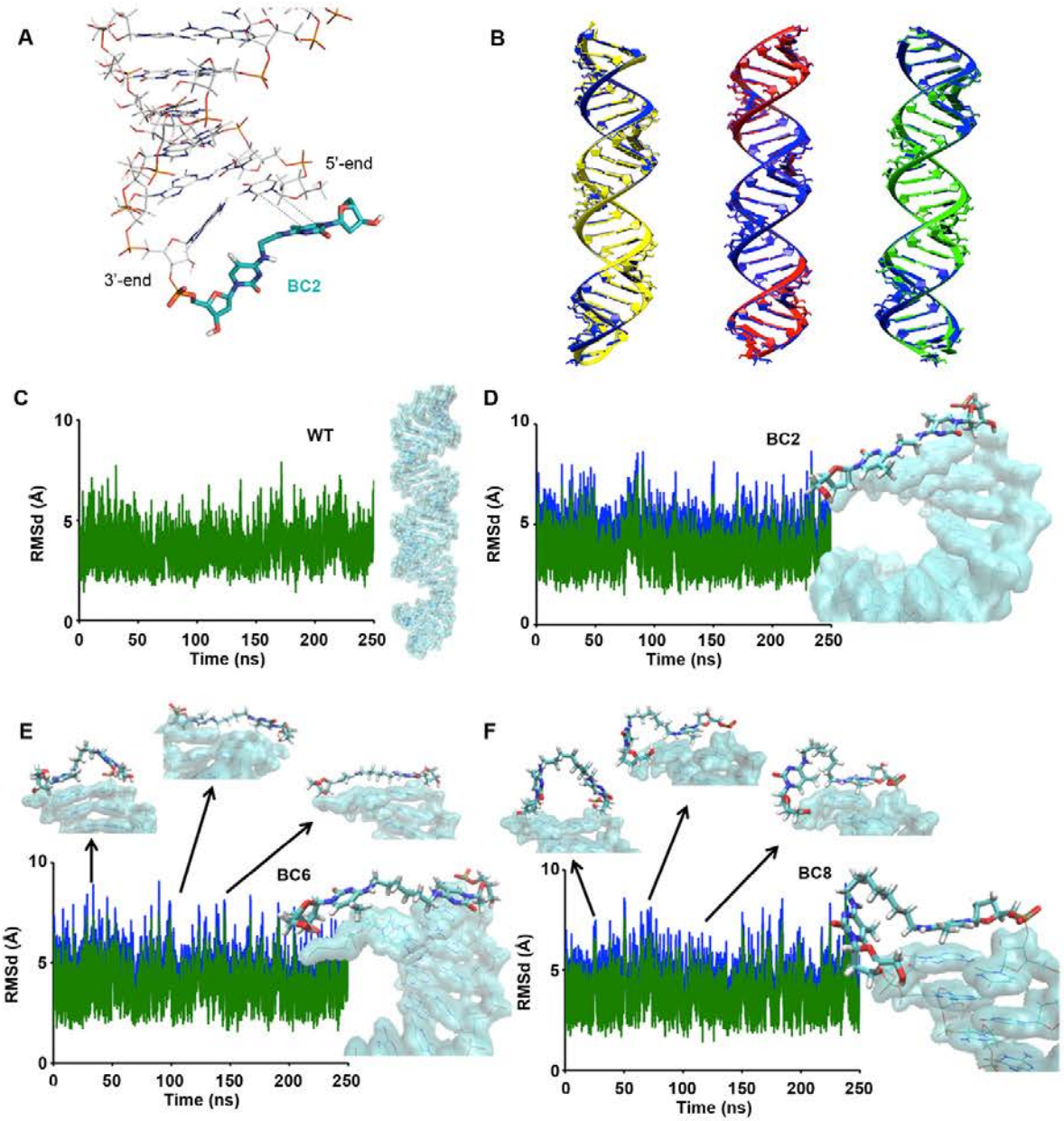



**Figure 2.** (**A**) Representative snapshot from the molecular dynamics (MD) trajectory of the RNA duplex 5′-UUUUUCUCCUUCUUCAGAU**BC2**–3′:3′-**BC2**AAAAAGAGGAAGAAGUCUA-5′ (**1**; Figure 1) showing a completely extended conformation of the BC2 dimer, with stacking interactions between its second subunit and the 5′-terminal nucleobase of the complementary strand. (**B**) Comparison of average structures of BC2- (yellow), BC6- (red) and BC8-dumbbells (green) with linear dsRNA (**LdsRNA29**; blue) from MD simulations. (**C**–**F**) RMSd plots of (**C**) linear **LdsRNA29**, (**D**) BC2-, (**E**) BC6- and (**F**) BC8-dumbbell structures with sequence 5′-AACCAUUUUUUCUCCUUCUUCAGAUUUGA-3′:3′-UUGGUAAAAAAGAGGAAGAAGUCUAAACU-5′. Representative snapshots are shown on the right side of each plot. Blue line shows RMSd evolution considering all residues, while green line shows only the linear part of the dumbbells. Linker structures are drawn with licorice representation, while the linear part is drawn in blue surface representation. Several of BC6- and BC8-dumbbell snapshots are shown to demonstrate the higher flexibility of those structures compared with BC2-linker.

5′-ends connected by a N-alkyl-N bridged nucleoside. To investigate this possibility, we proceeded to perform MD simulations of the predicted closed structure. It has been reported that the efficiency of Dicer processing increases with the length of the stem of the duplex (28). In particular, 29 bp dumbbells with loops of natural nucleotides have been found to act as better substrates for Dicer than shorter analogues. Thus, the final dumbbell-shaped structures were designed to have 29 base pairs targeting the same *Renilla* mRNA region. To identify the appropriate structural conditions needed for minimal constraints within the double-stranded RNA, three dumbbells with different alkyl chain lengths—2, 6 and 8 carbon atoms—were subjected to investigation (BC2-, BC6- and BC8-loop dumbbells **2a-c**, respectively; Figures 1 and 2B, D–F). As a control, we run MD simulations of their linear counterpart (**LdsRNA29**; Figures 1 and 2C). Interestingly, the capping of the dsRNA through BC2, BC6 and BC8 cross-linking did not alter the global geometry of the double helix. Average structures from 250 ns MD simulations (Figure 2B) revealed minimal differences in linear parts of the dumbbells compared with the control simulation (**LdsRNA29**). RMSd plots demonstrated stability of the proposed structures with neglectable effect of designed linkers on the linear part of the dumbbells (Figure 2, panels C–F). Represented snapshots in Figure 2D–F and Supplementary Figure S2 show higher flexibility of BC6- and BC8- linkers with one of its bases being stacked most of time. As expected, BC8- explored bigger conformational space taking into account its higher degrees of freedom. In the case of the BC2- linker, both of its bases kept stacking with the neighboring base-pair during the whole simulation. In terms of helical characteristics linear part of dumbbell structures are indistinguishable to its analogue counterpart, **LdsRNA29** (see Supplementary Figure S1).

### Synthesis of N-alkyl-N capped dumbbell RNAs

Encouraged by these observations, we explored the possibility of formation of these BC2-, BC6- and BC8-loop 29 bp dumbbells (**2a–c**; Figure 1). Our synthetic approach involved double ligation of a pair of hairpins with BCn loops and 5′-phosphorylated dangling ends (Figure 3A). The construction of these nanostructures began by synthesizing the BCn-loop dimeric nucleosides according to methods described previously (see the Supporting Information) (27,40). The resulting dimers were converted to the desired phosphoramidites (Scheme S1), which were incorporated into the BCn-loop internal position of a set of pairs of hairpins (pairs **3a-c**:**4a-c**; Figure 3A) by using an automated DNA/RNA synthesizer and 2′-*O*-TBDMS-protected phosphoramidites of natural ribonucleotides.

Each pair of hairpins (**3a:4a**, **3b:4b** and **3c:4c**) was incubated with T4 RNA ligase 2. Analysis of the reactions of the BC2- and BC6-loop pairs (**3a:4a** and **3b:4b**) by PAGE suggested the formation of the desired closed structures (dumbbells **2a** and **2b**, respectively; Figure 3B) as major products. In both cases, we observed the formation of a new band that migrated more slowly than the starting hairpins. 5′- and 3′-exonuclease assays (Figure 3C and D) confirmed the desired double-ligated closed structure of these RNAs. After isolation (by PAGE) and incubation with the 3′-exonuclease snake venom phosphodiesterase I (SNVPD), both products of ligation (**2a** and **2b**) remained untouched after 2 days of incubation (Figure 3D). The same occurred when they were treated with calf intestinal phosphatase (CIAP) and 5′-exonuclease bovine spleen phosphodiesterase (BSP) (Figure 3C). In contrast to this, pre-ligated hairpins **3a** and **4a** and a synthetic BC2-dumbbell-shaped RNA with an internal nick (**6**; Figure 1; that had been prepared as a control) were hydrolyzed in the presence of BSP and SNVPD (Figure 3C and D). In the case of the ligation of the hairpins containing a more flexible loop (**3c** and **4c**; BC8 loop), we observed the formation of a major product of much lower mobility than the products of the former ligations (**2a** and **2b**) and an almost undetectable band of similar mobility to that of the BC2 and BC6-loop dumbbells **2a** and **2b** (Figure 3B), which could not be isolated. After isolation of the product of lower mobility (major product) treatment with CIAP–BSP and SNVPD led to rapid hydrolysis, (Figure 3C and D), confirming its nicked structure. Taken together, these results suggest that the processes of hairpin formation and ligation are less favorable in the cases of RNAs having more flexible loops (BC8), which could be explained by the lower degree of BC loop-RNA nucleobase stacking predicted by our calculations for the final BC8-loop structure.

Having established conditions for dumbbell formation, we applied them to the synthesis of a second group of dumbbells. With the aim of comparing our dumbbell design with a known biostable dumbbell-shaped RNA (28), we focused on the synthesis of BC2- and BC6-loop versions of an anti-Firefly mRNA 24 bp dumbbell with loops of natural nucleotides (**10**, Figure 1) described in the literature (dumbbells **9a,b**; Figures 1 and 3A). In this case, the reactions of ligation of the corresponding pairs of hairpins (**7a**:**8a** and **7b**:**8b**; BC2- and BC6-loops, respectively) proceeded with high efficiency to give the desired closed structures (**9a,b**; Figure 3B), which were confirmed by the 5′-exo- and 3′-exonuclease assays (Figure 3C and D). This can be attributed to the higher G:C content of this sequence, which might favor the process of formation of the hairpins, due to their higher thermal stability.

### Stability of RNA dumbbells and linear controls in human serum and cell extracts

To determine their biological stability, the BCn-loop dumbbells and their corresponding linear siRNA and nicked controls were incubated in 50% human serum at 37°C. To compare them with known nuclease-resistant dumbbell structures, the serum stability of the 7 nt-loop version of dumbbells **9a,b** [**10** (Figure 1 and Supplementary Figure S3), which was synthesized according to the literature (28)] was also evaluated. Figure 4 shows the degradation profile of the most representative examples: the BC2- and BC6-dumbbells **9a** and **9b**, their siRNA, 1-nicked and 7 nt-loop versions (**siRNAII**, **6** and **10**, respectively) and the 3′-BC2 modified linear analogue **11** (Figure 1), which was synthesized as a control. Unmodified **siRNA II** displayed very low stability. Complete degradation was observed in only 4 h. The 1-nicked analogue **8** was hydrolyzed even faster. Interestingly, the BC-loop structures displayed strongly enhanced stability. The integrated intensities of the gel bands

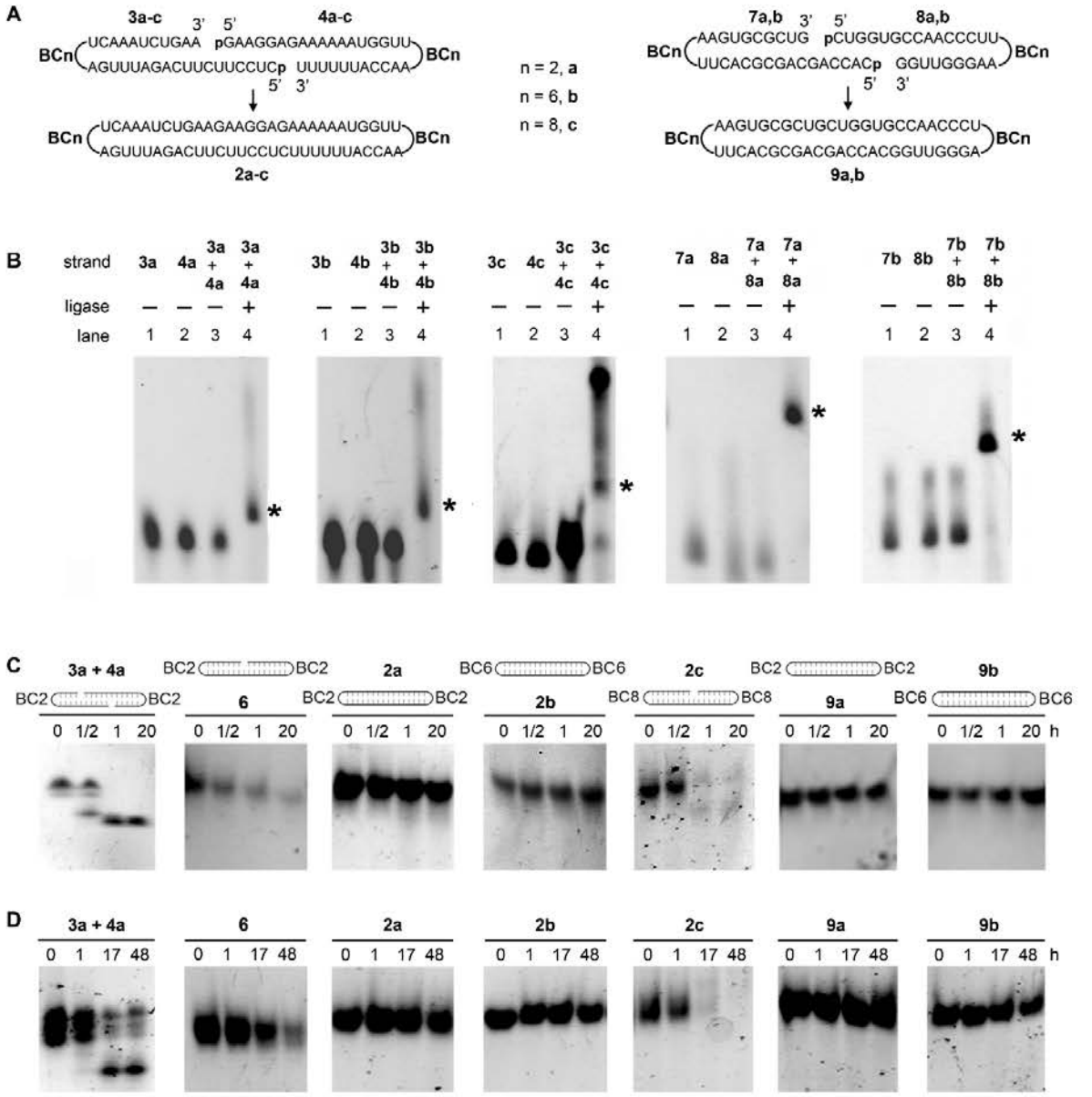



**Figure 3.** (**A**) Reaction of formation of the two sets of luciferase dumbbells used in our studies [targeting *Renilla* (**2a-c**) and Firefly (**9a,b**) mRNAs]. (**B**) Analysis of the ligation reactions by denaturing 15% PAGE containing 25% formamide and 7 M urea in 1X TBE. * means double-ligated dumbbell-shaped structure. (**C** and **D**) Incubation of RNAs with (C) calf intestinal phosphatase and bovine spleen phosphodiesterase and (D) snake venom phosphodiesterase I (SNVPD). Analysis of the exonuclease digestions by 15% non-denaturing PAGE. All oligonucleotides were withdrawn at indicated points, separated and visualized with SYBR Gold.

showed that ~35% and 5% of the BC2-loop dumbbell **9a** remained untouched after 24 and 48 h of incubation, respectively. Its BC6-loop analogue (**9b**) displayed similar stability, with 40% and 5% of intact RNA after 24 and 48 h. In contrast, its 7 nt-loop dumbbell version (**10**) and the 3′-BC2-modified dsRNA **11** were degraded in 12 and 24 h, respectively.

To confirm that the BCn-loops also protect the RNA from nuclease degradation in complex cellular mixtures, we studied the degradation of RNAs **siRNA II**, **6**, **9a**, **9b**, **10** and **11** in 10% S100 cytosolic extract from HeLa cells (Supplementary Figure S4) (41). As observed with human serum (Figure 4), the BCn-loop-modified dumbbells (**9a** and **9b**) were significantly more stable than the corresponding 3′-

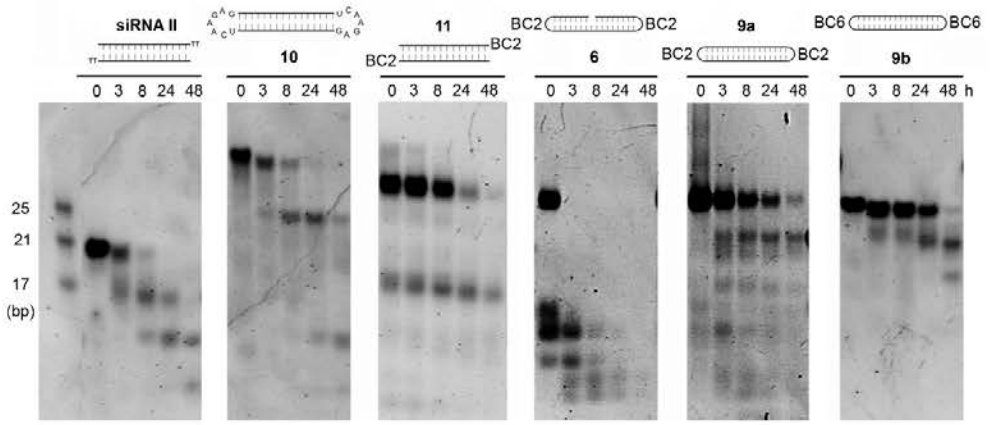



**Figure 4.** 15% non-denaturing polyacrylamide gels of unmodified **siRNA II**, 7 nt-loop dumbbell **10**, 3′-BC2-modified linear dsRNA **11**, 1-nicked structure **6**, BC2-loop and BC6-loop dumbbells **9a** and **9b** (respectively) incubated in phosphate buffered saline (PBS) containing 50% human serum at 37°C. All oligonucleotides were withdrawn at indicated points, separated by 15% native PAGE and visualized with SYBR Gold.

BC2-modified linear analogue and even higher than their 7 nt-loop version (**10**).

### Thermal denaturation and circular dichroism studies

Connection of the ends of the duplex through a BC loop gave a great increase in thermal stability (Supplementary Figure S5). The melting temperatures ($T_m$) of the 24 bp BC2- and BC6-loop dumbbells **9a** and **9b** were >25°C higher than that of their linear counterpart (**LdsRNA24**) (>95°C for **9a,b** versus 73°C for **LdsRNA24**) and slightly higher than that of their 7 nt-loop analogue **10** (~95°C; Supplementary Figure S5A). Similar results were obtained for the second group of 29 bp structures ($T_m$s of 79°C and 77°C for BC2- and BC6-dumbbells **2a** and **2b**, versus 54°C for **LdsRNA29**; Supplementary Figure S5B). Moreover, CD analysis confirmed that the siRNA BC2- and BC6- capping does not alter the structure of the double-helix. The CD spectra of all the synthesized dumbbells superimposed well to those of their siRNA analogues (Supplementary Figure S6), confirming an A-type RNA-like structure.

### RNA recognition and cleavage by Dicer

Dicer is a RNase III-like multi-domain protein composed of a DExH/DEAH RNA helicase domain, a PAZ signature, two neighboring RNAse III-like domains and a dsRNA-binding domain (3–5,42). The helicase domain promotes translocation of the enzyme along the dsRNA and structural rearrangement of the substrate required for cleavage, whereas PAZ binds to the 2 nt 3′ overhang of its dsRNA substrate (43). It has been proposed that canonical Dicers preferentially cleave dsRNAs possessing free termini, by measuring from the 3′-end (bound to the PAZ domain) to the RNase III active site (4). However, despite the absence of free RNA ends in capped structures, dumbbell RNAs can be processed by Dicer (28,29,44). The similarity of the dumbbell stem with a canonical A-like RNA duplex found in our MD simulation suggest also that our constructs should be substrates of Dicer.

In order to investigate the effect of the N-alkyl-N dimeric nucleotides on Dicer recognition, all the synthesized BC-dumbbells were incubated with recombinant human Dicer and subjected to native gel electrophoresis. As shown in Figure 5, all the dumbbells were indeed recognized and digested by the enzyme. In the case of the shorter dumbbells (24 bp; **9a** and **9b**), RNAs shorter than 20 bp were obtained as the major products. Under the same conditions, the longer dumbbell **2b** (29 bp; BC6-loop) was digested almost completely to sequences of about 21 bp, which is in accordance with length of the Dicer products of long dsRNAs (3–5), whereas the less flexible 29 bp BC2-loop analogue **2a** was slowly digested to a mixture of shorter products. As expected (3), the treatment did not change **siRNA I**.

The above described results suggest that Dicer recognizes this class of capped dsRNAs and that a decrease in the lengths of the stem and the alkyl chain increases the degree of resistance to Dicer cleavage, leading to longer digestion times and shorter products. This could be attributed to steric effects caused by the terminal loops over the Dicer active site and to the lower flexibility of N-alkyl-N caps possessing short alkyl chains (see Figure 2D–F). Steric contacts between the loop and Dicer residues might force the enzyme to bind to internal regions located far away from the loops, which would lead to short processing products in the case of shorter duplexes. On the other hand, the higher flexibility predicted for the longer alkyl linkers might favor the structural rearrangements (helicase-promoted) involved in the digestion process.

### RNAi activities of BCn-loop dumbbells targeting *Renilla* and Firefly mRNAs

To investigate if our dumbbell designs can be processed by Dicer not only *in vitro*, but also *in vivo*, we carried out two parallel experiments with the two groups of BCn-loop dumbbells [targeting *Renilla* (**2a,b**) and Firefly (**9a,b**) luciferases], with their corresponding siRNA controls [**siRNA I** and **siRNA II** (Figure 1) as controls for dumbbells **2a,b**

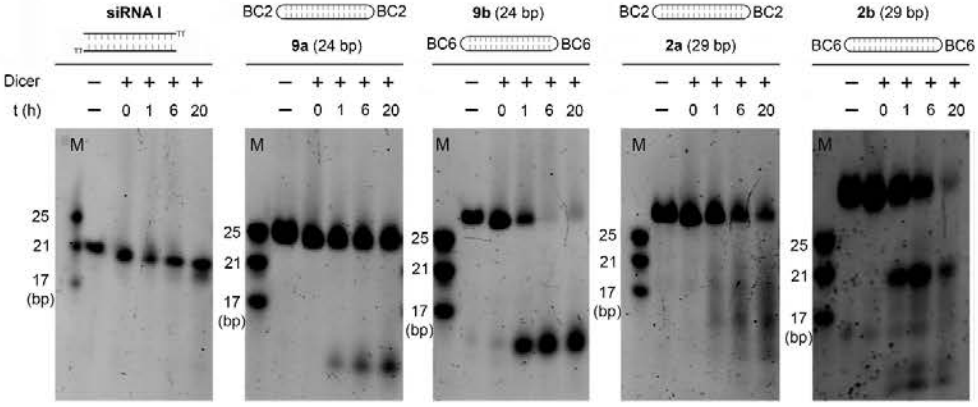



**Figure 5.** Analysis of the Dicer cleavage reaction of RNA dumbbells. Annealed RNAs were incubated with recombinant human Dicer at 37°C, and aliquots were withdrawn at 0, 1, 6 and 20 h. The reaction mixtures were analyzed by 15% native PAGE and visualized with SYBR Gold.

and **9a,b**, respectively], their 7 nt-loop dumbbell analogues [29 bp anti-*Renilla* dumbbell **5** and 24 bp anti-Firefly dumbbell **10**; Figure 1 and Supplementary Figure S3; synthesized by double ligation of 5′-phosphorylated dsRNAs, according to the literature (28)] and with a non-targeting siRNA (**scr**; Figure 1). HeLa cells were co-transfected with the dual reporter plasmids pRL-TK and pGL3 and with the various RNAs (20 nM), and the expression levels of the two luciferase genes were measured 24 h after transfection (Figure 6; panels A and B). Remarkably, all the dumbbells displayed significant target reduction. By comparing *Renilla* and Firefly dumbbells **2a,b** (Figure 6A) and **9a,b** (Figure 6B) with their corresponding siRNA controls [**siRNA I** (Figure 6A) and **siRNA II** (Figure 6B), respectively], we observed that a BC6 loop is better tolerated than the shorter BC2 loop. On the other hand, comparing RNAs **9a,b** (Figure 6B) with **2a,b**, (Figure 6A) and RNA **10** with **5**, we inferred that as the stem length increased (from 24 to 29 bp, respectively), the inhibitory effect (with respect to their unmodified siRNA analogues) significantly improved. Finally, by comparing BC2- and BC6-loop RNAs with their corresponding 7 nt-loop analogues, we observed that the activities of 7 nt-loop dumbbells are higher than that of the BC2-loop RNAs and slightly lower than that of their BC6-loop counterparts. [(98 ± 0.5)%, (78 ± 1.5)%, (91 ± 0.5)% and (88 ± 0.5)% gene knockwdown for the 24 bp RNAs **siRNA II**, **9a**, **9b** and **10**, respectively (Figure 6B), and (88 ± 2)%, (80 ± 0.5)%, (84 ± 1.5)% and (82 ± 0.5)% gene knockwdown for the 29 bp RNAs **siRNA I**, **2a, 2b** and **5** (Figure 6A)]. The most promising design corresponded to the 29 bp BC6-loop dumbbell **2b** (Figure 6A). Although it showed suppression levels slightly lower than that of its siRNA control [(84 ± 1.5)% for **2b** versus (88 ± 2)% for **siRNA I;** Figure 6A], a time-course experiment revealed that it had prolonged activity. The 7 nt-loop dumbbell analogue **5** also had RNAi effect longer than **siRNA I**, but the lifetime of its activity was lower than that of BC6-loop dumbbell **2b**. In a different experiment, HeLa H/P cells stably overexpressing the *Renilla* and Firefly vectors were transfected with **2b**, **5** and **siRNA I** (25 nM), and the RNAi activities of the three RNAs were compared over a period of 6 days (Figure 6C). Very interestingly, on day 6, the suppression activity of the BC6-dumbbell **2b** was significantly higher than that of RNAs **siRNA I** and **5** (7 nt-loop analogue)[(45 ± 1.5%) gene knockdown for **2b** versus (24 ± 2%) and (34 ± 1)% for **siRNA I** and **5**, respectively, $P < 0.0001$ and $P < 0.001$; Supplementary Figure S8].

Taken together, the results obtained from our RNAi experiments *in vivo* are in good agreement with the digestion pattern observed for *in vitro* Dicer cleavage (Figure 5), suggesting that capping the ends of the duplex with BCn loops leads to a slow release of the functional RNAs, permitting longer and then more effective RNAi effect. Our time-course experiments suggest that the presence of 7 nt-loops connecting the ends of the duplex causes a similar effect, although significantly smaller than that observed in the case of the BC6 loop.

### Gene silencing activity of dumbbell RNA targeting GRB7

To further determine the scope of application of our nanostructures, we decided to target an endogenous therapeutically relevant gene. We focused on GRB7 because of its important role in breast cancer biology and its supposed role in anti-cancer drug resistance (31,45–48). GRB7 is an adaptor protein involved in receptor tyrosine kinase signaling which has a key role in HER2 signaling, promoting cell survival and migration (31). It has been reported that expression of GRB7 in the HER2 overexpressed breast cancer subtype contributes to the aggressive nature of the tumor (45) and that HER2 signaling inhibition causes GRB7 upregulation. Moreover, it has also been demonstrated that knockdown of GRB7 by RNA interference potentiates the activity of HER2-targeting drugs (47), leading to decreases in cell proliferation (45).

The most promising dumbbell design, (29 bp stem and a BC6 loop), was used to synthesize a dumbbell targeting the 1019–1037 site of the GRB7 mRNA (GenBank code: BC006535.2), using a previously described sequence (dumbbell **14**, Figure 1 and Figures S3 and S7) (49). For comparison purposes, the corresponding 7 nt-loop ana-

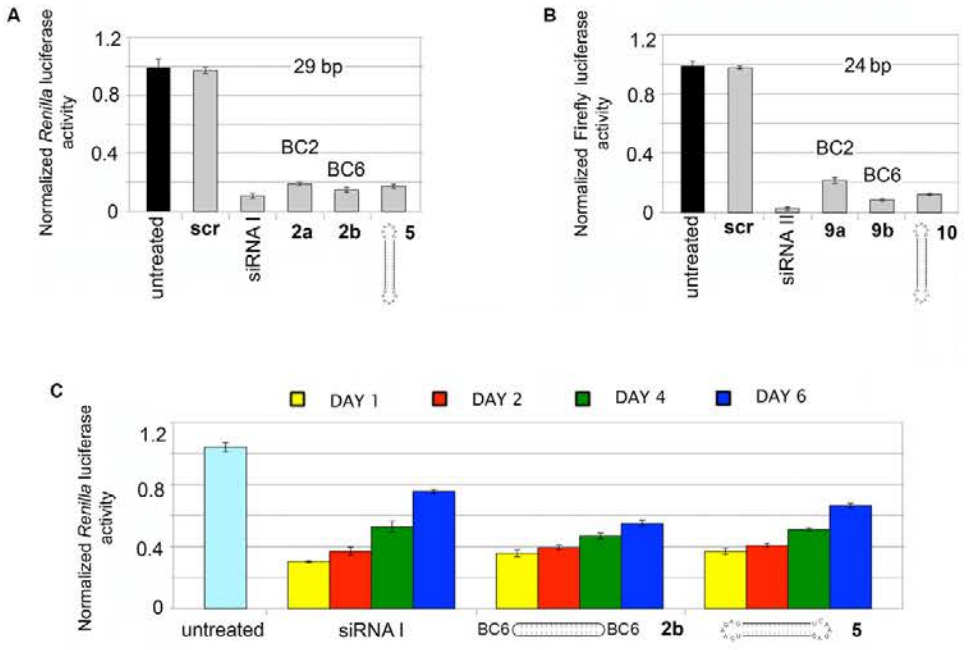



**Figure 6.** (**A** and **B**) Plots of specific activity for BCn- and 7 nt-loop dumbbells and unmodified siRNAs targeting the *Renilla* (**2a,b**, **5** and **siRNA I**; panel A) and the Firefly (**9a,b**, **10** and **siRNA II**; panel B) luciferase mRNAs in HeLa cells. Cells were co-transfected with the dual reporter plasmids pRL-TK and pGL3 and with the various RNAs, and the expression levels of the two luciferase genes were measured 24 h after transfection. Untreated cells: cells treated with plasmids alone. (**C**) Time-course experiments with BC6-dumbbell **2b**, 7 nt-loop dumbbell **5** and **siRNA I** targeting *Renilla* luciferase in HeLa H/P cells stably overexpressing the *Renilla* and Firefly vectors. Untreated cells: cells treated with the transfection agent alone. In all cases, bars indicate standard deviation.

logue (**15**) was also prepared [according to the literature (28); Supplementary Figure S3]. Transfection with the non-targeting control **siRNA I** (anti-*Renilla* luciferase siRNA) served as a negative control. Levels of GRB7 protein after treating the HER2+ breast cancer cell line SKBR3 with anti-GRB7 BC6- and 7 nt-loop dumbbells **14** and **15** and with a known anti-GRB7 siRNA as positive control (**siRNA III**; Figure 1) are shown in Figure 7A. In the three cases (**siRNA III** and dumbbells **14** and **15**), the inhibition of GRB7 protein expression was very small on day 1, whereas after 48 h, the GRB7 expression was reduced by 95%, 92% and 78% in the cases of **siRNA III** and dumbbells **14** and **15**, respectively. Very interestingly, the inhibitory effect of the BC6-loop dumbbell **14** was marked at 72 h and persisted up to 6 days, whereas unmodified **siRNA III**, starts losing efficiency after 72 h. Although the activity of the 7 nt-loop dumbbell **15** was also marked at 72 h (100% GRB7 suppression), it also started to decrease at this point, although significantly slower than unmodified **siRNA III**. On day 6, the suppression activity of the BC6-loop dumbbell **14** was 1.6-fold and 3.8-fold more potent than those of RNAs **15** (7 nt loop) and **siRNA III**, respectively, with cells treated with dumbbell **14** showing a marginal 5% expression of GRB7.

These results were further confirmed using a different siRNA sequence that had as target the 943–961 site of the same mRNA (GRB7) (GenBank code: BC006535.2) (45). The BC6-loop dumbbell directed against this second region (**18**; Figure 1; synthesized by double ligation of BC6-loop hairpins **16** and **17**; Supplementary Figure S7) induced changes in gene expression profiles over the time-course similar to the ones observed for anti-GRB7 BC6-loop dumbbell used in our first series of studies (**14**, see panels A and B in Figure 7). As observed in the first case (dumbbell **14**; Figure 7, panel A), on day 6, this second BC6-loop dumbbell construct (**18**) displayed RNAi activity significantly higher than its corresponding 7 nt-loop analogue [**19**; which was also synthesized (28), for comparison purposes (Supplementary Figure S3)] and much higher than the corresponding unmodified siRNA of the same sequence (**siRNA IV**), which was used as positive control (5% of GRB7 expression for BC6-loop dumbbell **18** versus 15% and 68% for 7 nt-loop dumbbell **19** and unmodified **siRNA IV**, respectively). In contrast to what had been observed for the first series of anti-GRB7 dumbbells (**14** and **15**), both dumbbells **18** and **19** presented their maximum activity at 48 h. BC6-dumbbell remained still completely active at 72 h, whereas the 7 nt-loop analogue **15** underwent a slight decrease in activity at this point (95% and 100% GRB7 suppression at time points 72 and 48 h, respectively). This difference in the delay of activity observed for the two sets of anti-GRB7 dumbbells could be attributed to possible sequence-dependent effects on Dicer cleavage (50,51), which might influence in the rate of processing and release of the active RNA species.

Taken together, our results suggest that the longer-lived RNAi activity observed for the BC6-loop dumbbell design

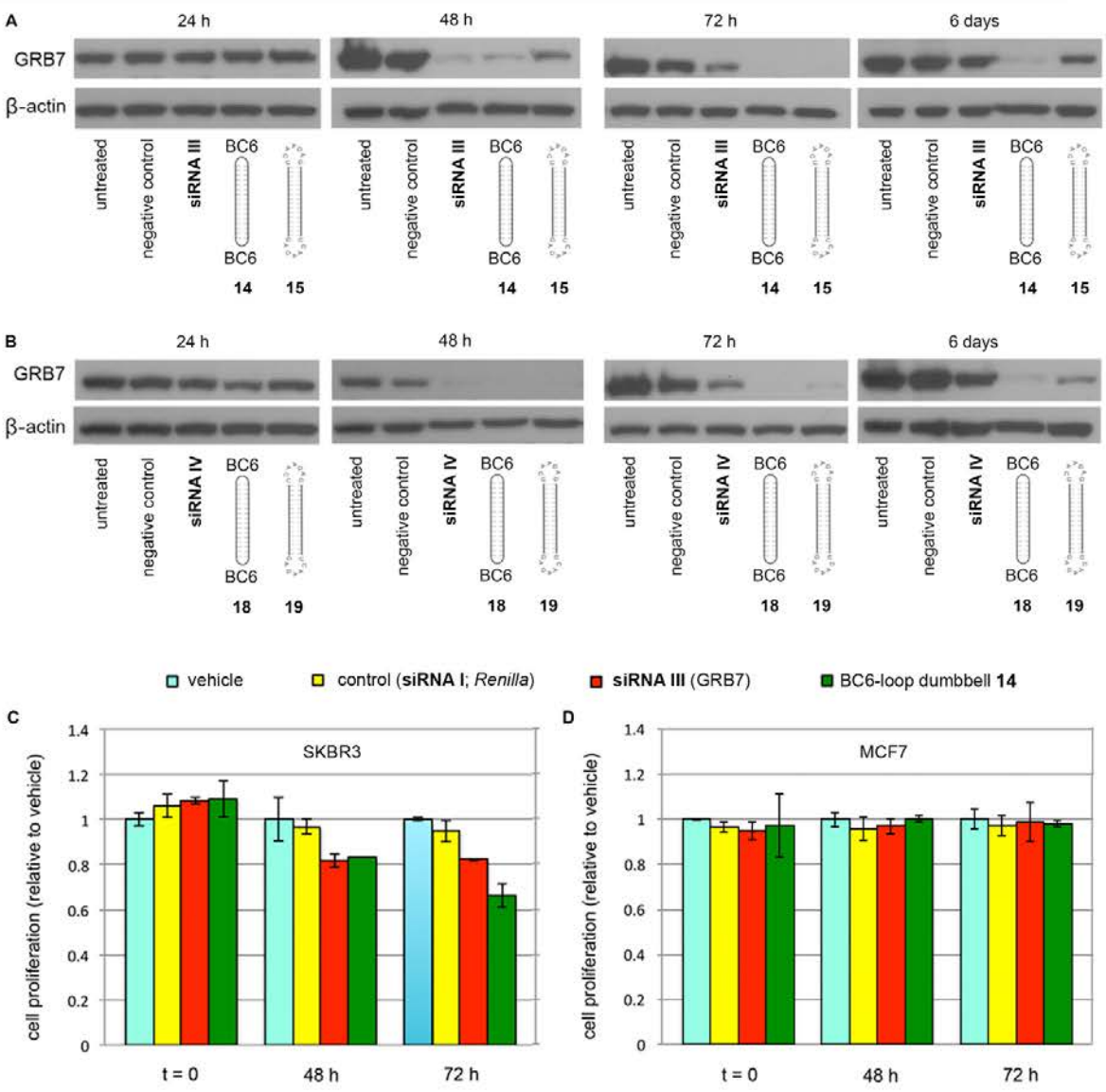



**Figure 7.** (**A** and **B**) Representative immunoblots for GRB7 and β-actin (internal control) from SKBR3 cells treated with (A) BC6-dumbbell **14**, 7 nt-loop dumbbell **15** and **siRNA III** targeting the 1019–1037 site of GRB7 mRNA and with (B) BC6-dumbbell **18**, 7 nt-loop dumbbell **19** and **siRNA IV** targeting the 943–961 site of GRB7 mRNA. In both cases (A and B) non-targeting **siRNA I** (60 nM) was used as negative control. (**C** and **D**) Proliferation assay after transfection with BC6-dumbbell **14**, **siRNA III** targeting the GRB7 mRNA and non-targeting control **siRNA I** (60 nM). The growth of (C) SKBR3 and (D) MCF7 cells were assessed using crystal violet assay and plotted as a percentage of proliferation relative to the vehicle control cells.

is due to the unique BCn-loop modification connecting the ends of the RNA duplex. Although a standard dumbbell RNA design (7 nt-loop; **15** and **19**) confers greater duration of action than a wt-siRNA, (see Figures 6C, 7A and B), our results demonstrate that the BCn-loop design leads to even longer-lived activity.

Finally, we evaluated the effect of dumbbell-mediated GRB7 knockdown on cell proliferation, quantified on 48 and 72 h after RNA transfection using the crystal violet cell viability assays in SKBR3 cells. Viability assays in cells transfected with anti-GRB7 **siRNA III** and dumbbell **14** (targeting the 1019–1037 site) and with the negative (non-targeting) control **siRNA I** (Figure 7C) revealed that 72 h after transfection, proliferation was significantly lower in **14**-transfected cells compared with unmodified **siRNA III**-transfected cells [(66 ± 5)% cell proliferation for **14** versus (82 ± 1)% for **siRNA III**; $P < 0.05$; Supplementary Figure S9]. On the contrary, MCF7 cells, that

do not have HER2 and GRB7 amplification, and express very low levels of GRB7, were unaffected when treated under the same conditions (Figure 7D). Very similar effects were observed when we used the second group of anti-GRB7 RNAs (**siRNA IV**, **18** and **19**; targeting the 943–961 site; see Supplementary Figure S10). In all cases, the dumbbell structures displayed higher anti-proliferative activity than natural siRNAs. In summary, BC6-loop dumbbell structures have excellent long-term inhibitory properties of the synthesis of GRB7—probably related to their higher biostability—which is translated in an excellent antiproliferation profile.

## CONCLUSIONS

In summary, computational studies on dsRNAs containing 3′-terminal bridged N-alkyl-N dimeric nucleotides (BC dimers) have allowed us to design new dumbbell-shaped BC-loop RNA architectures with higher biostability than their linear 3′-BC-modified version and the 7 nt-loop dumbbells described in the literature. The best dumbbell design, corresponding to a BC-loop dimer with a 6 carbon atoms alkyl linker and a stem length of 29 bp, could be used for targeting the relevant GRB7 oncogene in SKBR3 breast cancer cells with longer duration of action and higher antiproliferative effects than an unmodified siRNA. This class of alteration represents a complementary siRNA modification approach that might offer an avenue for the development of new potentially active derivatives. As the BC-loop N-alkyl-N bridged nucleosides have two free hydroxyl groups, they could participate in the conjugation with other biomolecules acting as delivery systems. The study of these and other potential biomedical applications are currently underway in our laboratory. Moreover, as our modification is located at the ends of the RNA duplex, it could be used in combination with conventional siRNA chemistries at selected internal positions (51) to optimize the properties of the siRNA molecule, giving rise to therapeutic RNAs with even higher nuclease resistance.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The HeLa H/P cell line was kindly provided by Dr A. Alagia (IQAC, CSIC, Barcelona, Spain). We thank Dr M. Soler-López, Dr. M. Royo, Prof. Dr R. Eritja, Prof. Dr E. Pedroso, Prof. Dr A. Grandas, Dr A. Aviñó, A. Fàbregas and Dr I. Faustino for their help and valuable comments. IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain). All the experimental work was performed in the EBL laboratory.

## FUNDING

Instituto de Salud Carlos III [Miguel Servet Program, CP13/00211, 205024141 to M.T.]; Spanish MINECO [BIO2012–32869 and BIO2015-64802-R to M.O.]; AGAUR (to M.O.); ERC Council (SimDNA, grant 291433, to M.O.). M.O. is an ICREA Academia fellow. Funding for open access charge: ERC Council [grant 291433 (simDNA)].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Fire,A., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.
2. Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Weber,K. and Tuschl,T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.
3. Kim,D.-H., Behlke,M.A., Rose,S.D., Chang,M.-S., Choi,S. and Rossi,J.J. (2005) Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nat. Biotechnol.*, **23**, 222–226.
4. MacRae,I.J., Zhou,K.H., Li,F., Repic,A., Brooks,A.N., Cande,W.Z., Adams,P.D. and Doudna,J.A. (2006) Structural basis for double-stranded RNA processing by Dicer. *Science*, **311**, 195–198.
5. Park,J.E., Heo,I., Tian,Y., Simanshu,D.K., Chang,H., Jee,D., Patel,D.J. and Kim,V.N. (2011) Dicer recognizes the 5′ end of RNA for efficient and accurate processing. *Nature*, **475**, 201–205.
6. Elbashir,S.M., Lendeckel,W. and Tuschl,T. (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.*, **15**, 188–200.
7. Elbashir,S.M., Martinez,J., Patkaniowska,A., Lendeckel,W. and Tuschl,T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J.*, **20**, 6877–6888.
8. Deleavey,G. and Damha,M.J. (2012) Designing chemically modified oligonucleotides for targeted gene silencing. *Chem. Biol.*, **19**, 937–954.
9. Watts,J.K., Deleavey,M.J. and Damha,M.J. (2008) Chemically modified siRNA: tools and applications. *Drug Discov.Today*, **13**, 842–855.
10. Choung,S., Kim,Y.J., Kim,S., Park,H.-O. and Choi,Y.-C. (2006) Chemical modification of siRNAs to improve serum stability without loss of efficacy. *Biochem. Biophys. Res. Commun.*, **342**, 919–927.
11. Shaw,J.-P., Kent,K., Bird,J., Fishback,J. and Froehler,B. (1991) Modified deoxyoligonucleotides stable to exonuclease degradation in serum. *Nucleic Acids Res.*, **19**, 747–750.
12. Harborth,J., Elbashir,S.M., Vanderburgh,K., Manninga,H., Scaringe,S.A., Weber,K. and Tuschl,T. (2003) Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev.* **13**, 83–105.
13. Wilds,C.J. and Damha,M.J. (2000) 2′-Deoxy-2′-fluoro-β-D-arabinonucleosides and oligonucleotides (2′F-ANA): synthesis and physicochemical studies. *Nucleic Acids Res.*, **28**, 3625–3635.
14. Deleavey,G.F., Watts,J.K., Alain,T., Robert,F., Kalota,A., Aishwarya,V., Pelletier,J., Gewirtz,A.M., Sonenberg,N. and Damha,M.J. (2010) Synergistic effects between analogs of DNA and RNA improve the potency of siRNA-mediated gene silencing. *Nucleic Acids Res.*, **38**, 4547–4557.
15. Chiu,Y.L. and Rana,T.M. (2003) siRNA function in RNAi: A chemical modification analysis. *RNA*, **9**, 1034–1048.
16. Braasch,D.A., Jensen,S., Liu,Y., Kaur,K., Arar,K, White,M.A. and Corey,D.R. (2003) RNA interference in mammalian cells by chemically-modified RNA. *Biochemistry*, **42**, 7967–7975.
17. Czauderna,F., Fechtner,M., Dames,S., Aygün,H., Klippel,A., Pronk,G.J., Giese,K. and Kaufmann,J. (2003) Structural variations and stabilising modifications of synthetic siRNAs in mammalian cells. *Nucleic Acids Res.*, **31**, 2705–2716.
18. Amarzguioui,M., Holen,T., Babaie,E. and Prydz,H. (2003) Tolerance for mutations and chemical modifications in a siRNA. *Nucleic Acids Res.* **31**, 589–595.
19. Eckstein,F. (1985) Nucleoside phosphorothioates. *Annu. Rev. Biochem.*, **54**, 367–402.
20. Eckstein,F. (2000) Phosphorothioate oligodeoxynucleotides: what is their origin and what is unique about them? *Antisense Nucleic Acid Drug Dev.*, **10**, 117–121.
21. Viazovkina,E., Mangos,M.M., Elzagheid,M.I. and Damha,M.J. (2002) Synthesis of 2'-fluoroarabinonucleoside phosphoramidites and their use in the synthesis of 2'F-ANA. *Curr. Prot. Nucleic Acid Chem.* John Wiley & Sons, Vol. **4.15**, pp. 1–22.

22. Watts,J.K. and Damha,M.J. (2008) 2′F-Arabinonucleic acids (2′F-ANA)-History, properties and new frontiers. *Can. J. Chem.*, **86**, 641–656.
23. Watts,J.K., Katolik,A., Viladoms,J. and Damha,M.J. (2009) Studies on the hydrolytic stability of 2′-fluoroarabinonucleic acid (2′F-ANA). *Org. Biomol. Chem.*, **7**, 1904–1910.
24. Martin-Pintado,N., Deleavey,G.F., Portella,G., Campos-Olivas,R., Orozco,M., Damha,M.J. and González,C. (2013) Backbone FC-H···O hydrogen bonds in 2′F-substituted nucleic acids. *Angew. Chem. Int. Ed.*, **52**, 12065–12068.
25. Morrissey,D., Blanchard,K., Shaw,L., Jensen,K., Lockridge,J., Dickinson,B., McSwiggen,J., Vargeese,C., Bowman,K., Shaffer,C. *et al.* (2005) Activity of stabilized short interfering RNA in a mouse model of hepatitis B virus replication. *Hepatology*, **41**, 1349–1356.
26. de Fougerolles,A., Vornlocher,H.P., Maraganore,J. and Lieberman,J. (2007) Interfering with disease: a progress report on siRNA-based therapeutics. *Nature Rev. Drug Discov.*, **6**, 443–453.
27. Terrazas,M., Alagia,A., Faustino,I., Orozco,M. and Eritja,R. (2013) Functionalization of the 3′-ends of DNA and RNA strands with N-ethyl-N coupled nucleosides: A promising approach to avoid 3′-exonuclease-catalyzed hydrolysis of therapeutic oligonucleotides. *ChemBioChem*, **14**, 510–520.
28. Abe,N., Abe,H. and Ito,Y. (2007) Dumbbell-shaped nanocircular RNAs for RNA interference. *J. Am. Chem. Soc.*, **129**, 15108–15109.
29. Wei,L., Cao,L. and Xi,Z. (2013) Highly potent and stable capped siRNAs with picomolar activity for RNA interference. *Angew. Chem. Int. Ed.*, **52**, 6501–6503.
30. Hamasaki,T., Suzuki,H., Shirohzu,H., Matsumoto,T., D'Alessandro-Gabazza,C.N., Gil-Bernabe,P., Boveda-Ruiz,D., Naito,M., Kobayashi,T., Toda,M. *et al.* (2012) Efficacy of a novel class of RNA interference therapeutic agents. *PLoS One*, **7**, e42655.
31. Han,D.C., Shen,T.-L. and Guan,J.-L. (2001) The Grb7 family proteins: structure, interaction with other signaling molecules and potential cellular functions. *Oncogene*, **20**, 6315–6327.
32. Beaucage,S. and Caruthers,M. (1981) Deoxynucleoside phosphoramidites - a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.*, **22**, 1859–1862.
33. Frisch,M.J., Trucks,G.W., Schlegel,H.B., Scuseria,G.E., Robb,M.A., Cheeseman,J.R., Scalmani,G., Barone,V., Mennucci,B. and Petersson,G.A. (2009) *Gaussian 09, revision A. 02*. Gaussian, Inc., Wallingford, Vol. **19**, pp. 227–238.
34. Wang,J., Cieplak,P. and Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.
35. Case,D.A., Babin,V., Berryman,J., Betz,R.M., Cai,Q., Cerutti,D.S., Cheatham,T.E., Darden,T.A., Duke,R.E., Gohlke,H. *et al.* (2014) *AMBER 14*, University of California, San Francisco.
36. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
37. 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.
38. Smith,D.E. and Dang,L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
39. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
40. Noronha,A.M., Noll,D.M., Wilds,C.J. and Miller,P.S. (2002) $N^4$C-Ethyl-$N^4$C cross-linked DNA: Synthesis and characterization of duplexes with interstrand cross-links of different orientations. *Biochemistry*, **41**, 760–771.
41. Uhler,S., Cai,D., Man,Y., Figge,C. and Walter,N.G.J. (2003) RNA degradation in cell extracts: real-time monitoring by Fluorescence Resonance Energy Transfer. *J. Am. Chem. Soc.*, **125**, 14230–14231.
42. MacRae,I.J., Li,F., Zhou,K., Cande,W.Z. and Doudna,J.A. (2006) *Structure of Dicer and mechanistic implicatios for RNAi. Cold Spring Harbor Symposia on Quantitative Biology*, Cold Spring Harbor Laboratory Press, NY, Vol. **LXXI**, pp. 73–80.
43. Hutvágner,G. and Zamore,P.D. (2002) A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, **297**, 2056–2060.
44. Zhang,H., Kolb,F.A., Brondani,V., Billy,E. and Filipowicz,W. (2002) Human Dicer preferentially cleaves dsRNAs at their termini without requirement of ATP. *EMBO J.*, **21**, 5875–5885.
45. Pradip,D., Bouzyk,M., Dey,N. and Leyland-Jones,B. (2013) Dissecting GRB7-mediated signals for proliferation and migration in HER2 overexpressing HER2 breast tumor cells: GTP-ase rules. *Am. J. Cancer Res.*, **3**, 173–195.
46. Bai,T. and Luoh,S.-W. (2007) GRB-7 facilitates HER2/Neu-mediated signal transduction and tumor formation. *Carcinogenesis*, **29**, 473–479.
47. Nencioni,A., Cea,M., Garuti,A., Passalacqua,M., Raffaghello,L., Soncini,D., Zoppoli,G., Pistoia,V., Patrone,F. and Ballestrero,A. (2010) Grb7 upregulation is a molecular adaptation to HER2 signaling inhibition due to removal of Akt-mediated gene repression. *PLoS One*, **5**, e9024.
48. Ramsey,B., Bai,T., Newel,A.H., Troxell,M., Park,B., Olson,S., Keenan,E. and Luoh,S.-W. (2011) GRB7 protein over-expression and clinical outcome in breast cancer. *Breast Cancer Res. Treat.*, **127**, 659–669.
49. Kao,J. and Pollack,J.R. (2006) RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes. *Genes Chromosomes Cancer*, **45**, 761–769.
50. Starega-Roslan,J., Galka-Marciniak,P. and Krzyzosiak,W.J. (2015) Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer. *Nucleic Acids Res.*, **43**, 10939–10951.
51. Collingwood,M.A., Rose,S.D., Huang,L., Hillier,C., Amarzguioui,M., Wiiger,M.T., Soifer,H.S., Rossi,J.J. and Behlke,M.A. (2008) Chemical modification patterns compatible with high potency Dicer-substrate small interfering RNAs. *Oligonucleotides*, **18**, 187–200.

**Supporting Information**

# Rational design of novel N-alkyl-N capped biostable RNA nanostructures for efficient long-term inhibition of gene expression

Montserrat Terrazas,[1,*] Ivan Ivani,[1] Núria Villegas,[1,2] Clément Paris,[3] Cándida Salvans,[1] Isabelle Brun-Heath,[1] and Modesto Orozco[1,4,*]

[1]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Joint IRB-BSC Program in Computational Biology, Barcelona (Spain)

[2] Barcelona Supercomputing Center, Barcelona (Spain)

[3] Department of Organic Chemistry and IBUB, University of Barcelona, Barcelona (Spain)

[4] Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona (Spain)

*To whom correspondence should be addressed. Email: (montserrat.terrazas@irbbarcelona.org, modesto.orozco@irbbarcelona.org)

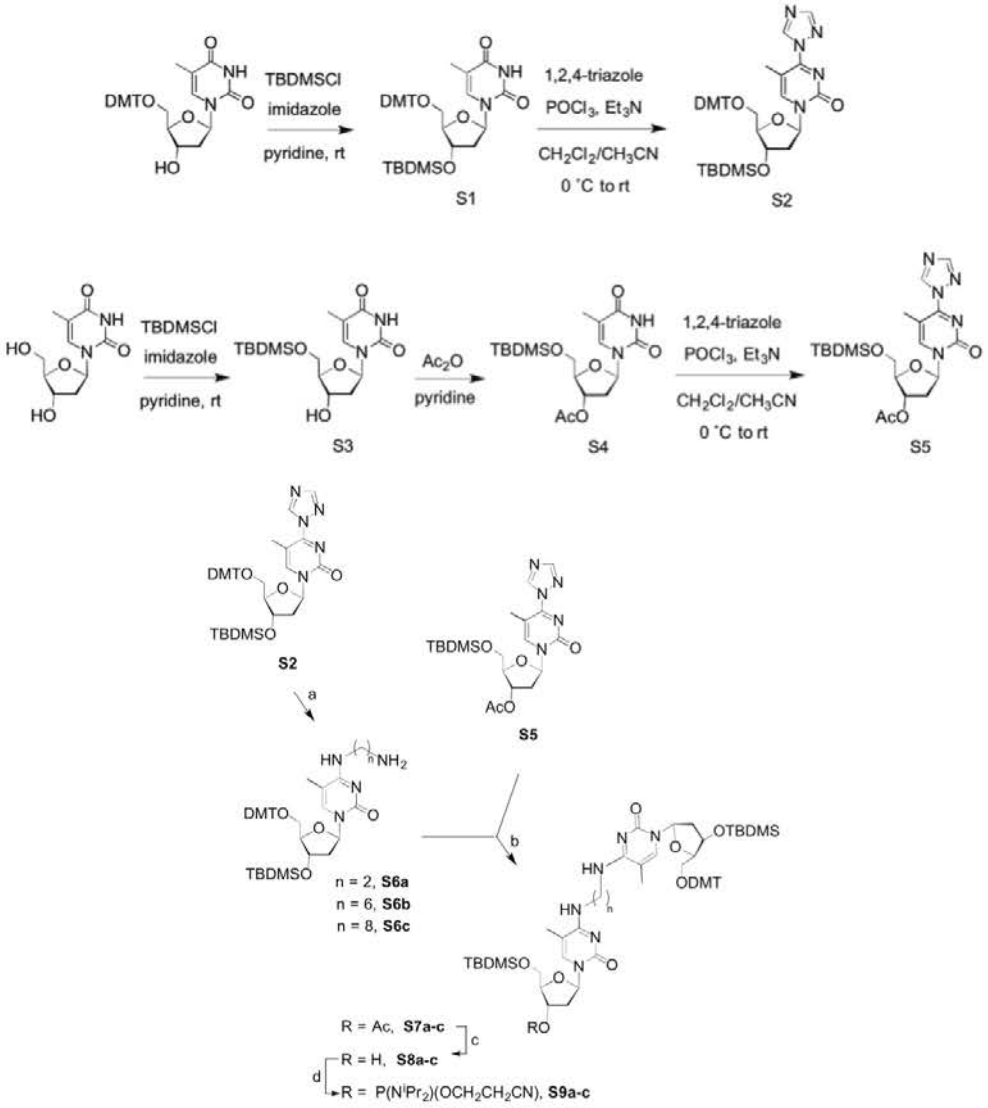
**General experimental methods**

Common chemicals and solvents in addition to 2-cyanoethyl diisopropyl-phosphoramidochloridite were purchased from commercial sources and used without further purification. Anhydrous solvents and deuterated solvents ($CDCl_3$) were obtained from reputable sources and used as received.

Phosphodiesterase I from *Crotalus adamanteus* venom (SNVPD) and Phosphodiesterase II from bovine spleen were purchased from Sigma-Aldrich and used without further purification. Calf intestinal alkaline phosphatase and Turbo Dicer enzyme were purchased from Invitrogen and Genlantis, respectively.
All reactions were carried out under argon atmosphere in oven-dried glassware. Thin-layer chromatography was carried out on aluminium-backed Silica-Gel 60 $F_{254}$ plates. Column chromatography was performed using Silica Gel (60 Å , 230 x 400 mesh). NMR spectra were measured on Varian Mercury-400, Varian-500 or Varian-500 instruments. Chemical shifts are given in parts per million (ppm); *J* values are given in hertz (Hz). All spectra were internally referenced to the appropriate residual undeuterated solvent.

HRMS and ESI spectra were performed on a LC/MSD-TOF (Agilent Technologies) mass spectrometer.

MALDI-TOF spectra were performed using a Perspective Voyager DETMRP mass spectrometer, equipped with nitrogen laser at 337 nm using a 3ns pulse. The matrix used contained 2,4,6-trihydroxyacetophenone (THAP, 10 mg/mL in $CH_3CN$/water 1:1) and ammonium citrate (50 mg/mL in water).




**Scheme S1.** Synthesis of N-alkyl-N dimeric nucleosides BCn

**3'-*O*-*tert*-Butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)thymidine (S1):** 5'-*O*-(4,4'-Dimethoxytrityl)thymidine (Peninsula Laboratories, Inc.) (800 mg, 1.47 mmol) and imidazole (330 mg, 4.85 mmol) were dissolved in DMF (5 mL) and stirred for 5 min at rt. TBDMSCl (377 mg, 2.5 mmol) was added and the reaction stirred for 17 h at rt. The reaction mixture was diluted with EtOAc and washed with water. The aqueous layer was extracted with EtOAc. The organic layers were combined, dried over $Na_2SO_4$, and

evaporated under reduced pressure. The residue that was obtained was purified by column chromatography eluting with $CH_2Cl_2$/MeOH (98:2) to give **S1** as a white foam (968 mg, 99%). $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.68 (m, 1H), 7.45-7.27 (m, 9H), 6.87 (d, *J* = 9.2 Hz, 4H), 6.39 (dd, *J* = 6.4 Hz, *J* = 6.8 Hz, 1H), 4.55 (m, 1H), 4.00 (m, 1H), 3.82 (s, 6H), 3.50 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.30 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 2.36 (ddd, *J* = 3.6 Hz, *J* = 6.0 Hz, *J* = 13.6 Hz, 1H), 2.25 (ddd, *J* = 6.4 Hz, *J* = 6.8 Hz, *J* = 13.6 Hz, 1H), 1.53 (s, 3H), 0.87 (s, 9H), 0.06 (s, 3H), (0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 163.8, 158.7, 150.3, 144.3, 135.6, 135.4, 135.4, 130.0, 130.0, 128.1, 127.9, 127.1, 113.2, 111.0, 86.8, 86.7, 84.8, 72.1, 62.9, 55.2, 46.1, 41.5, 25.7, 17.9, 11.9, -4.7, -4.9. HRMS (ES+): calculated for $C_{37}H_{46}N_2O_7Si$ $[M+H]^+$ 659.3147, found 659.3130.

**3'-*O*-*tert*-Butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-4-(*N*-1-triazolyl)thymidine (S2):** A suspension of 1,2,4-triazole (1.39 g, 20.11 mmol) in a $CH_2Cl_2$/$CH_3CN$ mixture (1:1, 20 mL) was treated with $Et_3N$ (4 mL, 28.91 mmol) and the mixture was stirred at 0 ºC for 5 min. Phosphorous oxychloride (293 µL, 3.14 mmol) was slowly added. After stirring at 0 ºC for 30 min, a solution of **S1** (828 mg, 1.26 mmol) in $CH_2Cl_2$ (3.4 mL) was added. After stirring at rt for 50 min, the starting material was completely converted to the $O^4$-triazolyl intermediate **S2** as evidenced by TLC. The reaction mixture was diluted with $CH_2Cl_2$ and washed with 5% $NaHCO_3$ followed by saturated NaCl. The aqueous layers were extracted with $CH_2Cl_2$. The organic layers were combined, dried over $MgSO_4$ and evaporated under reduced pressure to give the $O^4$-triazolyl intermediate **S2** as a yellow foam (895 mg), which was used without further purification. $^1$H NMR ($CDCl_3$, 400 MHz) δ 9.34 (s, 1H), 8.49 (s, 1H), 8.14 (s, 1H), 7.46-7.32 (m, 9H), 6.89 (d, *J* = 9.2 Hz, 4H), 6.35 (dd, *J* = 5.2 Hz, *J* = 6.4 Hz, 1H), 4.56 (m, 1H), 4.12 (m, 1H), 3.84 (s, 6H), 3.65 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.36 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 2.71 (m, 1H), 2.37 (ddd, *J* = 4.8 Hz, *J* = 6.4 Hz, *J* = 13.6 Hz, 1H), 2.01 (s, 3H), 0.87 (s, 9H), 0.07 (s, 3H), (0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ158.7, 158.1, 153.9, 153.2, 146.6, 144.9, 144.1, 135.2, 130.0, 128.1, 127.9, 127.2, 113.2, 133.1, 105.7, 87.5, 87.1, 86.8, 70.7, 62.0, 55.2, 45.8, 42.2, 25.6, 17.8, 16.4, 8.6, -4.7, -5.0. HRMS (ES+): calculated for $C_{39}H_{47}N_5O_6Si$ $[M+H]^+$ 710.3368, found 710.3351.

**3'-*O*-*tert*-Butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-$N^4$-(2-aminoethyl)-2'-deoxycytidine (S6a):** A solution of $O^4$-triazolyl intermediate **S2** (220 mg, 0.31 mmol) in anhydrous pyridine (14 mL) was treated dropwise with ethylenediamine (890 mL, 13.3 mmol). After stirring for 15 h at rt, the solvents were evaporated. Residual pyridine was removed by co-evaporation with toluene followed by ethanol. The crude product was purified by silica gel chromatography with 20% MeOH and 4% $Et_3N$ in $CH_2Cl_2$ to give **S6a** as a yellow oil (214 mg, 99%). $^1$H NMR ($CDCl_3$, 400 MHz) δ 8.19 (s, 1H), 7.73 (s, 1H), 7.45-7.24 (m, 9H), 6.86 (d, *J* = 8.8 Hz, 4H), 6.73 (bs, 2H), 6.35 (dd, *J* = 6.0 Hz, *J* = 6.1 Hz, 1H), 4.50 (m, 1H), 3.96 (m, 1H), 3.80 (s, 6H), 3.74 (t, *J* = 5.2, 2H), 3.53 (dd, *J* = 2.4 Hz, *J* = 10.4 Hz, 1H), 3.27 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.11 (t, *J* = 5.2, 2H), 2.41 (m, 1H), 2.21 (m, 1H), 1.55 (s, 3H), 0.82 (s, 9H), 0.01 (s, 3H), (-0.05 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 163.5, 158.6, 156.6, 148.9, 144.4, 136.9, 135.5, 130.1, 129.0, 128.1, 127.9, 127.0, 113.2, 113.1, 102.9, 86.5, 86.2, 85.5, 71.2, 62.5, 55.2, 42.0, 40.9, 39.8, 25.7, 17.9, 12.6, -4.7, -5.0. HRMS (ES+): calculated for $C_{39}H_{52}N_4O_6Si$ $[M+H]^+$ 701.3729, found 701.3726.

**3'-*O*-*tert*-Butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-$N^4$-(2-aminohexyl)-2'-deoxycytidine (S6b):** Compound **S2** (239 mg, 0.34 mmol) was dissolved in anhydrous pyridine (14 mL). 1,6-diaminohexane (1.7 g, 14.4 mmol) was slowly added and the reaction stirred for 15 h at rt. The solvents were evaporated and residual pyridine was removed by co-evaporation with toluene followed by ethanol. The crude product was purified by silica gel chromatography with 20% MeOH and 4% $Et_3N$ in $CH_2Cl_2$ to give **S6b** as a yellow oil (239 mg, 94%). $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.78 (s, 1H), 7.51-7.23 (m, 9H), 6.90 (d, *J* = 8.8 Hz, 4H), 6.73 (bs, 2H), 6.41 (dd, *J* = 6.0 Hz, *J* = 6.4 Hz, 1H), 5.41 (bs, 2H), 4.55 (m, 1H), 3.99 (m, 1H), 3.86 (s, 6H), 3.59-3.57 (m, 3H), 3.31 (dd, *J* = 2.8 Hz, *J* = 10.4 Hz, 1H), 2.98-2.95 (m, 2H), 2.47 (m, 1H), 2.26 (m, 1H), 1.75-1.45 (m, 11H), 0.87 (s, 9H), 0.06 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 163.2, 158.6, 156.5, 144.4, 136.7, 135.5, 130.2, 129.0, 128.2, 127.9, 127.0, 113.2, 113.1, 102.1, 86.5, 86.1, 85.5, 71.1, 62.4, 55.2, 42.0, 40.8, 30.0, 29.9, 28.7, 26.1, 25.9, 25.7, 17.8, 12.6, -4.5, -4.9. HRMS (ES+): calculated for $C_{43}H_{60}N_4O_6Si$ $[M+H]^+$ 757.4360, found 757.4364.

**3'-*O*-*tert*-Butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-$N^4$-(2-aminooctyl)-2'-deoxycytidine (S6c):** A solution of $O^4$-triazolyl intermediate **S2** (700 mg, 0.99 mmol) in anhydrous pyridine (14 mL) was treated dropwise with 1,8-diaminooctane (6.1 g, 42.33 mmol). After stirring for 15 h at rt, the solvents were evaporated. Residual pyridine was removed by co-evaporation with toluene followed by ethanol. The crude product was purified by silica gel chromatography with 20% MeOH and 4% $Et_3N$ in $CH_2Cl_2$ to give **S6c** as a yellow oil (695 mg, 90%). $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.80 (s, 1H), 7.51-7.30 (m, 9H), 6.91 (d, *J* = 8.8 Hz, 4H), 6.43 (dd, *J* = 5.6 Hz, *J* = 6.0 Hz, 1H), 5.02 (m, 1H), 4.56 (m, 1H), 4.00 (m, 1H) 3.87 (s, 6H), 3.62-3.56 (m, 3H), 3.31 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.06 (bs, 4H), 2.88-2.84 (m, 2H), 2.50 (m, 1H), 2.28 (m, 1H), 1.69-1.59 (m, 4H), 1.53 (s, 3H), 1.40 (bs, 8H), 0.88 (s, 9H), 0.07 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 163.1, 158.6, 156.3, 144.4, 136.9, 135.6, 130.1, 128.9, 128.2, 127.9, 127.0, 113.2, 113.1, 101.6, 86.5, 86.1, 85.6, 71.0, 62.4, 55.2, 42.1, 41.4, 41.0, 31.7, 29.1, 29.0, 28.9, 26,6, 26.5, 25.6, 17.9, 12.5, -4.6, -5.0. HRMS (ES+): calculated for $C_{45}H_{64}N_4O_6Si$ $[M+H]^+$ 785.4673, found 785.4665.

**5'-*O*-*tert*-Butyldimethylsilylthymidine (S3):** Thymidine (800 mg, 3.3 mmol) and imidazole (522 mg, 7.66 mmol) were dissolved in DMF (11 mL) and stirred for 5 min at rt. TBDMSCl (597 mg, 3.96 mmol) was added and the reaction stirred for 22 h at rt. The reaction mixture was diluted with EtOAc and washed with water. The aqueous layer was extracted with EtOAc. The organic layers were combined, dried over $Na_2SO_4$, and evaporated under reduced pressure. The residue that was obtained was purified by column chromatography eluting with $CH_2Cl_2$/MeOH (95:5) to give **S3** as a white foam (915 mg, 78%). $^1$H NMR ($CDCl_3$, 400 MHz) δ 9.69 (bs, 1H), 7.54 (s, 1H), 6.40 (dd, *J* = 5.6 Hz, *J* = 8.4 Hz, 1H), 4.44 (m, 1H), 4.08 (m, 1H), 3.89 (dd, *J* = 2.4 Hz, *J* = 11.2 Hz, 1H), 3.83 (dd, *J* = 2.4 Hz, *J* = 11.2 Hz, 1H), 3.52 (bs, 1H), 2.40 (ddd, *J* = 2.0 Hz, *J* = 5.6 Hz, *J* = 13.2 Hz, 1H), 2.07 (ddd, *J* = 6.0 Hz, *J* = 8.4 Hz, *J* = 13.2 Hz, 1H), 1.90 (s, 3H), 0.91 (s, 9H),0.10 (s, 3H), 0.09 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 164.1, 150.7, 135.5, 110.9, 87.4, 85.0, 72.5, 63.6, 41.1, 25.9, 18.3, 12.5, -5.4, -5.5.

**3'-*O*-Acetyl-5'-*O*-*tert*-butyldimethylsilylthymidine (S4):** 5'-*O*-*tert*-Butyldimethylsilylthymidine (**S3**, 866 mg, 2.43 mmol) and 4-(dimethylamino)pyridine (30 mg, 0.24 mmol) were dissolved in pyridine (7.5 mL). Acetic anhydride (573 µL, 6.07 mmol) was added and the reaction mixture was stirred at room temperature. After 15 h, the reaction was quenched by addition of MeOH and the solvent was evaporated. The crude product was purified by silica gel column chromatography with 2% MeOH in $CH_2Cl_2$ to give **S4** (888 mg, 92%) as a white foam. $^1$H NMR ($CDCl_3$, 400 MHz) δ 9.01 (bs, 1H), 7.54 (s, 1H), 6.37 (dd, *J* = 5.6 Hz, *J* = 9.6 Hz, 1H), 5.25 (m, 1H), 4.10 (m, 1H), 3.92-3.91 (m, 2H), 2.41 (m, 1H), 2.11 (ddd, *J* = 5.6 Hz, *J* = 9.2 Hz, *J* = 14.2 Hz, 1H), 2.10 (s, 3H), 1.93 (s, 3H), 0.94 (s, 9H),0.14 (s, 6H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 170.6, 163.7, 150.4, 135.0, 111.2, 85.3, 84.7, 75.4, 63.6, 38.0, 25.9, 21.0, 18.3, 12.5, -5.4, -5.5. HRMS (ES+): calculated for $C_{18}H_{30}N_2O_6Si$ $[M+H]^+$ 399.1946, found 399.1946.

**3'-*O*-Acetyl-5'-*O*-*tert*-butyldimethylsilyl-4-(*N*-1-triazolyl)thymidine (S5):** A suspension of 1,2,4-triazole (1.98 g, 28.15 mmol) in a $CH_2Cl_2$/$CH_3CN$ mixture (1:1, 33 mL) was treated with $Et_3N$ (5.64 mL, 40.48 mmol) and the mixture was stirred at 0 ºC for 5 min. Phosphorous oxychloride (410 µL, 4.4 mmol) was slowly added. After stirring at 0 ºC for 30 min, a solution of **S4** (702 mg, 1.76 mmol) in $CH_2Cl_2$ (5.2 mL) was added. After stirring at rt for 1 h, the starting material was completely converted to the $O^4$-triazolyl intermediate **S5** as evidenced by TLC. The reaction mixture was diluted with $CH_2Cl_2$ and washed with 5% $NaHCO_3$ followed by saturated NaCl. The aqueous layers were extracted with $CH_2Cl_2$. The organic layers were combined, dried over $MgSO_4$ and evaporated under reduced pressure to give the $O^4$-triazolyl intermediate **S5** as a yellow foam (792 mg), which was used without further purification. $^1$H NMR ($CDCl_3$, 400 MHz) δ 9.22 (s, 1H), 8.20 (s, 1H), 8.06 (s, 1H), 6.27 (dd, *J* = 5.2 Hz, *J* = 8.0 Hz, 1H), 5.20 (m, 1H), 4.22 (m, 1H), 3.94 (dd, *J* = 2.0 Hz, *J* = 11.6 Hz, 1H), 3.88 (dd, *J* = 2.0 Hz, *J* = 11.6 Hz, 1H), 2.80 (m, 1H), 2.40 (s, 3H), 2.11-2.03 (m, 4H), 0.93 (s, 9H),0.06 (s, 3H), 0.05 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 170.3, 158.0, 153.7, 153.2, 146.2, 144.9, 105.3, 87.8, 86.4, 75.3, 63.3, 45.7, 39.4, 25.7, 20.8, 18.1, 17.1, 8.5, -5.6, -5.7. HRMS (ES+): calculated for $C_{20}H_{31}N_5O_5Si$ $[M+H]^+$ 450.2167, found 450.2178.

**1-{$N^4$-[3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[$N^4$-(3'-*O*-acetyl-5'-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]ethane (S7a):** To a solution of 3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-$N^4$-(2-aminoethyl)-2'-deoxycytidine (**S6a**, 200 mg, 0.29 mmol) in pyridine (3 mL) was added $Et_3N$ (226 µL, 1.62 mmol), followed by 3'-*O*-acetyl-5'-*O*-*tert*-butyldimethylsilyl-4-(*N*-1-triazolyl)thymidine (**S5**, 99 mg, 0.22 mmol). The reaction mixture was allowed to stir at rt for 15 h. The solution was evaporated to dryness. Silica gel column chromatography using 5% MeOH in $CH_2Cl_2$ yielded **S7a** (210 mg, 88%) as a white foam. $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.83 (bs, 1H), 7.68 (s, 1H), 7.53-7.28 (m, 9H), 6.89 (d, *J* = 8.8 Hz, 4H), 6.50 (d, *J* = 5.2 Hz, *J* = 9.2 Hz, 1H), 6.43 (dd, *J* = 6.4 Hz, *J* = 6.0 Hz, 1H), 5.31 (m, 1H), 4.50 (m, 1H), 4.14 (m, 1H), 4.02-3.82 (m, 13H), 3.53 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.30 (dd, *J* = 3.2 Hz, *J* = 10.8 Hz, 1H), 2.55-2.43 (m, 2H), 2.20 (m, 1H), 2.13-2.07 (m, 4H), 2.05 (s, 3H), 1.69 (s, 3H), 0.98 (s, 9H), 0.87 (s, 9H), 0.18 (s, 6H), 0.05 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 75 MHz) δ 170.6, 164.1, 164.0, 158.6, 156.1, 156.0, 144.5, 136.4, 135.9, 135.5, 130.0, 129.9, 128.0, 127.9, 126.9, 113.2, 113.1, 103.4, 102.9, 86.5, 86.2, 85.6, 85.4, 85.0, 75.6, 71.7, 63.5, 62.8, 55.2, 53.4, 42.5, 42.3, 42.0, 38.5, 25.9, 25.7, 21.0, 18.2, 17.9, 13.4, 12.9, -4.7, -5.0, -5.4, -5.6. HRMS (ES+): calculated for $C_{57}H_{80}N_6O_{11}Si_2$ $[M+H]^+$ 1081.5496, found 1081.5501.

**1-{$N^4$-[3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[$N^4$-(3'-*O*-acetyl-5'-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]hexane (S7b):** To a solution of 3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-$N^4$-(2-aminohexyl)-2'-deoxycytidine (**S6b**, 229 mg, 0.3 mmol) in pyridine (5 mL) was added $Et_3N$ (240 µL, 1.72 mmol), followed by 3'-*O*-Acetyl-5'-*O*-*tert*-butyldimethylsilyl-4-(*N*-1-triazolyl)thymidine (**S5**, 91 mg, 0.2 mmol). The reaction mixture was allowed to stir at rt for 15 h. The solution was evaporated to dryness. Silica gel column chromatography using 5% MeOH in $CH_2Cl_2$ yielded **S7b** (226 mg, 99%) as a white foam. $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.61 (bs, 1H), 7.52-7.24 (m, 10H), 6.87 (d, *J* = 8.8 Hz, 4H), 6.51 (d, *J* = 5.2 Hz, *J* = 9.2 Hz, 1H), 6.42 (dd, *J* = 6.4 Hz, *J* = 6.0 Hz, 1H), 5.30 (m, 1H), 4.53 (m, 1H), 4.15 (m, 1H), 3.96-3.95 (m, 2H), 3.83 (m, 6H), 3.51 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.33-3.26 (m, 5H), 2.51 (m, 1H), 2.40 (m, 1H), 2.33 (s, 3H), 2.20 (m, 1H), 2.14 (s, 3H), 2.06 (m, 1H), 1.95 (s, 3H), 1.13 (bs, 4H), 0.96 (s,

9H), 0.87 (s, 9H), 0.18 (s, 3H), 0.17 (s, 3H), 0.05 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 170.6, 163.3, 158.6, 158.5, 157.0, 156.9, 144.5, 135.6, 135.4, 134.9, 130.1, 129.9, 128.1, 127.8, 126.9, 113.2, 113.1, 105.0, 104.7, 86.5, 86.1, 85.4, 85.3, 85.0, 75.8, 71.6, 63.7, 62.8, 55.2, 53.4, 42.1, 42.0, 41.9, 38.7, 28.4, 28.3, 26.9, 25.9, 25.7, 21.0, 18.2, 17.9, 13.8, 13.2, -4.6, -4.9, -5.4, -5.5. HRMS (ES+): calculated for $C_{61}H_{88}N_6O_{11}Si_2$ $[M+H]^+$ 1137.6128, found 1137.6118.

**1-{*N*$^4$-[3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[*N*$^4$-(3'-*O*-acetyl-5'-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]octane (S7c):** To a solution of 3'-*O*-*tert*-Butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-*N*$^4$-(2-aminooctyl)-2'-deoxycytidine (**S6c**, 510 mg, 0.64 mmol) in pyridine (14 mL) was added $Et_3N$ (429 µL, 3.08 mmol), followed by 3'-*O*-Acetyl-5'-*O*-*tert*-butyldimethylsilyl-4-(*N*-1-triazolyl)thymidine (**S5**, 243 mg, 0.54 mmol). The reaction mixture was allowed to stir at rt for 15 h. The solution was evaporated to dryness. Silica gel column chromatography using 5% MeOH in $CH_2Cl_2$ yielded **S7c** (505 mg, 80%) as a white foam. $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.80 (s, 1H), 7.60 (s, 1H), 7.52-7.31 (m, 9H), 6.91 (d, *J* = 9.0 Hz, 4H), 6.55 (d, *J* = 5.2 Hz, *J* = 9.2 Hz, 1H), 6.45 (dd, *J* = 6.4 Hz, *J* = 6.0 Hz, 1H), 5.31 (m, 1H), 5.23 (bs, 1H), 5.00 (bs, 1H), 4.55 (m, 1H), 4.16 (m, 1H), 4.01-3.98 (m, 3H), 3.87 (s, 6H), 3.61-3.59 (m, 5H), 3.31 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 2.61 (m, 1H), 2.51 (m, 1H), 2.28 (m, 1H), 2.15 (s, 3H), 2.06 (m, 1H), 2.00 (s, 3H), 1.85-1.64 (m, 10H), 1.55 (s, 3H), 0.99 (s, 9H), 0.88 (s, 9H), 0.20 (s, 6H), 0.06 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 75 MHz) δ 170.7, 163.1, 163.0, 158.6, 156.2, 144.4, 136.9, 135.6, 135.5, 130.1, 130.0, 128.2, 127.9, 126.9, 113.2, 113.1, 101.9, 101.5, 86.5, 86.1, 85.7, 85.1, 75.6, 71.1, 63.6, 62.4, 55.2, 42.1, 41.0, 40.9, 38.7, 29.2, 29.1, 28.9, 28.8, 26.6, 26.5, 25.9, 25.7, 21.0, 18.3, 17.9, 13.2, 12.6, -4.6, -5.0, -5.4, -5.5. HRMS (ES+): calculated for $C_{63}H_{92}N_6O_{11}Si_2$ $[M+H]^+$ 1165.6441, found 1165.6419.

**1-{*N*$^4$-[3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[*N*$^4$-(5'-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]-ethane (S8a):** Compound **S7a** (203 mg, 0.19 mmol) was dissolved in saturated methanolic ammonia (5.9 mL) and was stirred for 15 h at rt. After removal of the solvents under reduced pressure, compound **S8a** was obtained in 87% yield (169 mg).

$^1$H NMR ($CDCl_3$, 400 MHz) δ 7.67 (s, 1H), 7.61 (bs, 1H), 7.55-7.22 (m, 10H), 6.89 (d, *J* = 8.8 Hz, 4H), 6.48 (dd, *J* = 6.0 Hz, *J* = 7.2 Hz, 1H), 6.42 (dd, *J* = 6.4 Hz, *J* = 6.0 Hz, 1H), 4.51-4.49 (m, 2H), 4.11 (m, 1H), 4.02 (m, 1H), 3.97-3.76 (m, 13H), 3.52 (dd, *J* = 2.8 Hz, *J* = 10.4 Hz, 1H), 3.31 (dd, *J* = 3.2 Hz, *J* = 10.4 Hz, 1H), 2.57-2.43 (m, 2H), 2.20 (ddd, *J* = 6.4 Hz, *J* = 6.8 Hz, *J* = 13.2 Hz, 1H), 2.09-2.02 (m, 4H), 1.65 (s, 3H), 0.97 (s, 9H), 0.87 (s, 9H), 0.16 (s, 3H), 0.15 (s, 3H), 0.06 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 164.1, 163.9, 158.5, 156.3, 156.1, 144.4, 136.5, 136.4, 135.5, 135.5, 130.0, 129.9, 128.1, 127.8, 126.9, 113.2, 113.1, 103.1, 102.9, 86.8, 86.5, 86.2, 85.7, 85.5, 72.1, 71.5, 63.6, 62.7, 55.2, 42.3, 41.9, 25.9, 25.7, 18.3, 17.8, 13.4, 12.8, -4.7, -5.0, -5.4, -5.5. HRMS (ES+): calculated for $C_{55}H_{78}N_6O_{10}Si_2$ $[M+H]^+$ 1039.5391, found 1039.5379.

**1-{*N*$^4$-[3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[*N*$^4$-(5'-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]-hexane (S8b):** Compound **S7b** (236 mg, 0.21 mmol) was dissolved in saturated methanolic ammonia (6.5 mL) and was stirred for 15 h at rt. After removal of the solvents under reduced pressure, compound **S8b** was obtained in 96% yield (218 mg). $^1$H NMR ($CDCl_3$, 400 MHz) δ 7.63 (s, 1H), 7.51-7.27 (m, 10H), 6.87 (d, *J* = 8.8 Hz, 4H), 6.48 (dd, *J* = 6.4 Hz, *J* = 7.2 Hz, 1H), 6.42 (dd, *J* = 6.4 Hz, *J* = 6.0 Hz, 1H), 4.53 (m, 1H), 4.48 (m, 1H), 4.09 (m, 1H), 4.00 (m, 1H), 3.96-3.87 (m, 2H), 3.84 (s, 6H), 3.52 (m 1H), 3.40-3.29 (m, 5H), 2.51 (m, 1H), 2.41 (m, 1H), 2.24-2.06 (m, 5H), 1.86 (s, 3H), 1.22 (bs, 4H), 0.97 (s, 9H), 0.87 (s, 9H), 0.16 (s, 3H), 0.15 (s, 3H), 0.05 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 163.3, 158.6, 156.9, 156.8, 144.4, 135.6, 135.5, 130.1, 130.0, 128.1, 127.9, 126.9, 113.2, 113.1, 104.1, 104.0, 86.7, 86.5, 86.2, 85.6, 85.4, 72.5, 71.5, 63.6, 62.7, 55.2, 42.1, 41.9, 41.7, 28.6, 26.8, 25.9, 25.8, 18.3, 17.9, 13.7, 13.1, -4.6, -4.9, -5.4, -5.5. HRMS (ES+): calculated for $C_{59}H_{86}N_6O_{10}Si_2$ $[M+H]^+$ 1095.6022, found 1095.6006.

**1-{*N*$^4$-[3'-*O*-*tert*-butyldimethylsilyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[*N*$^4$-(5'-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]-octane (S8c):** Compound **S7c** (470 mg, 0.403 mmol) was dissolved in saturated methanolic ammonia (12.5 mL) and was stirred for 15 h at rt. After removal of the solvents under reduced pressure, compound **S8c** was obtained in 98% yield (444 mg).

$^{1}$H NMR ($CDCl_3$, 400 MHz) δ 7.78 (s, 1H), 7.60 (s, 1H), 7.51-7.30 (m, 9H), 6.91 (d, *J* = 8.8 Hz, 4H), 6.49 (dd, *J* = 5.6 Hz, *J* = 7.6 Hz, 1H), 6.44 (dd, *J* = 6.4 Hz, *J* = 5.6 Hz, 1H), 5.40 (bs, 1H), 5.19 (bs, 1H), 4.56-4.51 (m, 2H), 4.14-3.86 (m, 10H), 3.60-3.56 (m, 6H), 3.31 (dd, *J* = 3.2 Hz, *J* = 10.8 Hz, 1H), 2.60 (ddd, *J* = 3.2 Hz, *J* = 6.0 Hz, *J* = 13.2 Hz, 1H), 2.49 (ddd, *J* = 5.2 Hz, *J* = 6.4 Hz, *J* = 13.2 Hz, 1H), 2.27 (m, 1H), 2.10 (m, 1H), 2.01 (s, 3H), 1.66 (bs, 4H), 1.56 (s, 3H), 0.98 (s, 9H), 0.87 (s, 9H), 0.18 (s, 3H), 0.17 (s, 3H), 0.06 (s, 3H), 0.00 (s, 3H). $^{13}$C NMR ($CDCl_3$, 100 MHz) δ 163.1, 158.6, 156.4, 156.3, 149.7, 144.4, 136.8, 136.7, 135.9, 135.6, 130.1, 130.0, 128.1, 127.9, 127.0, 123.7, 113.2, 113.1, 101.8, 101.7, 86.9, 86.5, 86.1, 85.8, 85.6, 72.2, 71.1, 63.6, 62.4, 55.2, 42.1, 41.0, 29.1, 29.0, 28.9, 28.8, 26.6, 26.5, 25.9, 25.6, 18.3, 17.9, 13.2, 12.6, -4.7, -5.0, -5.4, -5.5. HRMS (ES+): calculated for $C_{61}H_{90}N_6O_{10}Si_2$ $[M+H]^+$ 1123.6335, found 1123.6325.

**1-[*N*$^4$-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxycytidylyl)]-2-{*N*$^4$-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxycytidylyl-3'-*O*-(β-cyanoethyl-*N*,*N*'-diisopropyl)-phosphoramidite]}ethane (S9a):** To a solution of alcohol **S8a** (175 mg, 0.16 mmol) in $CH_2Cl_2$ (2 mL) at 0 ºC, DIPEA (44 µL, 0.25 mmol), and 2-cyanoethyl diisopropylphosphoramidochloridite (112 µL, 0.5 mmol) were added. After 15 min the reaction was allowed to reach rt and stirred for 1 h. The reaction was quenched with 5% $NaHCO_3$, extracted with $CH_2Cl_2$, dried over $MgSO_4$ and concentrated. The crude phosphoramidite (mixture of diastereomers) was used without further purification. $^{31}$P NMR ($CDCl_3$, 110 MHz) 148.0, 147.9.

**1-[*N*$^4$-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxycytidylyl)]-2-{*N*$^4$-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxycytidylyl-3'-*O*-(β-cyanoethyl-*N*,*N*'-diisopropyl)-phosphoramidite]}hexane (S9b):** To a solution of alcohol **S8b** (214 mg, 0.2 mmol) in $CH_2Cl_2$ (2.5 mL) at 0 ºC, DIPEA (51 µL, 0.3 mmol), and 2-cyanoethyl diisopropylphosphoramidochloridite (131 µL, 0.59 mmol) were added. After 15 min the reaction was allowed to reach rt and stirred for 1 h. The reaction was quenched with 5% $NaHCO_3$, extracted with $CH_2Cl_2$, dried over $MgSO_4$ and concentrated. The crude phosphoramidite (mixture of diastereomers) was used without further purification. $^{31}$P NMR ($CDCl_3$, 110 MHz) 147.9, 147.8.

**1-[$N^4$-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxycytidylyl)]-2-{$N^4$-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxycytidylyl-3'-*O*-(β-cyanoethyl-*N*,*N*'-diisopropyl)-phosphoramidite]}octane (S9c):** To a solution of alcohol **S8c** (267 mg, 0.24 mmol) in $CH_2Cl_2$ (11.6 mL) at 0 °C, DIPEA (144 μL, 0.83 mmol), and 2-cyanoethyl diisopropylphosphoramidochloridite (127 μL, 0.57 mmol) were added. After 15 min the reaction was allowed to reach rt and stirred for 1 h. The reaction was quenched with 5% $NaHCO_3$, extracted with $CH_2Cl_2$, dried over $MgSO_4$ and concentrated. The crude phosphoramidite (mixture of diastereomers) was used without further purification. $^{31}P$ NMR ($CDCl_3$, 110 MHz) 147.9, 147.8.

**Table S1.** Mass spectrometry analysis of synthesized oligonucleotides

| ON | Sequence | MW calcd. | MW found |
|---|---|---|---|
| **1-top** | 5'-AUCUGAAGAAGGAGAAAAA**BC2**-3' | 6775.0 (+ Na) | 6774.1 (+ Na) |
| **1-bottom** | 5'-UUUUUCUCCUUCUUCAGAU**BC2**-3' | 6407.1 (+ Na) | 6406.1 (+ Na) |
| **6** | 5'pUGGUGCCAACCCU**BC2**AGGGUUGGCACCAGCAG-CGCACU**BC2**AGUGCGCUGC-3' | 16581.6 | 16587.7 |
| **7a** | 5'- pCACCAGCAGCGCACUU**BC2**AAGUGCGCUG-3' | 8923.2 | 8928.1 |
| **8a** | 5'-pCUGGUGCCAACCCUU**BC2**AAGGGUUGG-3' | 8308.8 | 8312.0 |
| **7b** | 5'- pCACCAGCAGCGCACUU**BC6** AAGUGCGCUG-3' | 8979.2 | 8982.0 |
| **8b** | 5'-pCUGGUGCCAACCCUU**BC6** AAGGGUUGG-3' | 8368.8 | 8370.2 |
| **3a** | 5'-pCUCCUUCUUCAGAUUUGA**BC2**UCAAAUCUGAA-3' | 9730.6 | 9736.0 |
| **4a** | 5'-pGAAGGAGAAAAAAUGGUU**BC2**AACCAUUUUUU-3' | 9967.8 (+ Na) | 9966.8 (+ Na) |
| **3b** | 5'-pCUCCUUCUUCAGAUUUGA**BC6**UCAAAUCUGAA-3' | 9809.6 (+ Na) | 9810.5 (+ Na) |
| **4b** | 5'-pGAAGGAGAAAAAAUGGUU**BC6**AACCAUUUUUU-3' | 10023.8 (+ Na) | 10024.5 (+ Na) |
| **3c** | 5'-pCUCCUUCUUCAGAUUUGA**BC8**UCAAAUCUGAA-3' | 9758.5 | 9756.4 |
| **4c** | 5'-pGAAGGAGAAAAAAUGGUU**BC8**AACCAUUUUUU-3' | 9972.9 | 9978.0 |
| **11-top** | 5'-AAGUGCGCUGCUGGUGCCAACCCU**BC2**-3' | 8188.0 | 8190.2 |
| **11-bottom** | 5'-AGGGUUGGCACCAGCAGCGCACUU**BC2**-3' | 8251.9 | 8254.0 |
| **12** | 5'-pGAAGCUUGUUGGGCUUGA**BC6**UCAAGCCCAAC-3' | 9958.8 | 9982.8 (+ Na) |
| **13** | 5'-pAAGCUUCGAAAUGGCCAC**BC6**GUGGCCAUUUC-3' | 9918.8 | 9954.3 (+ Na) |
| **16** | 5'-pCACGUUGGACUCGUUCAC**BC6**GUGAACGAGUC-3' | 9935.0 | 9964.5 (+ Na) |
| **17** | 5'-pCAACGUGUACGUGGUGAC**BC6**GUCACCACGUA-3' | 9958.1 | 9987.6 (+ Na) |

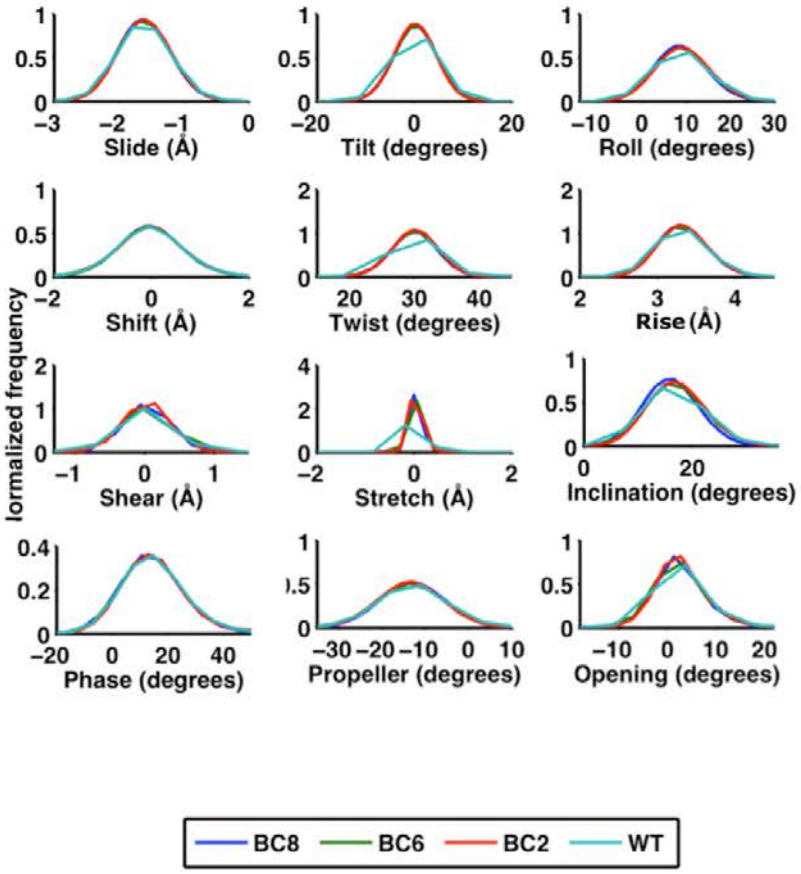



**Figure S1.** Helical parameters distribution at the 29-mer level obtained from linear (WT, light blue) and closed dumbbell with BC2 (red), BC6 (green) and BC8 (blue) linkers. All the frequencies were normalized.

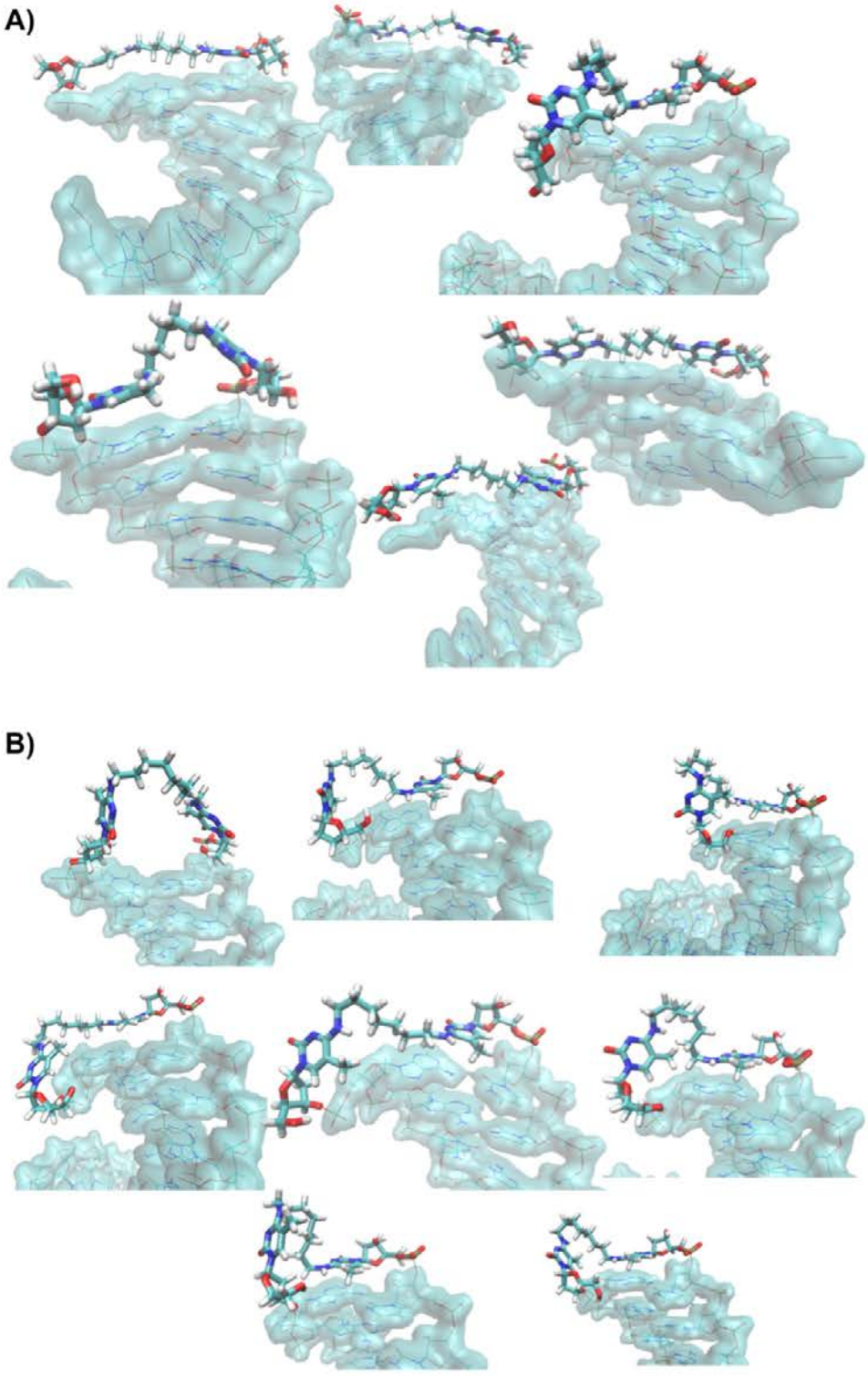



**Figure S2.** Snapshots from BC6 (A) and BC8 (B) dumbbells **2b** and **2c**, respectively.

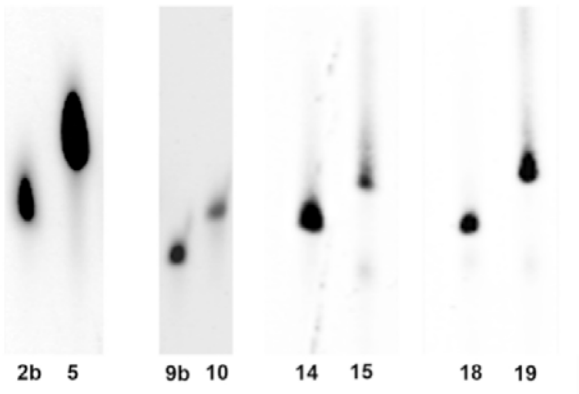



**Figure S3.** Denaturing PAGE analysis (10%PAGE, 25% formamide, 7 M urea in 1X TBE) of pure BC6-loop dumbbells **2b**, **9b**, **14** and **18** and the corresponding 7 nt-loop controls **5**, **10**, **15** and **19** after purification of the ligated products by denaturing PAGE, extraction and dialysis.

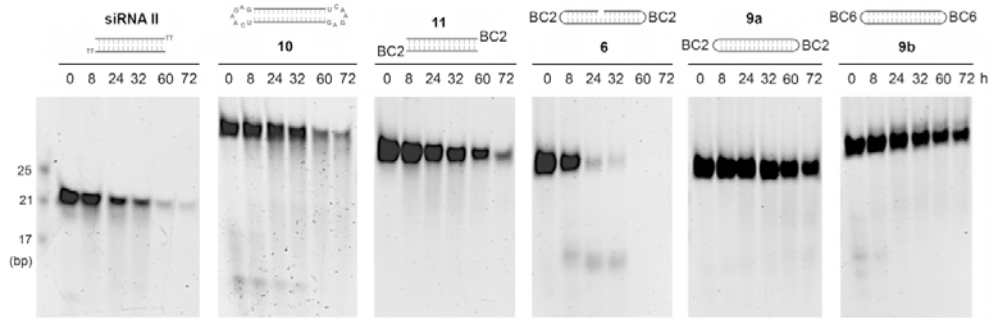



**Figure S4.** 15% non-denaturing polyacrylamide gels of unmodified **siRNA II**, 7nt-loop dumbbell **10**, 3'-BC2-modified linear dsRNA **11**, 1-nicked structure **6**, BC2-loop and BC6-loop dumbbells **9a** and **9b** (respectively) incubated in 20 mM HEPES-Na pH 7.9, 42 mM ammounium sulfate, 0.2 mM EDTA, 0.5 mM DTT, 20% glycerol buffer containing 10% HeLa cell cytosol extract (S100, human; Jena Bioscience) at 37 ºC. All oligonucleotides were withdrawn at indicated points, separated by 15% native PAGE and visualized with SYBR Gold.

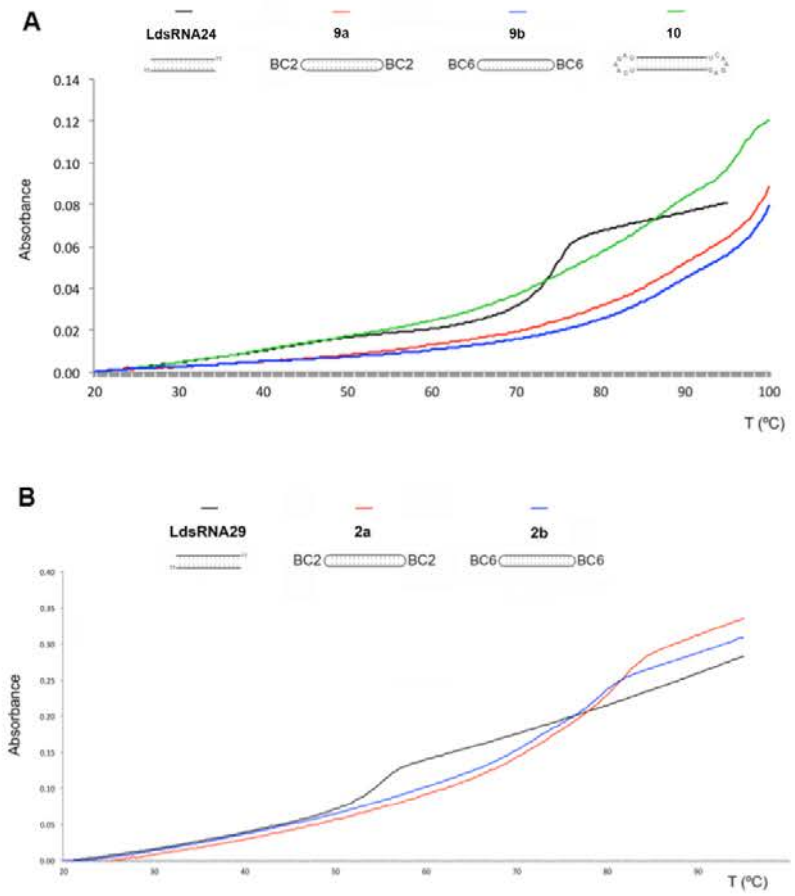



**Figure S5.** Thermal denaturation studies of dumbbell-shaped RNAs

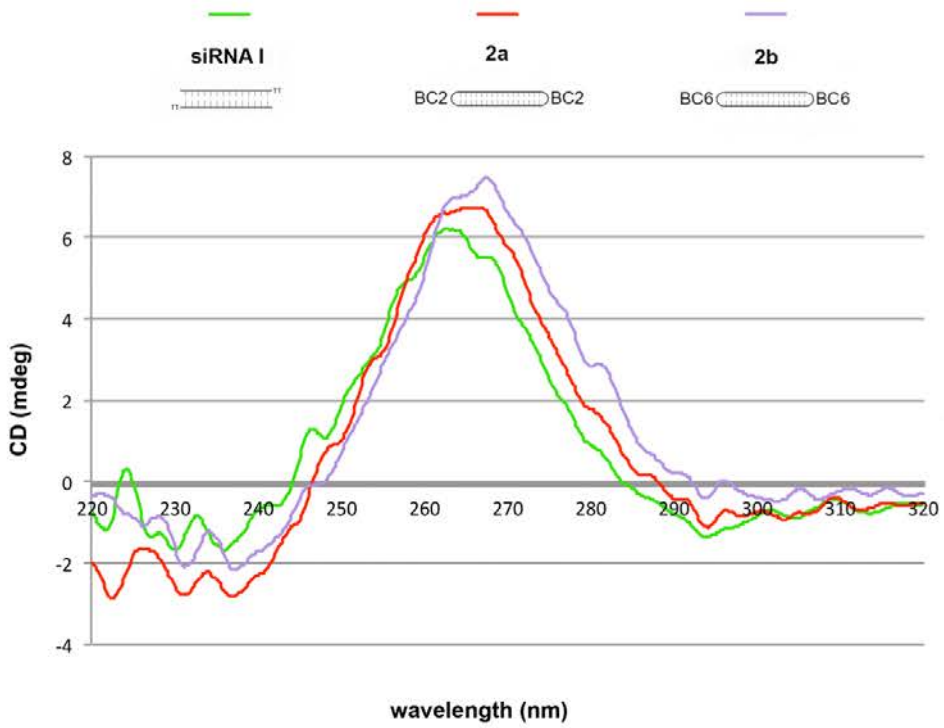



**Figure S6.** Circular dichroism spectra of **siRNA I**, BC2-loop dumbbell **2a** and BC6-dumbbell **2b** at room temperature.

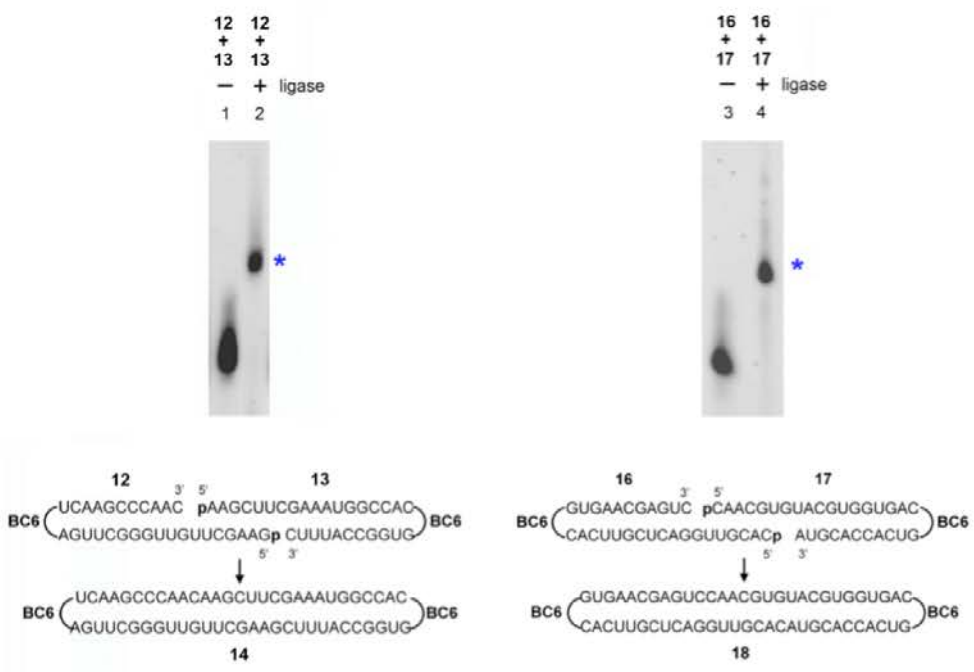



**Figure S7.** Synthesis of RNA dumbbells **14** and **18** targeting the 1019-1037 and 943-961 sites of the GRB7 mRNA. Denaturing 15 % PAGE containing 25% formamide and 7 M urea in 1X TBE. * means double-ligated dumbbell-shaped structures

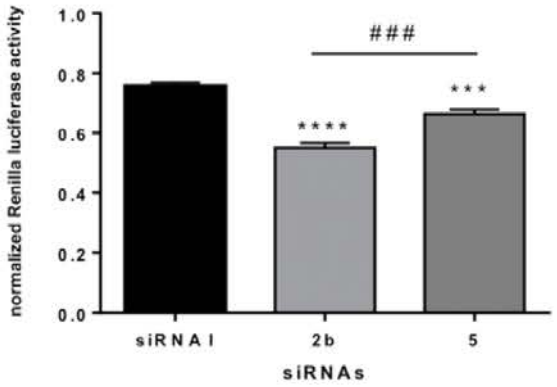



**Figure S8.** Separate gene silencing activities for unmodified **siRNA I** and dumbbells **2b** and **5** targeting the *Renilla* luciferase mRNA on day 6 of the time-course experiment in HeLa H/P cells stably overexpressing the *Renilla* and Firefly vectors (25 nM per well). A Bonferroni test was conducted to evaluate BC6 and 7 nt loops (RNAs **2b** and **5**) to the unmodified control (**siRNA I**) and to evaluate 7 nt loops (RNA **5**) to BC6 loops (RNA **2b**). *** ($P < 0.001$) and **** ($P < 0.0001$) indicates a significant change in *Renilla* expression from unmodified **siRNA 1**. ### ($P < 0.001$) indicates a significant change in *Renilla* expression from 7 nt-loop dumbbell **5**.

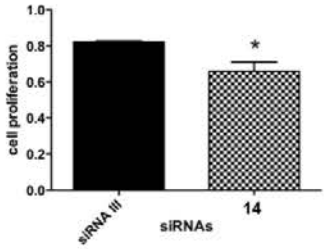



**Figure S9.** Cell proliferation levels 6 days after transfection of SKBR3 cells with BC6-dumbbell **14** and **siRNA III** targeting the GRB7 mRNA (60 nM). Cell proliferation was assessed using crystal violet assay and plotted as a percentage of proliferation relative to the vehicle control cells. Student's *t*-test was conducted to evaluate BC6-loop modification to the unmodified control (**siRNAIII**). * Indicate significant changes in *Renilla* luciferase expression from unmodified siRNA **III** ($P < 0.05$).

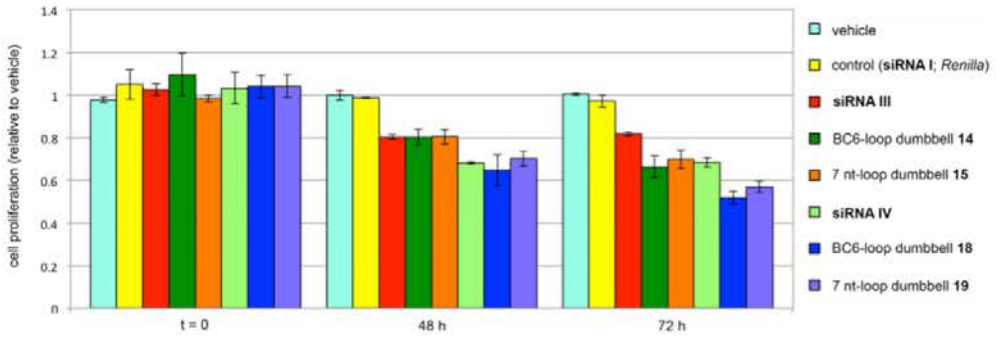



**Figure S10.** Proliferation assay after transfection with BC6-dumbbell **14**, 7 nt-loop dumbbell **15** and **siRNA III** targeting the 1019-1037 site of the GRB7 mRNA, with BC6-dumbbell **18**, 7 nt-loop dumbbell **19** and **siRNA IV** targeting the 943-961 site of the GRB7 mRNA, and with non-targeting control **siRNA I** (60 nM). The growth of SKBR3 cells were assessed using crystal violet assay and plotted as a percentage of proliferation relative to the vehicle control cells.

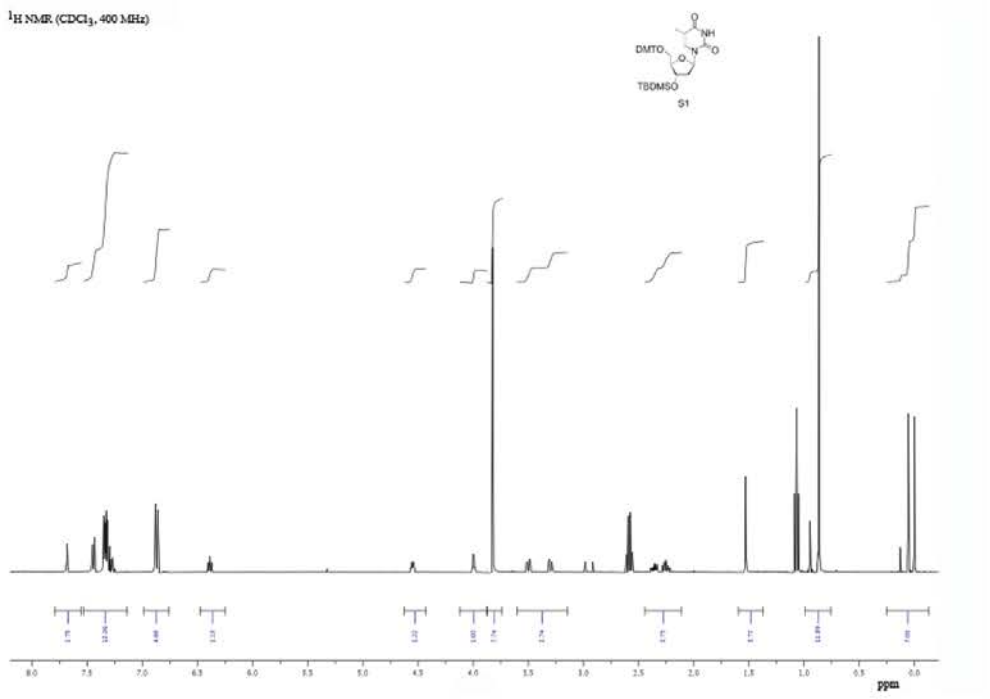



**Figure S11.** $^1$H NMR spectrum of compound **S1**

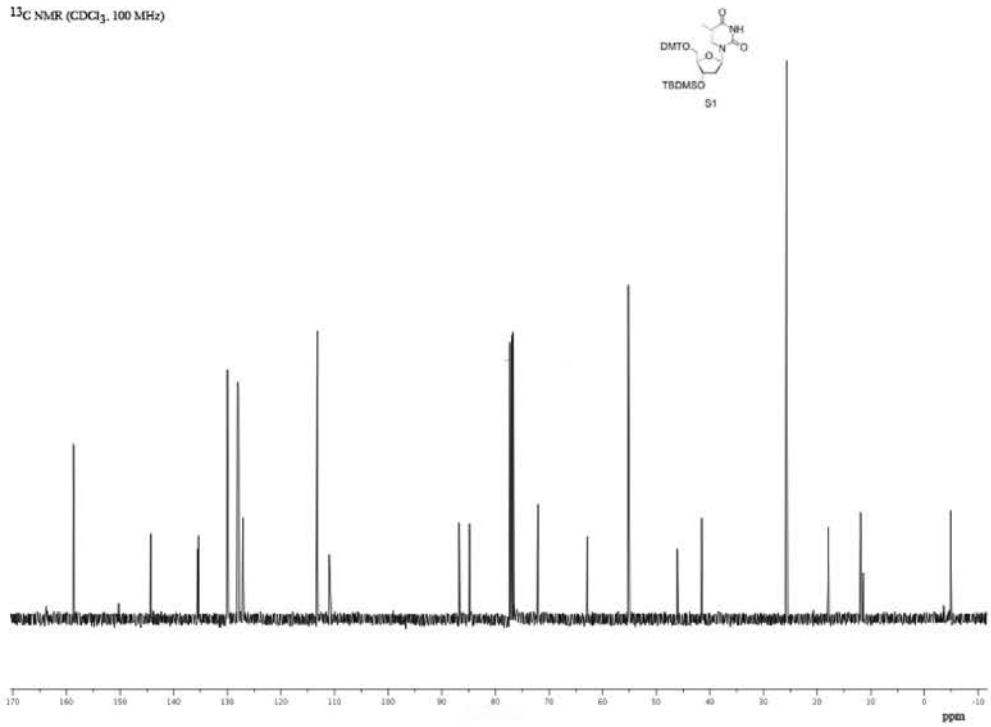



**Figure S12.** $^{13}$C NMR spectrum of compound **S1**

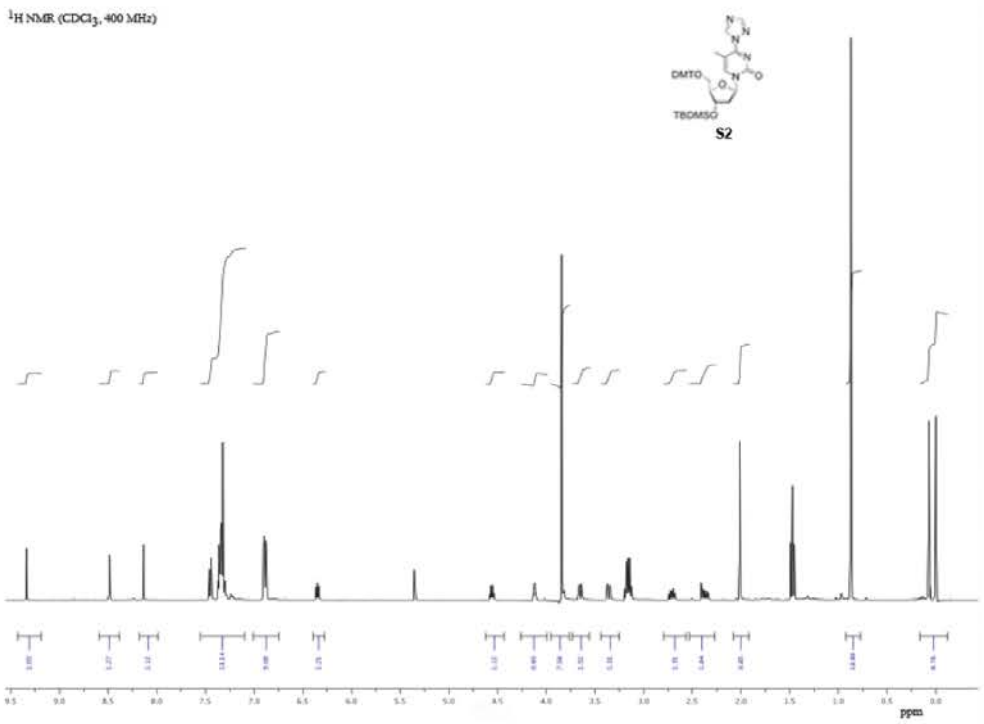



**Figure S13.** $^1$H NMR spectrum of compound **S2**

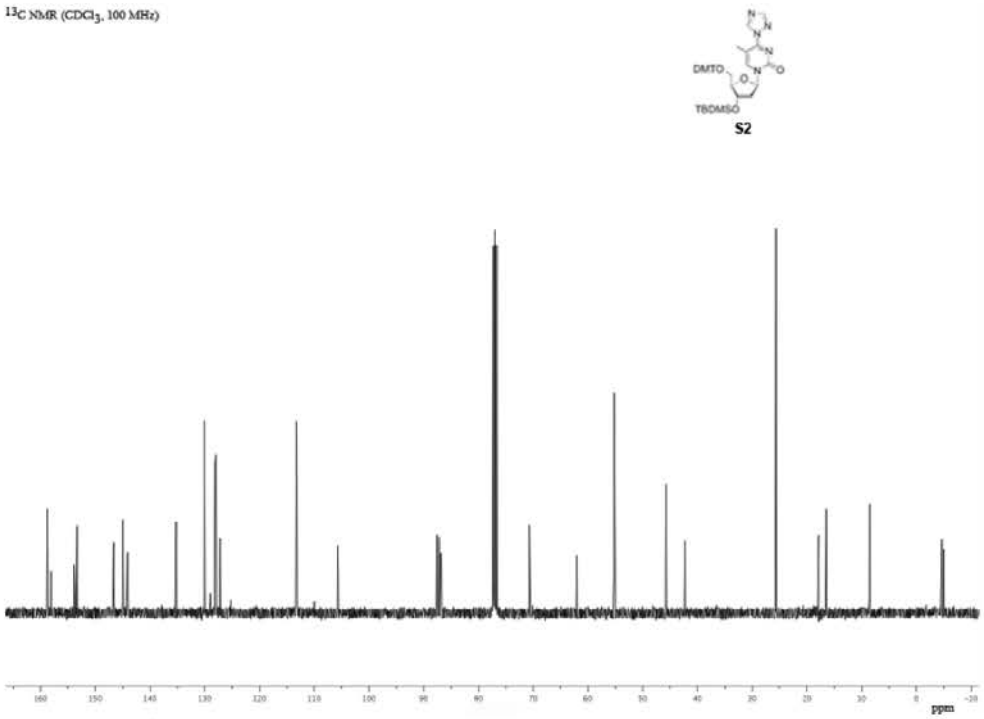



**Figure S14.** $^{13}$C NMR spectrum of compound **S2**

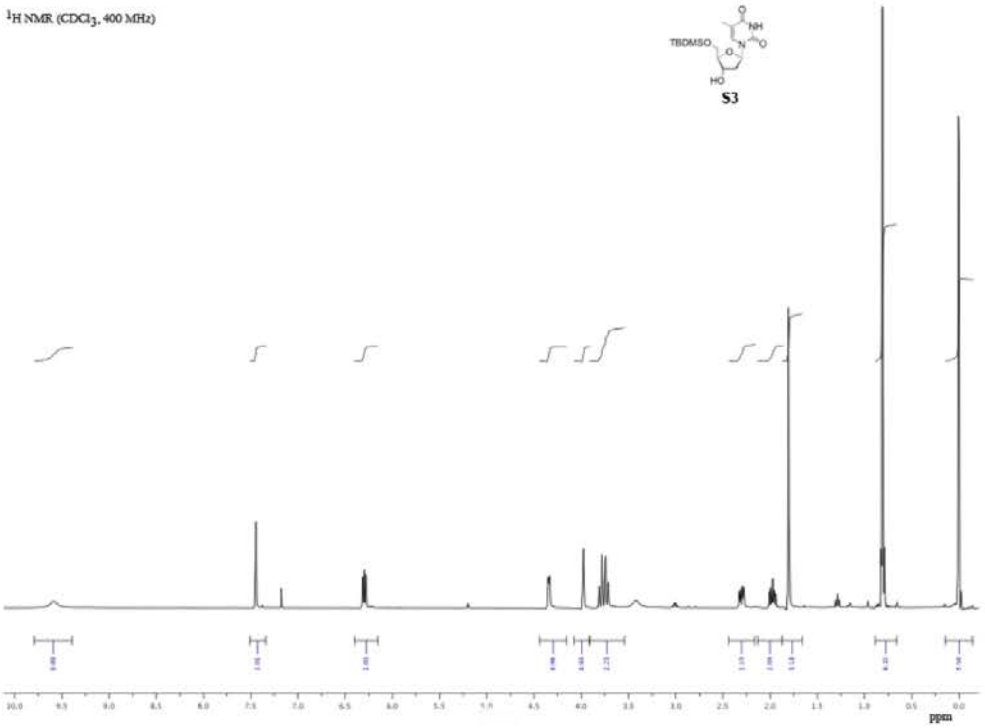



**Figure S15.** $^1$H NMR spectrum of compound **S3**

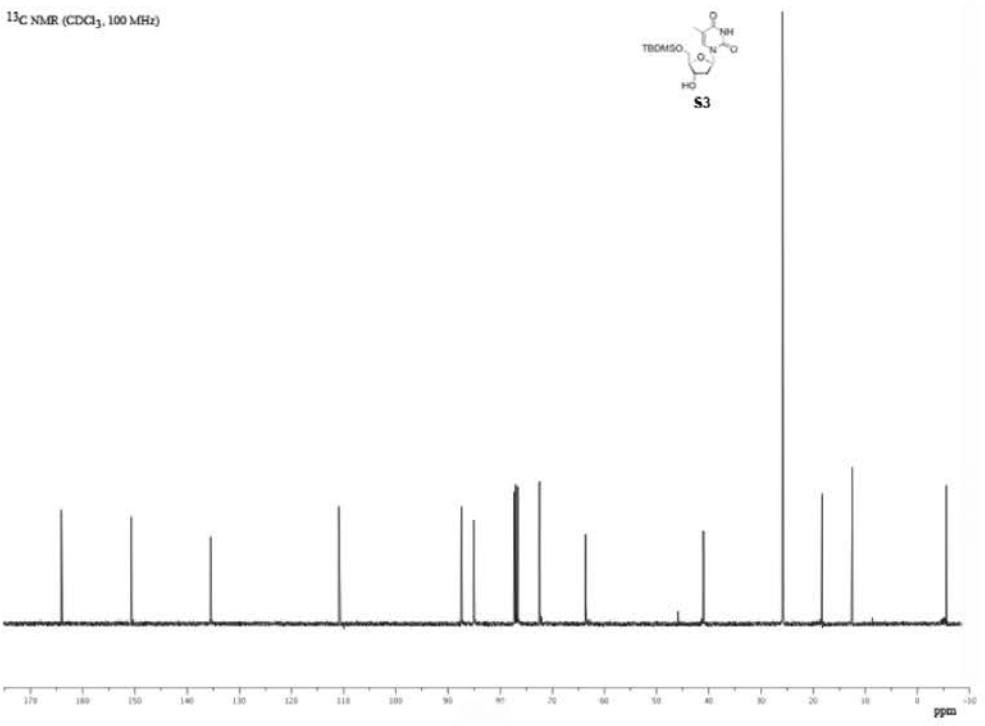



**Figure S16.** $^{13}$C NMR spectrum of compound **S3**

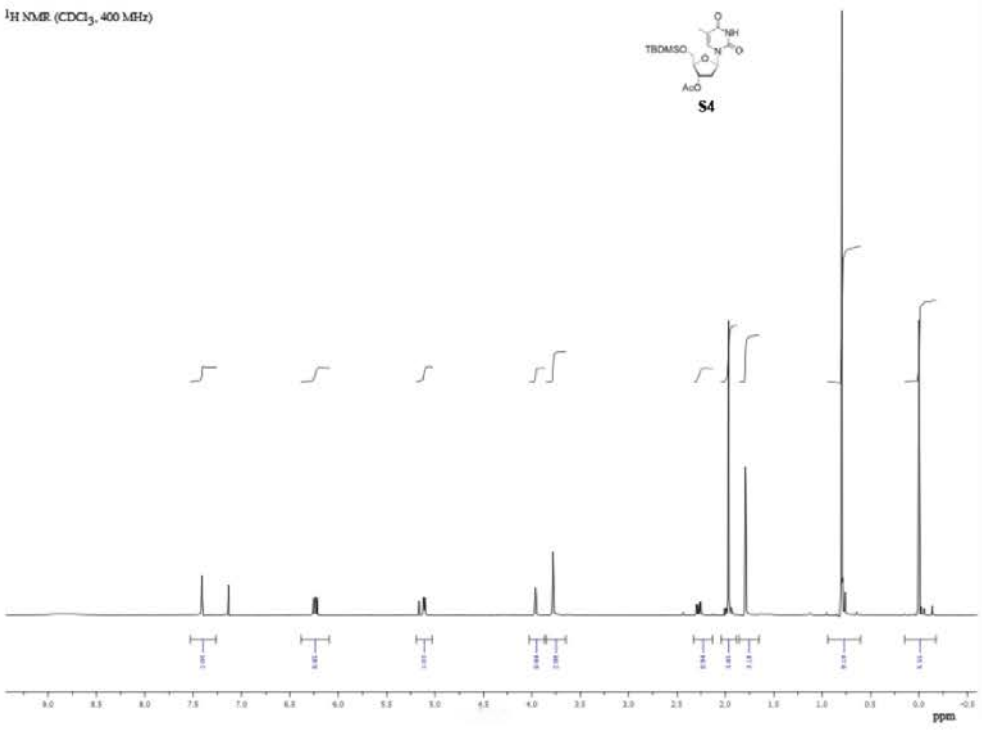



**Figure S17.** $^1$H NMR spectrum of compound **S4**

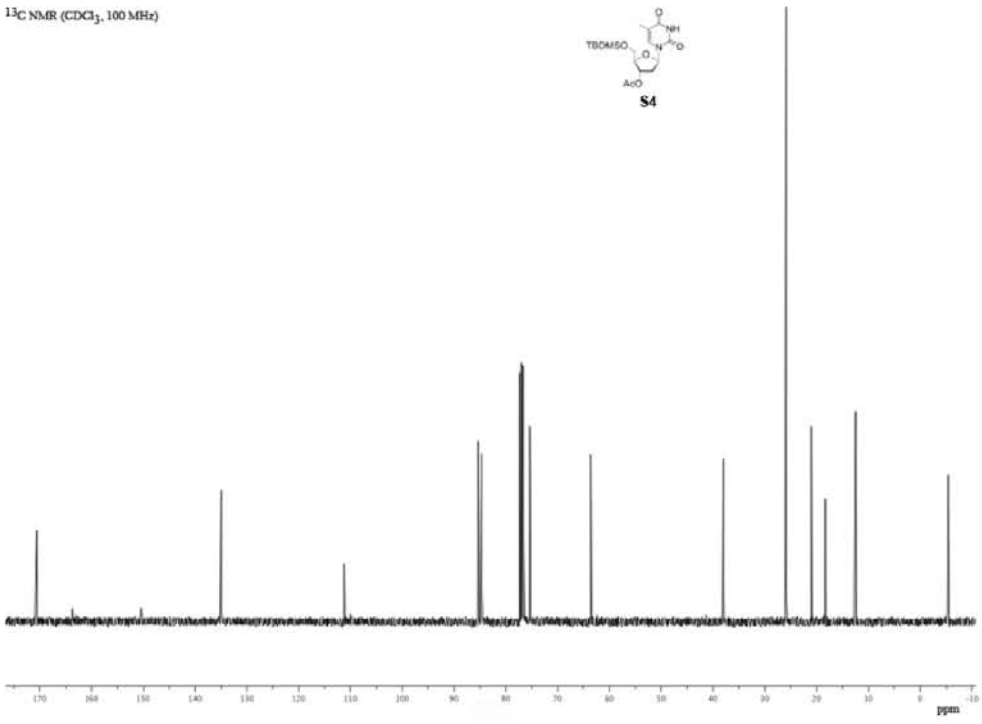



**Figure S18.** $^{13}$C NMR spectrum of compound **S4**

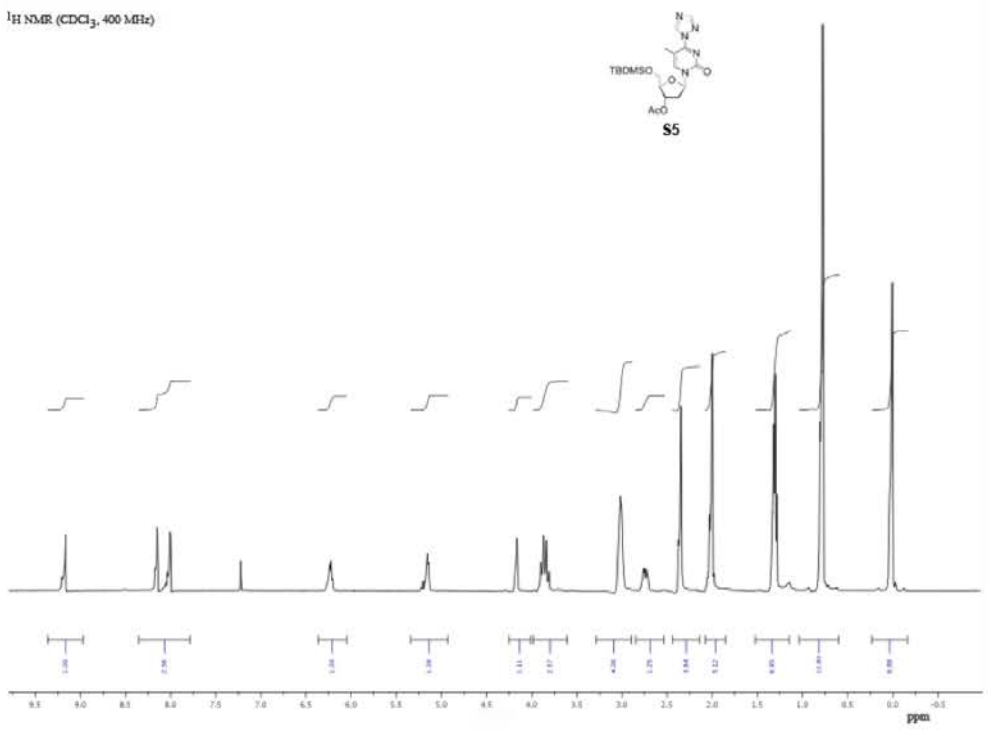



**Figure S19.** $^1$H NMR spectrum of compound **S5**

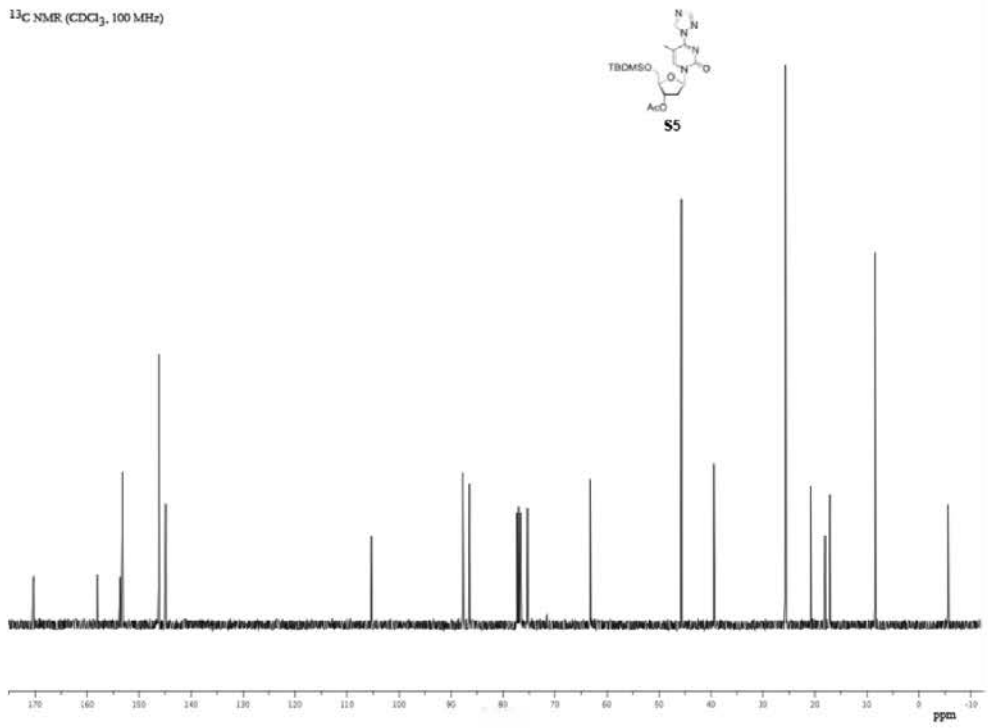



**Figure S20.** $^{13}$C NMR spectrum of compound **S5**

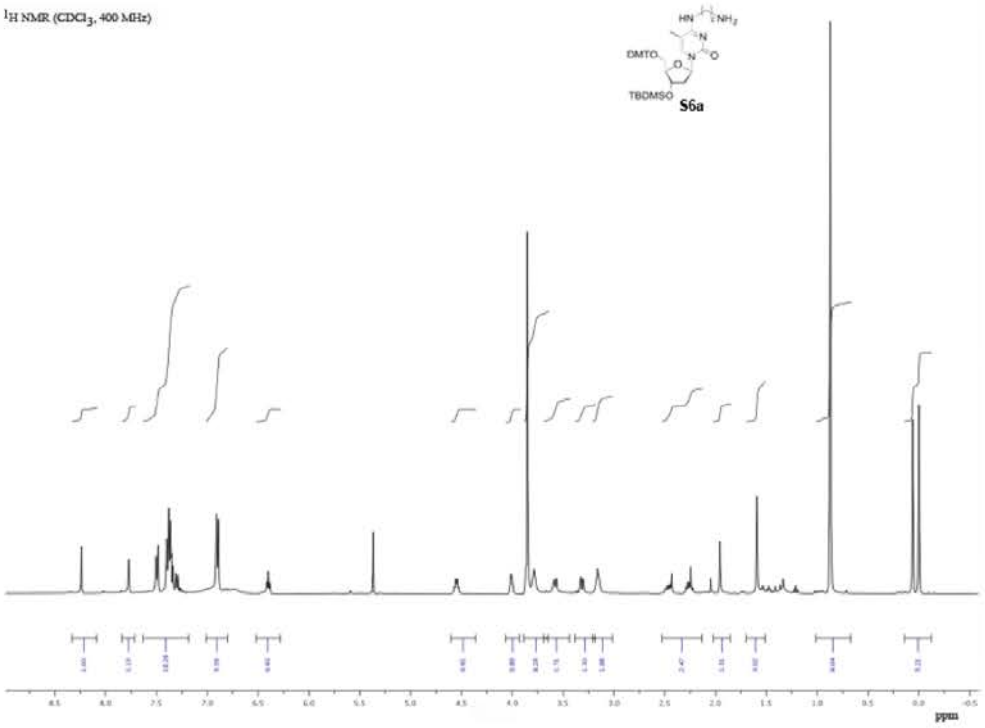



**Figure S21.** $^1$H NMR spectrum of compound **S6a**

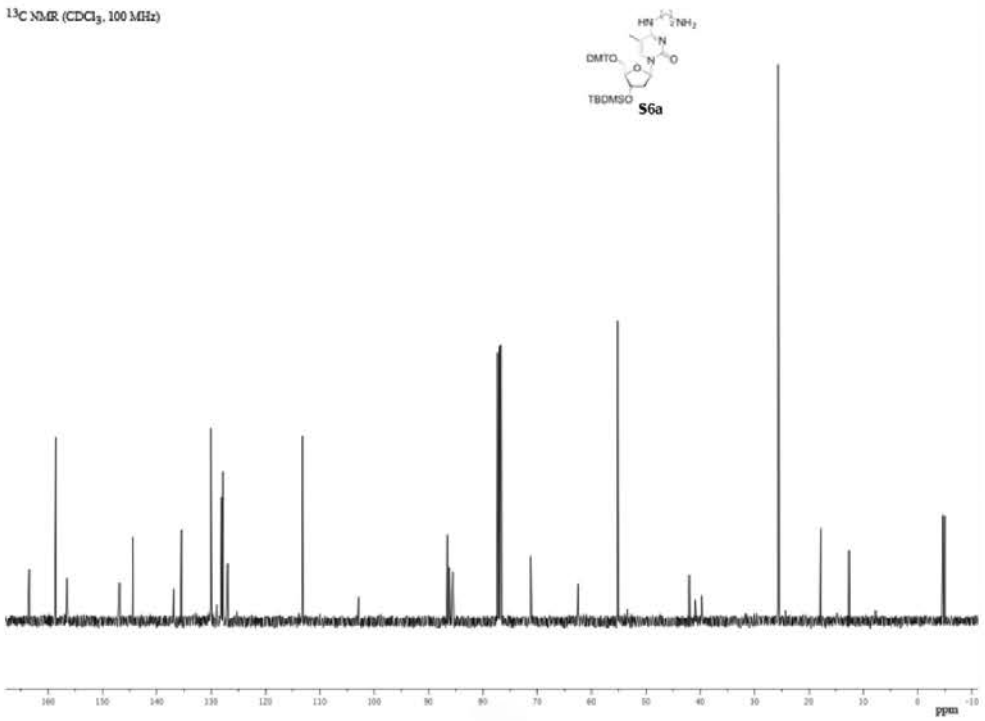



**Figure S22.** $^{13}$C NMR spectrum of compound **S6a**

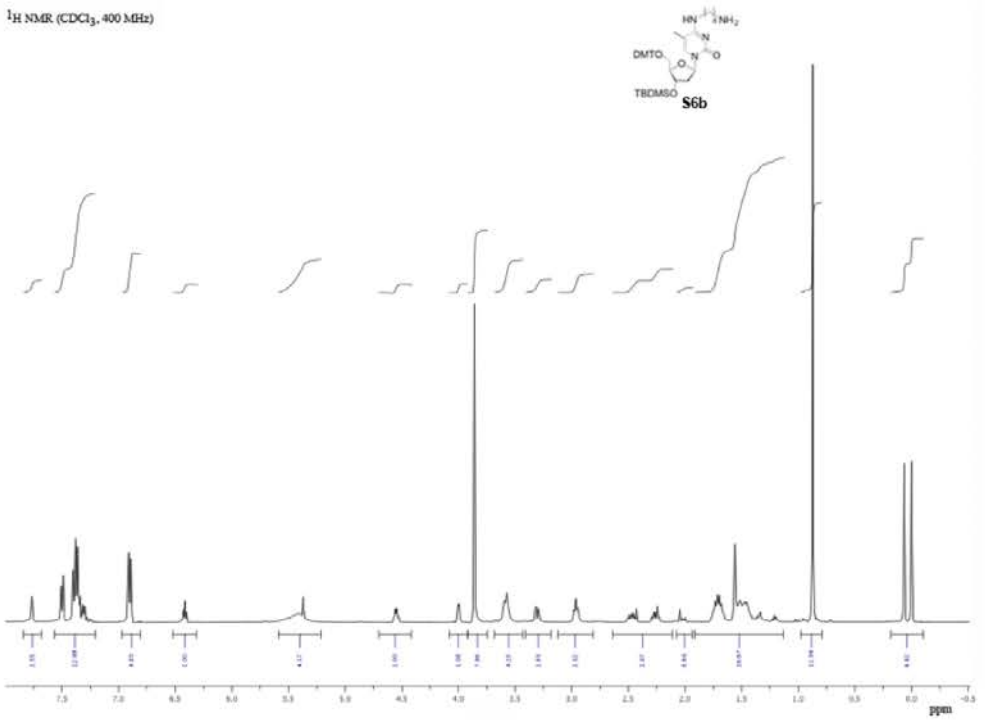



**Figure S23.** $^1$H NMR spectrum of compound **S6b**

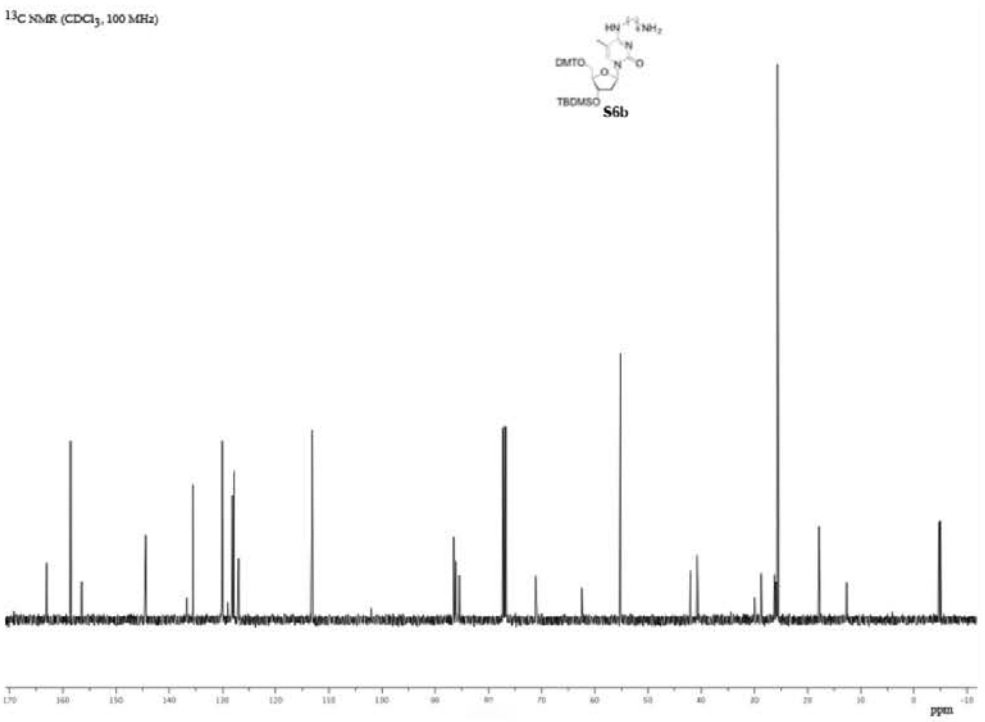



**Figure S24.** $^{13}$C NMR spectrum of compound **S6b**

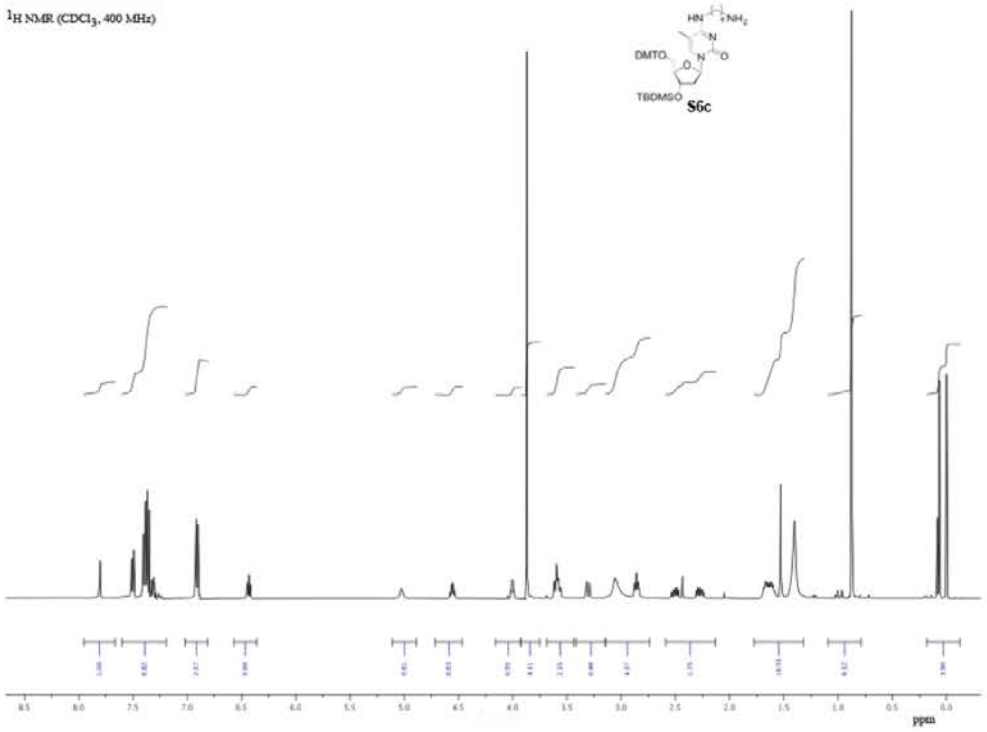



**Figure S25.** $^1$H NMR spectrum of compound **S6c**

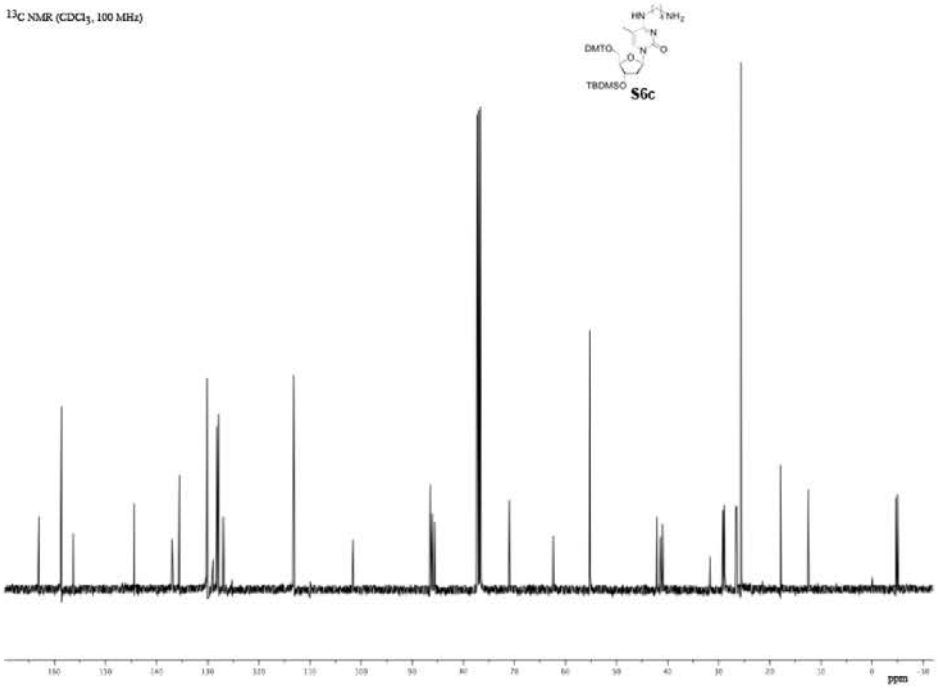



**Figure S26.** $^{13}$C NMR spectrum of compound **S6c**

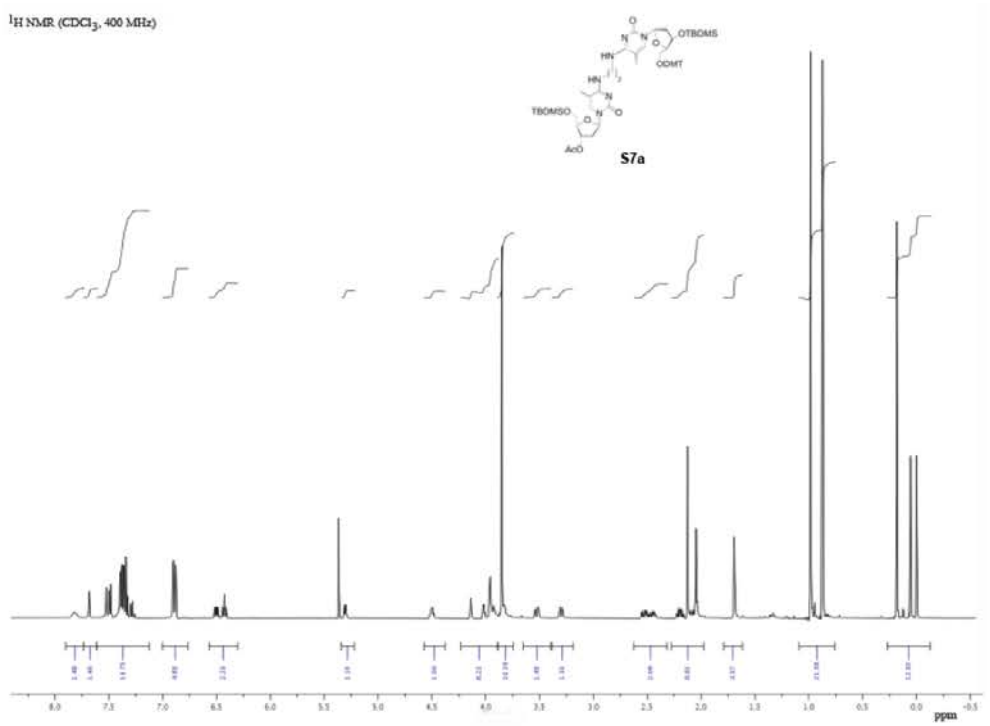



**Figure S27.** $^1$H NMR spectrum of compound **S7a**

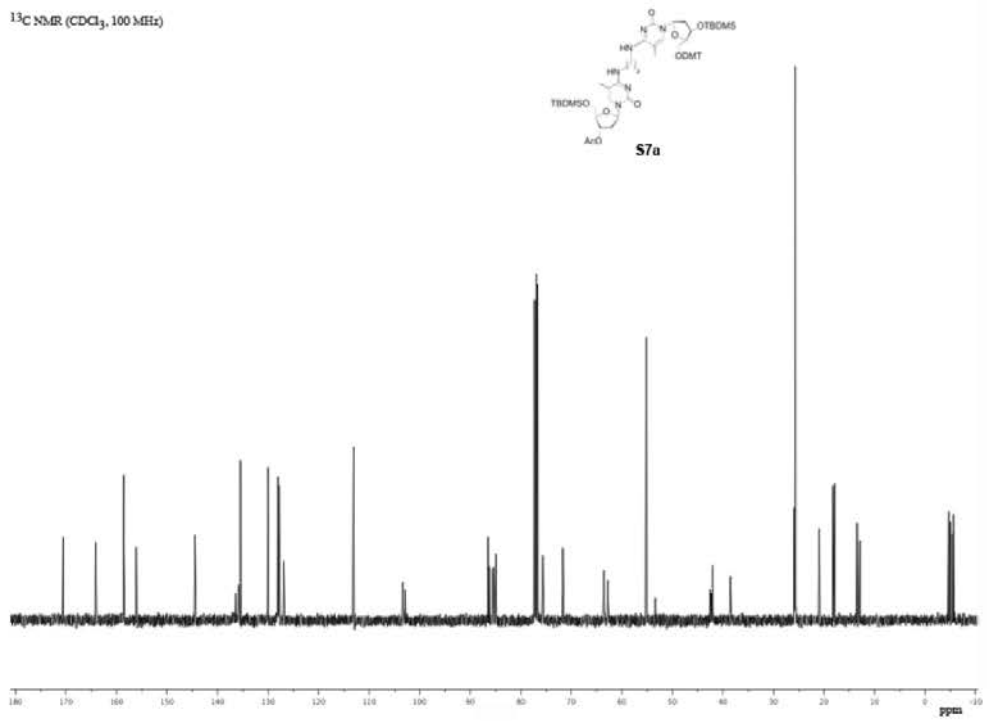



**Figure S28.** $^{13}$C NMR spectrum of compound **S7a**

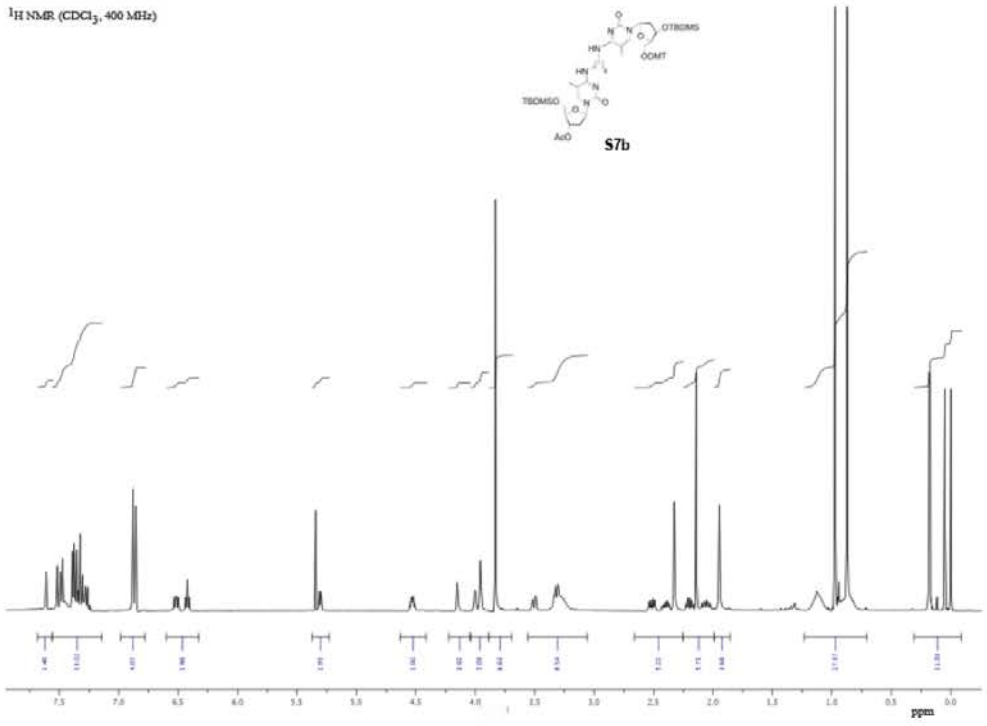



**Figure S29.** $^1$H NMR spectrum of compound **S7b**

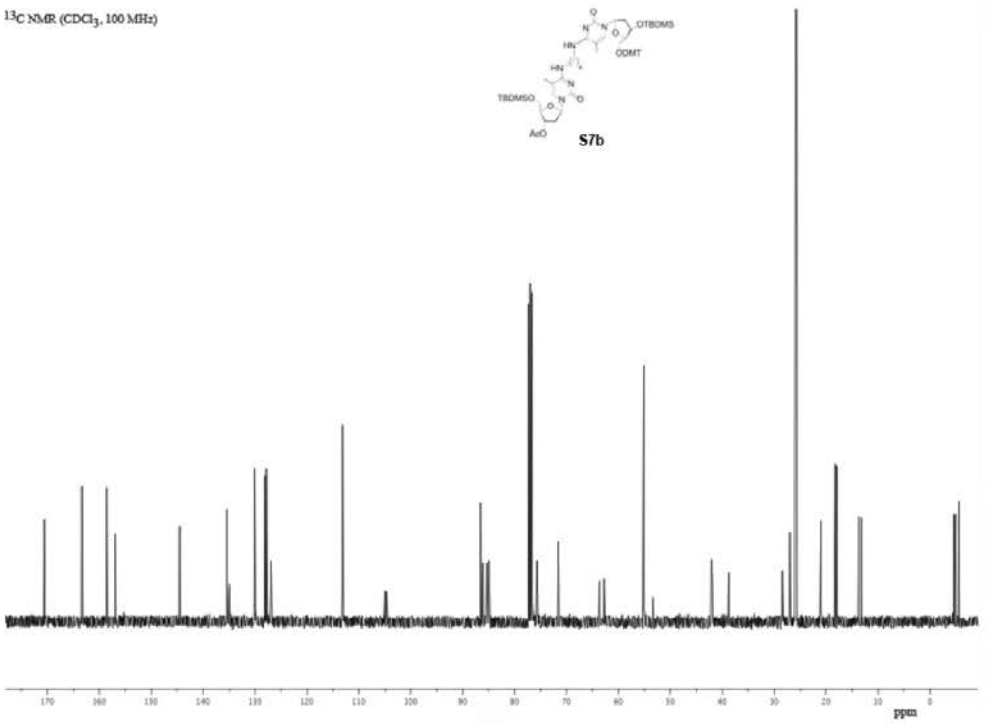



**Figure S30.** $^{13}$C NMR spectrum of compound **S7b**

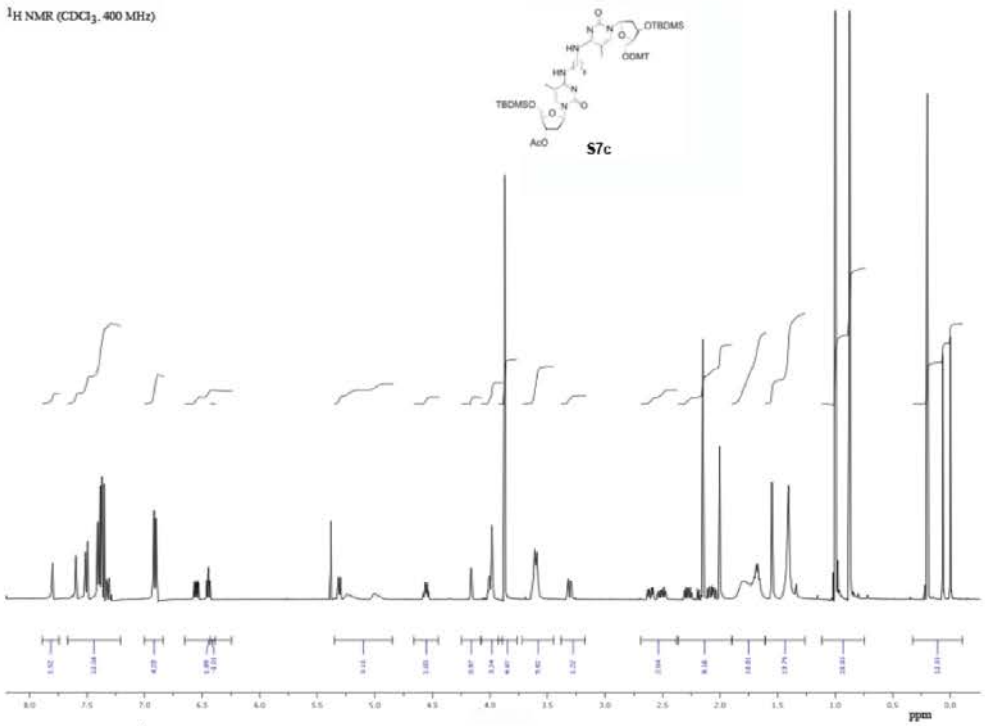



**Figure S31.** $^1$H NMR spectrum of compound **S7c**

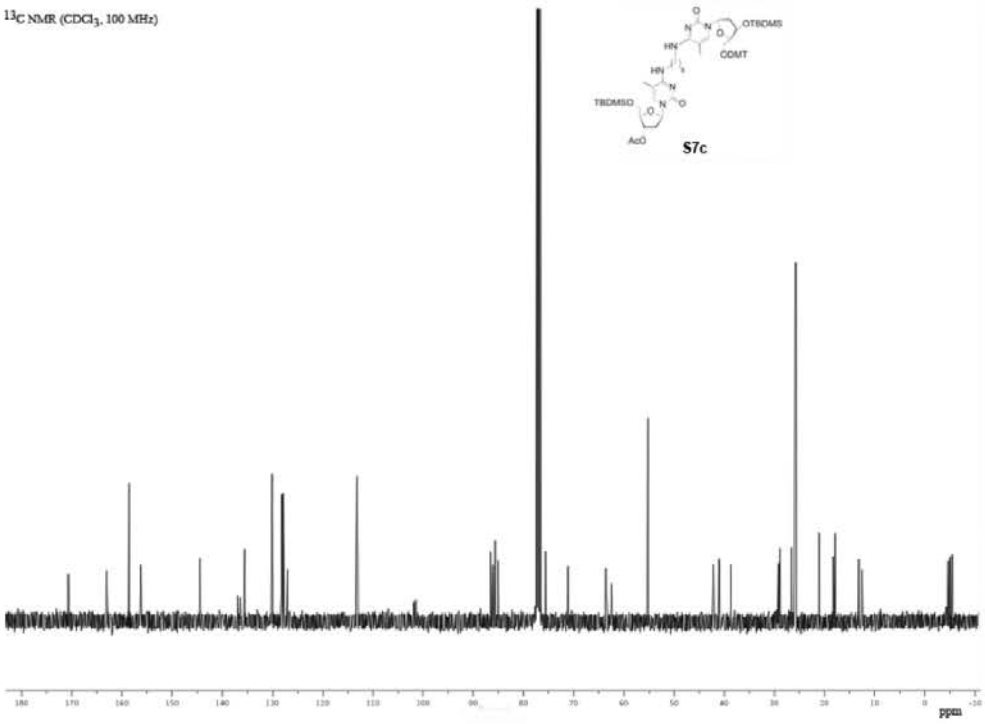



**Figure S32.** $^{13}$C NMR spectrum of compound **S7c**

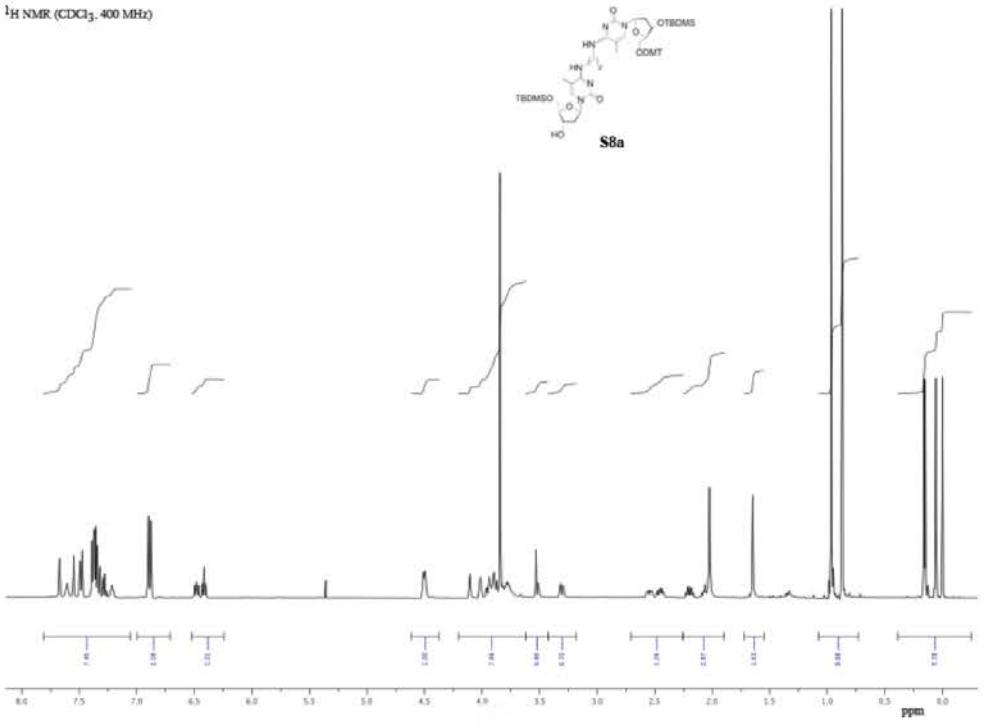



**Figure S33.** $^1$H NMR spectrum of compound **S8a**

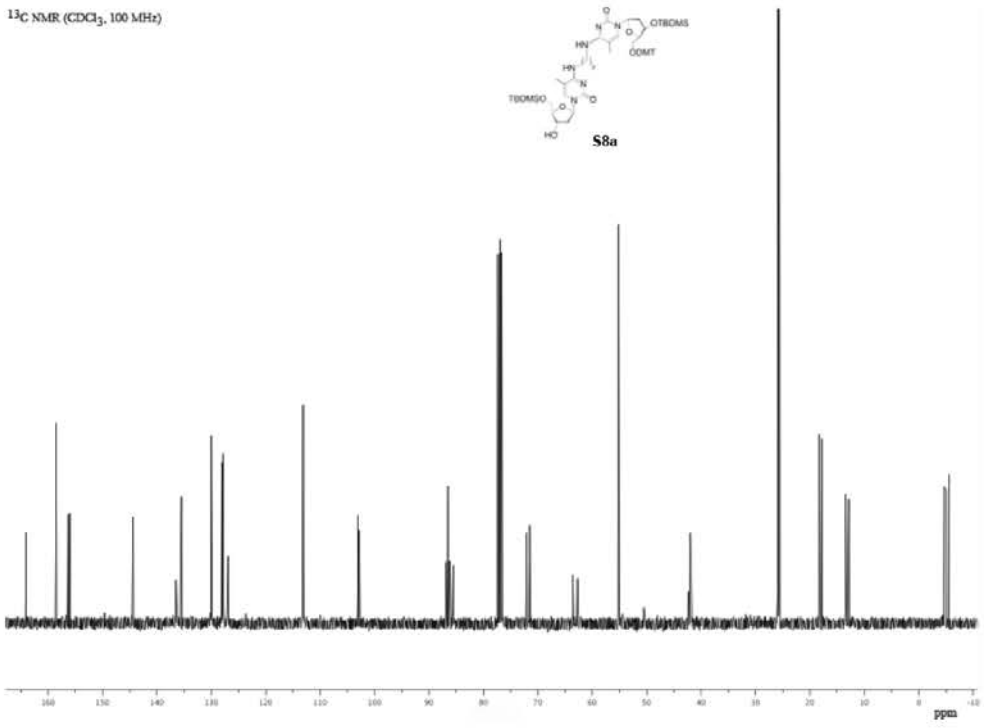



**Figure S34.** $^{13}$C NMR spectrum of compound **S8a**

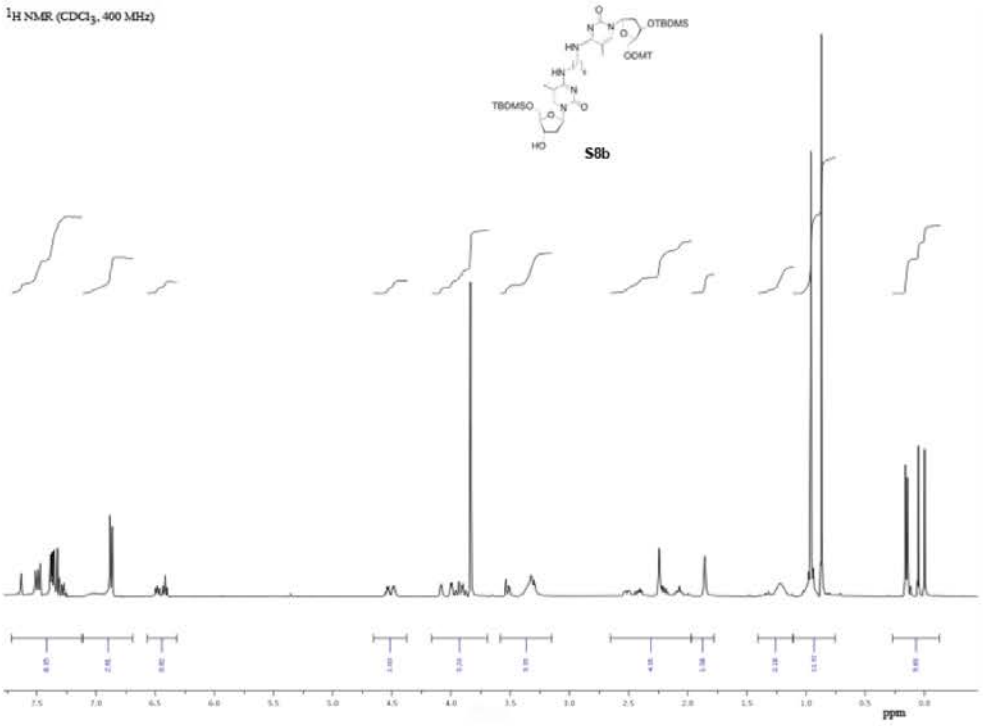



**Figure S35.** $^1$H NMR spectrum of compound **S8b**

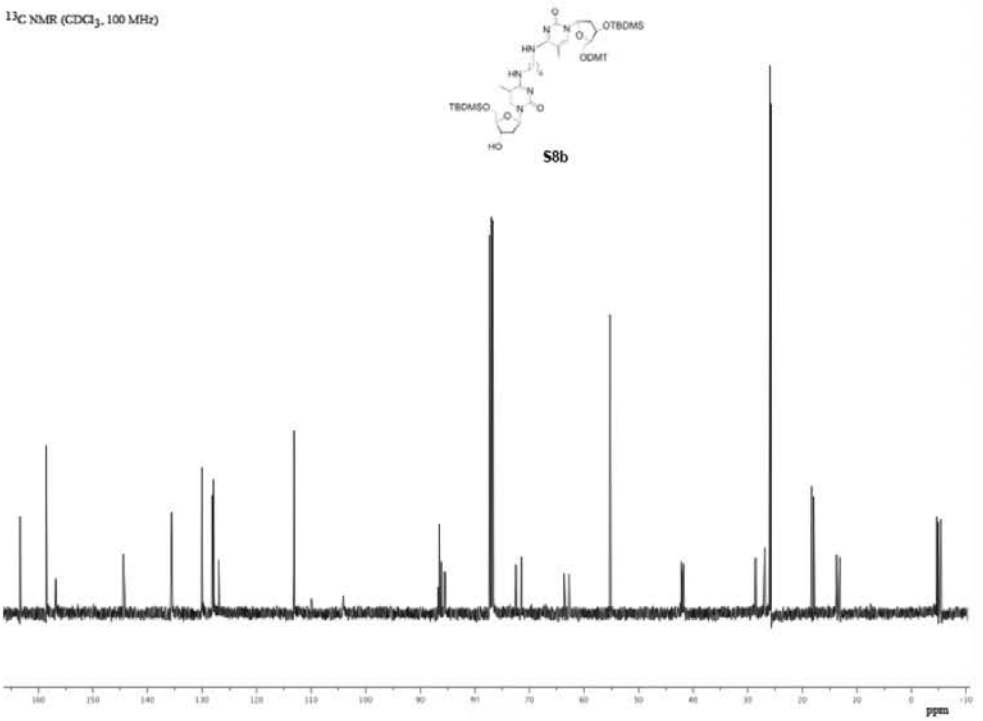



**Figure S36.** $^{13}$C NMR spectrum of compound **S8b**

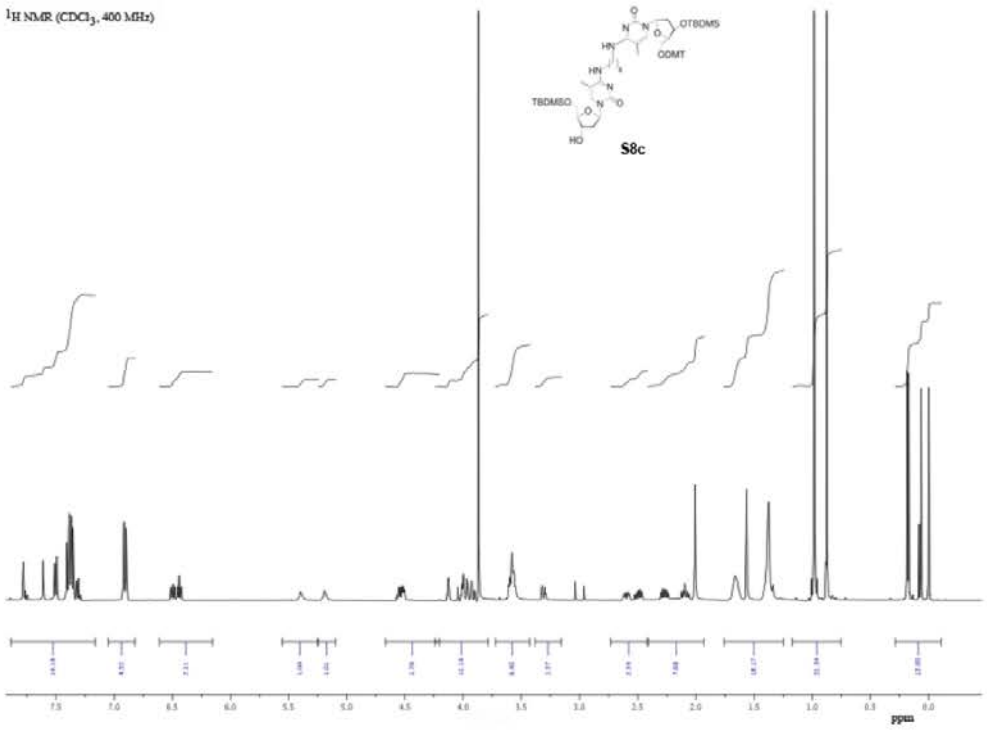



**Figure S37.** $^1$H NMR spectrum of compound **S8c**

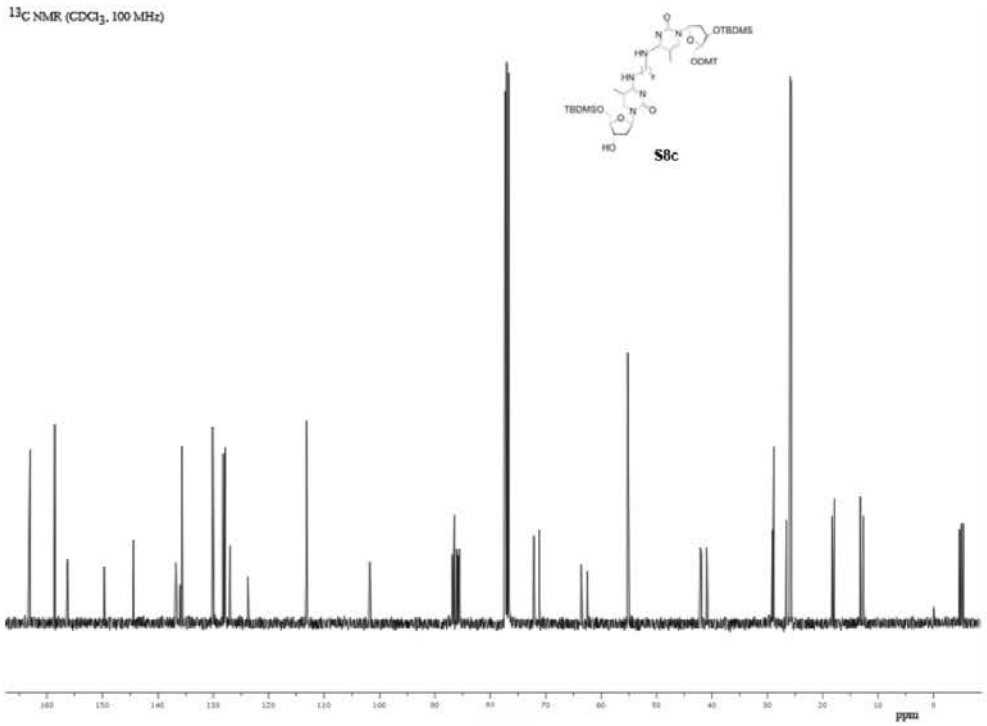



**Figure S38.** $^{13}$C NMR spectrum of compound **S8c**

# Bibliography to Chapter 5

Auffinger, P. & Westhof, E., 1997. Rules governing the orientation of the 2'-hydroxyl group in RNA. *Journal of Molecular Biology*, 274(1), pp.54–63.

Auffinger, P. & Westhof, E., 1997. Rules governing the orientation of the 2′-hydroxyl group in RNA. *Journal of Molecular Biology*, 274(1), pp.54–63.

Bader, R.F.W., 1998. A Bond Path: A Universal Indicator of Bonded Interactions. *The Journal of Physical Chemistry A*, 102(37), pp.7314–7323.

Bader, R.F.W., 1991. A quantum theory of molecular structure and its applications. *Chemical Reviews*, 91(5), pp.893–928.

Bader, R.F.W., 1994. *Atoms in molecules: a quantum theory*, Oxford [England] : New York: Clarendon Press ; Oxford University Press.

Banáš, P. et al., 2010. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *Journal of Chemical Theory and Computation*, 6(12), pp.3836–3849.

Bergonzo, C. et al., 2015. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA (New York, N.Y.)*, p.rna.051102.115-.

Caetano-Anollés, G. & Caetano-Anollés, D., 2015. Computing the origin and evolution of the ribosome from its structure — Uncovering processes of macromolecular accretion benefiting synthetic biology. *Computational and Structural Biotechnology Journal*, 13, pp.427–447.

Cheatham III, T.E., Cieplak, P. & Kollman, P.A., 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 16(4), pp.845–862.

Chen, A.A. & García, A.E., 2013. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 110(42), pp.16820–16825.

Cornell, W.D. et al., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19), pp.5179–5197.

Cubero, E. et al., 1999. Hydrogen Bond versus Anti-Hydrogen Bond: A Comparative Analysis Based on the Electron Density Topology. *The Journal of Physical Chemistry A*, 103(32), pp.6394–6401.

D, V. & D, A., 2014. AIM-UC: An application for QTAIM analysis. *Journal of Computational Methods in Sciences and Engineering*, (1–3), pp.131–136.

Dang, L.X., 1995. Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *Journal of the American Chemical Society*, 117(26), pp.6954–6960.

Dang, L.X. & Kollman, P.A., 1995. Free Energy of Association of the K+:18-Crown-6 Complex in Water: A New Molecular Dynamics Study. *The Journal of Physical Chemistry*, 99(1), pp.55–58.

Denning, E.J. et al., 2011. Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *Journal of computational chemistry*, 32(9), pp.1929–1943.

Denning, E.J. & MacKerell Jr, A.D., 2012. Intrinsic contribution of the 2′-hydroxyl to RNA conformational heterogeneity. *Journal of the American Chemical Society*, 134(5), pp.2800–2806.

Egli, M., Portmann, S. & Usman, N., 1996. RNA Hydration: A Detailed Look [†,‡].

*Biochemistry*, 35(26), pp.8489–8494.
Elbashir, S.M., Lendeckel, W. & Tuschl, T., 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes and Development*, 15(2), pp.188–200.
Fire, A. et al., 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669), pp.806–811.
Fohrer, J., Hennig, M. & Carlomagno, T., 2006. Influence of the 2′-hydroxyl group conformation on the stability of A-form helices in RNA. *Journal of molecular biology*, 356(2), pp.280–287.
Gil-Ley, A., Bottaro, S. & Bussi, G., 2016. Empirical corrections to the Amber RNA force field with Target Metadynamics. *Journal of Chemical Theory and Computation*, 12(6), pp.2790–98.
Ivani, I. et al., 2015. Parmbsc1: a refined force field for DNA simulations. *Nature methods*, 13(1), pp.55–58.
Leontis, N.B. & Westhof, E. eds., 2012. *RNA 3D structure analysis and prediction*, Heidelberg ; New York: Springer.
Marenich, A. V, Cramer, C.J. & Truhlar, D.G., 2009. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B*, 113(18), pp.6378–6396.
Martin-Pintado, N. et al., 2013. Backbone FC H···O Hydrogen Bonds in 2′F-Substituted Nucleic Acids. *Angewandte Chemie International Edition*, 52(46), pp.12065–12068.
Martín-Pintado, N. et al., 2013. Dramatic effect of furanose C2′ substitution on structure and stability: Directing the folding of the human telomeric quadruplex with a single fluorine atom. *Journal of the American Chemical Society*, 135(14), pp.5344–5347.
Mládek, A. et al., 2014. Energies and 2′-hydroxyl group orientations of RNA backbone conformations. Benchmark CCSD(T)/CBS database, electronic analysis, and assessment of DFT methods and MD simulations. *Journal of Chemical Theory and Computation*, 10(1), pp.463–480.
Pérez, A. et al., 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11), pp.3817–3829.
Petrov, A.S. et al., 2014. Evolution of the ribosome at atomic resolution. *Proceedings of the National Academy of Sciences*, 111(28), pp.10251–10256.
Petrov, A.S. et al., 2015. History of the ribosome and the origin of translation. *Proceedings of the National Academy of Sciences*, 112(50), pp.15396–15401.
Pianna, S., 2016. Improvements of RNA force-field. *Workshop of RNA: Structure, dynamics and function*, Trieste.
Saint-Leger, A. et al., 2016. Saturation of recognition elements blocks evolution of new tRNA identities. *Science Advances*, 2(4), pp.e1501860–e1501860.
Smith, D.E. & Dang, L.X., 1994. Computer simulations of NaCl association in polarizable water. *The Journal of Chemical Physics*, 100(5), pp.3757–3766.
Soliva, R. et al., 1999. Role of sugar re-puckering in the transition of A and B forms of DNA in solution. A molecular dynamics study. *Journal of Biomolecular Structure and Dynamics*, 17(1), pp.89–99.
Terrazas, M. et al., 2013. Functionalization of the 3’-ends of DNA and RNA strands with N-ethyl-N-coupled nucleosides: a promising approach to avoid 3’-exonuclease-catalyzed hydrolysis of therapeutic oligonucleotides. *Chembiochem : a European journal of chemical biology*, 14(4), pp.510–20.

Tubbs, J.D. et al., 2013. The Nuclear Magnetic Resonance of CCCC RNA Reveals a Right-Handed Helix, and Revised Parameters for AMBER Force Field Torsions Improve Structural Predictions from Molecular Dynamics. *Biochemistry*, 52(6), pp.996–1010.

Yildirim, I. et al., 2011. Benchmarking AMBER force fields for RNA: Comparisons to NMR spectra for single-stranded r(GACC) are improved by revised X torsions. *Journal of Physical Chemistry B*, 115(29), pp.9261–70.

Zgarbova, M. et al., 2011. Refinement of the Cornell et al. nucleic acid force field based on reference quantum chemical calculations of torsion profiles of the glycosidic torsion. *J. Chem. Theory Comput.*, 7, pp.2886–2.

Zgarbová, M. et al., 2011. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *Journal of Chemical Theory and Computation*, 7(9), pp.2886–2902.

Zhang, J. & Ferré-D'Amaré, A., 2016. The tRNA Elbow in Structure, Recognition and Evolution. *Life*, 6(1), p.3.

*"There is no real ending. It's just the place*
*where you stop the story."*
*Frank Herbert*

# 6| SUMMARY AND GENERAL DISCUSSION

## 6.1. SUMMARY OF THE RESULTS

**1) Parmbsc1**

During the development of the new force field for DNA simulations we:

- Obtained high-level QM profiles of coupled ε/ζ and χ torsions, as well as the sugar pucker.
- Fitted the high-level QM profiles to the then gold-standard force field, parmbsc0, naming the new force field parmbsc1.

This newly obtained parameters, we:

- Tested and benchmarked on more than 100 system of a big variety.
- Compared with direct experimental observables obtained in solution like NOEs, RDCs, or WAXS spectra.

Looking at the performance of parmbsc1 we observed:

- Significant improvements or equal results in all tested systems when compared with parmbsc0 results.
- Clear improvements in averages and profiles of helical parameters and $B_{II}$ populations.
- Better description of terminal residues with a decrease in terminal fraying.
- Excellent agreement with experimental observables.

**2) Drew-Dickerson dodecamer study in a variety of solution models**

When comparing the simulations of DDD using various solvent and salt models, we found that:

- Profiles of helical parameters showed no significant differences in regard of the ion and solvent model used.
- High similarity in DNA flexibility between the simulations with different ion and solvent models used.
- All of this suggests that parmbsc1 is robust with respect to the selection of ion and solvent force field.

In the extended 10 μs trajectory of DDD, we observed:

- Preservation of the good performance parmbsc1 showed on 1 μs timescale, with good reproduction of helical parameters' profiles, no terminal defects that were previously observed with parmbsc0, and clear bimodality of some base pair steps (especially CG step), all in good agreement with experimental observations.
- No significant differences between 2 μs segments, suggesting a full convergence, coming from principle component analysis and entropy analysis.

**3) DNA force field benchmark**

In the "de novo" validation of parmbsc1, we:

- Compared the results of the simulations with experimental results, namely *de novo* obtained NMR data.
- Found that amongst all tested force fields parmbsc1 provides the best fit to various experimental data.
- Report that parmbsc1 (together with OL15) best reproduces averages and profiles of helical parameters.
- Observed that parmbsc1 highlights the noise expected in NMR-derived structures when focused at the base-pair resolution level.
- Strongly suggest using parmbsc1 for simulations of DNA duplexes.

**4) $C_{2'}$-OH study**

When extending our understanding of 2'-OH orientation and effects on the RNA conformational space, we found:

- Three preferred orientations of 2'-OH, whose preference is coupled with sugar puckering.
- That QM study suggests the lowest energy orientation being orientation towards $O_{4'}$ in South, which indicates an ability of 2'-OH to induce changes in sugar puckering and subsequently global changes of the RNA structure.
- A variety of different techniques agree in the 2'-OH acting as the trigger for a general novel induced fit mechanism of protein-RNA recognition.

**5) RNA dumbbells**

In the design of linkers for RNA dumbbell structures, we conclude that:

- Newly designed BC6-linker dumbbell shows higher biostability than other derivatives found in the literature.
- The new dumbbell could be used in gene regulation taking advantage of its improved ADME properties.

## 6.2. GENERAL DISCUSSION

- **New gold-standard force field**

Difficulty in developing a new force field consists of trying to improve known problems while keeping the good aspects of the latest 'state of the art' force field. Parmbsc1 improved generally accepted glitches of parmbsc0 force field, like under-twisting of B-DNAs, lower population of $B_{II}$ states, improper representation of non-canonical structures, and excessive terminal fraying, while managing not to alter generally good results of parmbsc0, like sequence dependent profiles of helical parameters, and reproduction of (among others) dielectric response of DNA and cooperative binding. The valuable ability of parmbsc1 to reproduce well experimental observables did not come with a cost, meaning that parmbsc1 did not 'freeze' the DNA structures to its experimental average, which is confirmed by entropy and normal mode analysis. Moreover, studying conformational transitions we were able to reproduce some large transitions, like unfolding of DNA in pyridine solution or effective folding of a small hairpin. Additionally, parmbsc1 was able to recognize instability of structures like antiparallel d(G-G•C) triplex or anti-parallel Hoogsteen DNA, Z-DNA in physiological conditions. Together with its ability to correctly represent protein-DNA complexes and other non-canonical structures, parmbsc1 showed unprecedented results in the world of force field development. Benchmark of relevant DNA force fields showed that parmbsc1 best fit experimental data (coming from different sources), especially *de novo* obtained NMR data, among all the tested force fields. We believe that parmbsc1 will serve as the new gold-standard force field for simulation of various DNA systems, and that it is possible close to the limit of accuracy for a pairwise additive force field.

- **More details on Drew-Dickerson dodecamer**

The choice of solvent and ion force field models ($Na^+Cl^-$ or $K^+Cl^-$) had little impact on the global structure or flexibility of DDD, proving the robustness of parmbsc1 with respect to the selection of ion and solvent force field. No significant changes in the motions of 10 μs timescale and complete convergence of 2 μs segments, suggests that there is not much left to gain from longer simulations at the present simulation timescales, considering there are no significant motions in sub-ms timescale for DNA dynamics (Galindo-Murillo et al. 2014). Accurate agreement of

helical parameter profiles and averages imply that further small improvements are likely to require the inclusion of polarization. Overall, parmbsc1 provides an improved representation of DDD dynamics, including the subtle choreography of changes happening related to bimodality of certain steps.

- **Efforts on RNA force field**

Despite the big effort of some groups, RNA force fields still lack a great amount of accuracy. Most of the approaches include refinement of χ torsion, but without any significant improvements for many problematic, usually non-canonical, RNA systems. The approach from Chen and Garcia included scaling of base stacking interaction in order to fold RNA hairpin loop, but this modification remained case-specific. Probable cause of this unfavorable outcome of most parameterization efforts is lack of understanding of 2'-OH mechanism of orientation and its influence to overall structure. From our study, we have clearly seen that 2'-OH conformational change can influence sugar puckering reorientation and indirectly induce global structural changes in RNA molecules. Proper description of this torsion is of high importance for further RNA force field development and is crucial to understand the mechanism of induce fitting linked to protein-recognition of RNA structures.

- **RNA dumbbells**

Degradation resistant siRNAs are biologically relevant molecules because of their role in gene regulation. Previous results showed that a structure, called dumbbell, consisting of replacing the natural dinucleotide overhangs with dimeric N-ethyl-N bridges, called BC*n* dimer, showed a resistance to nuclease digestion. Following the previous suggestion we designed dumbbell structures including longer BC-linkers, BC6- and BC8-linkers. Out of the two, BC6-linker proved to be more practical, since dumbbell with BC8-linker was hard to synthesized, mainly due to its high flexibility. Despite all the existing shortcomings, simulation techniques allowed to predict RNA properties and design a new dumbbell structure, with good RISC- activity and large biostability.

- **The known caveats of our force field.**

Even being close to convergence, we believe that classical non-polarizable DNA force fields (including ours) might experience still some refinement. For example, we have been demonstrated the ability of parmbsc1 to fold some basic DNA motives and distinguish between stable and unstable conformations of DNA, but it is not clear if they will be able to reproduce in a systematic way, for example it is not clear that we are going to be able to reproduce absolute melting temperatures. Furthermore, it is not clear whether or not currently used non-bonded terms based on simple combination rules can finely reproduce DNA-protein interactions. In fact, some authors have pointed intrinsic shortcoming in these terms that might need specific correction. For example, Chen and Garcia (Chen & García 2013) and later Elcock's group (Brown et al. 2015) have suggested that nucleobase-nucleobase stacking is overestimated in

water by around 1 kcal/mol, suggesting a linear scaling of the van der Waals term. Unfortunately, as we see from our benchmark study, the resulting force field is not well balanced and produces suboptimal results, highlighting the complexity of correcting force field artifacts. The issue of the potential inaccuracy of the stacking, and its impact on the structure of DNA is under debate, as existing experimental evidences are quite contradictory to each other (Bommarito et al. 2000; Guckian et al. 2000). Similar criticisms have been raised by Case and coworkers (Steinbrecher et al. 2012), and by Cheatham's group (Galindo-Murillo et al. 2014) on the accuracy of the phosphate non-bonded term, which might affect the quality of DNA-protein interactions. It is likely that, as the range of applicability of MD simulations extends, and simulations reach multi-microsecond scales in a systematic way, more errors are going to emerge, and surely new versions of parmbsc1 will tackle them.

As we discussed it before, the van der Waals term includes a mixture of interaction terms, some of them escaping from the pair-additive paradigm, like polarization. The use of rotationally-averaged charges neglects intramolecular polarization effects coupled with structural movements (Basma et al. 2001). For that reason, several groups have developed polarizable force fields, most notable based on Drude particle formalism. For example, MacKerell's group worked extensively on developing polarizable force fields for CHARMM, based on Drude oscillator model, but as we saw from our studies, they are far from perfection. Lastly, we should mention that the introduction of polarization into pair-additive force fields would have a small drawback, as the simulation speed would dramatically decrease (in our experience, at least 5x decrease in speed). Nevertheless, we strongly support the introduction of polarization effects in new generation force fields.

Finally, we believe that it is very important not to overestimate our expectations on classical pair-additive force fields, as there are many intrinsic problems in the basic formalism, which will always restrict their accuracy. For example, the use of atom-centered point charges is a cheap strategy, which shows its shortcomings in reproducing accurate electrostatic interactions when compared with QM potentials (Alemán et al. 1994). Whether spherical van der Waals potentials are able to reproduce in all cases anisotropic dispersion interactions is unclear. Furthermore, we cannot ignore the existence of some interactions involving charge-transfer as is the case of some ion-DNA complexes (Dans et al. 2014; Savelyev & Alexander D MacKerell 2015a; Savelyev & Alexander D MacKerell 2015b; Savelyev & Alexander D. MacKerell 2015). As we see from our Drew-Dickerson study, refined two-body descriptions might be accurate for $Na^+$ and $K^+$, which are the most prevalent cations in physiological conditions, but smaller ions can be much more difficult to represent due to the neglect of polarization (Savelyev & Alexander D MacKerell 2015a; Savelyev & Alexander D MacKerell 2015b; Savelyev & Alexander D. MacKerell 2015), and bivalent ions, especially $Mg^{2+}$, which transfer large part of their charge to the DNA are nearly impossible to reproduce, even for polarizable force fields.

# BIBLIOGRAPHY TO CHAPTER 6

Alemán, C., Orozco, M. & Luque, F.J., 1994. Multicentric charges for the accurate representation of electrostatic interactions in force field calculations for small molecules. *Chemical Physics*, 189(3), pp.573–84.

Basma, M. et al., 2001. Solvated ensemble averaging in the calculation of partial atomic charges. *Journal of computational chemistry*, 22(11), pp.1125–37.

Bommarito, S., Peyret, N. & SantaLucia, J., 2000. Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic acids research*, 28(9), pp.1929–34.

Brown, R.F., Andrews, C.T. & Elcock, A.H., 2015. Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment. *Journal of Chemical Theory and Computation*, 11(5), pp.2315–28.

Chen, A.A. & García, A.E., 2013. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 110(42), pp.16820–16825.

Dans, P.D. et al., 2014. Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic acids research*, 42(18), pp.11304–20.

Galindo-Murillo, R., Roe, D.R. & Cheatham III, T.E., 2014. On the absence of intrahelical DNA dynamics on the μs to ms timescale. *Nature communications*, 5.

Guckian, K.M. et al., 2000. Factors Contributing to Aromatic Stacking in Water: Evaluation in the Context of DNA. *Journal of the American Chemical Society*, 122(10), pp.2213–22.

Savelyev, A. & MacKerell, A.D., 2015a. Competition among Li(+), Na(+), K(+), and Rb(+) monovalent ions for DNA in molecular dynamics simulations using the additive CHARMM36 and Drude polarizable force fields. *The journal of physical chemistry. B*, 119(12), pp.4428–40.

Savelyev, A. & MacKerell, A.D., 2015. 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. *Journal of Chemical Theory and Computation*, 11(9), pp.4473–85.

Savelyev, A. & MacKerell, A.D., 2015b. Differential Impact of the Monovalent Ions Li(+), Na(+), K(+), and Rb(+) on DNA Conformational Properties. *The journal of physical chemistry letters*, 6(1), pp.212–16.

Steinbrecher, T., Latzer, J. & Case, D.A., 2012. Revised AMBER parameters for bioorganic phosphates. *Journal of Chemical Theory and Computation*, 8(11), pp.4405–12.

# CONCLUSIONS

1. The new force field parameters, called parmbsc1, were able to significantly improve MD simulations of big variety of DNA molecules ranging from canonical B-DNA duplexes to unusual DNA structures, correcting known errors of previous force fields.

2. Parmbsc1 results show high accuracy matching with experimental values, mainly direct experimental observables like NOEs, RDCs, WAXS spectra and behaves well in reproducing long transitions, and a variety of unusual DNA structure.

3. Parmbsc1 is robust with respect to the selection of ion and solvent force field and is able to reproduce long-time scale conformational movements, the minor states of DNA, and the mechanical properties of DNA.

4. Parmbsc1 has predictive power and is not contaminated by overtraining artifacts as show by *de novo* predictions on previously unknown structures of DNA.

5. Development of RNA force field is complicated by the largest accessible conformational space and by the extra complexity introduced by the 2'-OH group.

6. The conformational states of the 2'-OH provides the RNA with a unique mechanism to control structure and to react to the presence of interacting proteins.

7. Despite all the existing shortcomings, simulation techniques allow to predict RNA properties and design a new dumbbell structure, with good RISC-activity and large biostability that promise large impact in siRNA regulation of gene activity.