ABSTRACT

PAN, FENG. Structure, Stability, and Dynamics of Select DNA and RNA Double Helices: Trinucleotide Repeats and Ion Distributions. (Under the direction of Dr. Celeste Sagui and Dr. Christopher Roland.)

Trinucleotide repeats (TRs) belong to the family of simple sequence repeats (SSRs), that comprises all sequences with core motifs of 1 to 6 (and even 12) nucleotides that are repeated up to 30 times (and more for pathological cases). SSRs exhibit "dynamic mutations" that do not follow Mendelian inheritance (which asserts that mutations in a single gene are stably transmitted between generations). To date, approximately 30 DNA expandable SSR diseases have been identified and the list is expected to grow. In particular, the dynamic mutations in human genes associated with TRs cause severe neurodegenerative and neuromuscular disorders, known as Trinucleotide (or Triplet) Repeat Expansion Diseases (TREDs). This thesis focuses on the computational study of DNA/RNA duplexes which are directly related with TREDs. Molecular dynamic (MD) simulations can show the atomic structure in details as well as the dynamic properties of those duplexes in microsecond time scale. In addition, the investigation of the Free Energy (FE) landscape can find the global minimum of different structures.

First, we studied the ion atmosphere around nucleic acids. We used large-scale MD simulations to perform a comparative study of the ion distribution around (5'-CGCGCGCGCGCG-3')$_2$ dodecamers in solution in B-DNA, A-RNA, Z-DNA and Z-RNA forms. Our results quantitatively describe the characteristics of the ionic distributions for different structures at varying ionic strengths as well as several binding pockets with rather long ion residence times. Besides, We explored the use of a fast laser melting simulation approach combined with atomistic molecular dynamics simulations in order to determine the melting and healing responses of B-DNA and Z-DNA (5'-CGCGCGCGCGCG-3')$_2$ dodecamers. The frequency of the laser pulse is specifically tuned to disrupt Watson-Crick hydrogen bonds, thus inducing melting of the DNA duplexes. Subsequently, the structures relax and partially refold, depending on the field strength. In addition to the inherent interest of the nonequilibrium melting process, we propose that fast melting by an infrared laser pulse could be used as a technique for a fast comparison of relative stabilities of same-sequence oligonucleotides with different secondary structures with full atomistic detail of the structures and solvent.

Then we targeted on the DNR/RNA duplexed with TRs, which include CAG/GAC, CCG/GCC and CGG/GGC repeats. CAG TRs are known to cause ten late-onset progressive neurodegenerative disorders as the repeats expand beyond a threshold, while GAC repeats are associated with skeletal dysplasias and expand from the normal 5 to a maximum of 7 repeats. Expansions of both GGC and GCC sequences also lead to a number of expandable, TR neurodegenerative diseases. We have carried out free energy and molecular dynamics studies to determine the preferred conformations of the A-A non-canonical pairs in $(CAG)_n$ and $(GAC)_n$ TRs (n=1, 4), C-C or G-G non-canonical pairs in $(GCC)_n$ and $(GCC)_n$, and the consequent changes in the overall structure of the RNA and DNA duplexes. We find that the global free energy minimum corresponds to A-A pairs stacked inside the core of the helix with anti-anti conformations in RNA and (high-anti)-(high-anti) conformations in DNA. For $(GGC)_n$ and $(GCC)_n$, each TR has two reading frames, which results in eight non-equivalent DNA/RNA homoduplexes, characterized by CpG or GpC steps between the Watson-Crick basepairs. Free energy maps for the eight homoduplexes indicate the C-mismatches prefer anti-anti conformations while G-mismatches prefer anti-syn conformations. Besides, the cytosine mismatches in C-rich homoduplexes are weakly bonded for TRs as well as hexanucleotide repeats (HRs) Those cytosines may flip out of the helix core to form e-motif structure. We have performed molecular dynamics simulations of C-rich TR and HR DNA homoduplexes in order to characterize the conformations, stability and dynamics of formation of the e-motif, where the mismatched cytosines symmetrically flip out in the minor groove, pointing their base moieties towards the 5'-direction in each strand.

Structure, Stability, and Dynamics of Select DNA and RNA Double Helices: Trinucleotide Repeats and Ion Distributions

by
Feng Pan

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Physics

Raleigh, North Carolina

2017

APPROVED BY:

Dr. Robert Riehn	Dr. Robert B. Rose
Dr. Celeste Sagui Co-chair of Advisory Committee	Dr. Christopher Roland Co-chair of Advisory Committee

DEDICATION

To my beloved parents,

Qing Pan and Huiqin Cheng,

for all their love and support and putting me through the best education possible.

I appreciate your sacrifices and I wouldn't have been able to get to this stage without you.

BIOGRAPHY

Born in a small city of Anhui located east of China, on May 19, 1990, to Qing Pan and Huiqin Cheng, Feng Pan is the only child in the family. As his father is a middle school teacher, Pan was affected by him and had the chance to read a lot of books in different areas. In 2007, Pan enrolled in one of the most prestigious universities in China, University of Science and Technology of China (USTC), and got his bachelor degree in physics in 2011.

Pan worked in a computational materials laboratory under the advice of Dr. Zhenyu Li and also finished his undergraduate thesis in an experimental condensed matter lab under the advice of Dr. Zengming Zhang. In 2012, with more interest on computational physics, Pan entered the doctoral program in physics at North Carolina State University (NCSU), where he joined the computational biophysics group under the advice of Dr. Celeste Sagui and Dr. Christopher Roland. To date, his research has been focused on structural and dynamic characterization of various nucleic acids associated with trinucleotide repeats disorders by using all-atom molecular dynamics.

ACKNOWLEDGEMENTS

The completion of this project could not have been possible without the support, patience and generosity of many people.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Celeste Sagui. I appreciate all her contributions of time and ideas to make my Ph.D. experience productive and stimulating. The joy and enthusiasm she has for her research were contagious and motivational for me, especially during tough times in the Ph.D. pursuit. She has provided me an excellent example of a successful physicist and professor. I also want to thank my advisor Prof. Christopher Roland for the continuous support of my Ph.D. study and related research, for his patience, encouragement, and immense knowledge. I could not have imagined having better advisors and mentors for my Ph.D. study.

Besides, I would like to thank the rest of my thesis committee: Dr. Robert Riehn and Prof. Robert Rose, for their insightful comments.

My sincere thanks also go to my current and former colleagues: Dr. Mahmoud Moradi, Dr. Viet Man, and Dr. Yuan Zhang for the stimulating discussion, and for all the fun we have had in the last several years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Nucleic acids including DNA and RNA, are some of the most important macromolecules central to life itself in biophysical research. DNA is the carrier of genes and RNA plays an important role in coding, decoding, regulation and expression of genes. DNA and RNA research has mainly targeted the sequence, structure, dynamics and the interactions of these molecules with others such like DNA-binding protein. The most common structure for nucleic acids is double helix, which is formed by two complementary strands held together by Watson-Crick (WC) pairs [1]. However, nucleic acids can also form other secondary or tertiary structures such as stem-loop structures [2], pseudoknots [3], tetraloop [4], quadruplex [5], etc. Since nucleic acids are polyanionic in nature, water and counterions are crucial for their stability. In a recent study, counterions are known to play important roles in regulate genome packing [6, 7], ribozyme activity [8], RNA folding [9–11], and aid in mediating DNA-protein interactions [12]. Besides of regular DNA structure with WC pairs, atypical secondary structures with non-WC mismatches have been found associated with expansion in trinucleotide repeats (TRs) and hexanucleotide repeast (HRs) sequences that underlie approximately 30 expandable simple sequence repeats (SSRs) diseases [13–15]. It is recognized the stable atypical DNA secondary structure in the expanded repeats is "a common and causative factor for expansion in human disease" [13]. In addition, mutant transcripts also contribute to the pathogenesis of Trinucleotide (or Triplet) Repeat Expansion Diseases (TREDs) through toxic RNA gain-of-function [15–21]. In this mechanism, the RNA TRs sequester proteins that are generally involved in pre-mRNA splicing and regulation. Thus, the understanding of these diseases involves the structural characterization of the atypical DNA and RNA structures.

In this thesis, related studies of DNA and RNA are presented including (1) the ion distribution around left- and right-handed DNA and RNA duplexes; (2) the application of a new fast laser melting on the relative stability of B-DNA and Z-DNA; (3) structure and dynamics of atypical DNA and RNA duplexes of trinucleotide repeats: CAG, CCG and CGG; (4) e-motif formed by extrahelical cytosine bases in atypical structures.

1.1 Ion distributions around left- and right-handed DNA and RNA duplexes

Helices formed by natural amino acids and nucleotides are predominantly right-handed, while left-handed forms, such as PPII helices in proteins and Z-DNA and Z-RNA duplexes, are relatively rare. A left-handed, double-helix DNA with two antiparallel chains joined by Watson-Crick (WC) base pairs was first revealed by a crystal structure of $d(CGCGCG)_2$ in 1979 [22]. The term "Z-DNA" was coined for this structure because the sugar-phosphate backbone displays a characteristic zig-zag pattern. Z-DNA is formed by dinucleotide repeats, and is a characteristic of sequences that alternate purines and pyrimidines, mainly CG or GC. These kinds of base pairs give rise to an anti-syn alternation, which is due to rotation of the guanine residue around its glycosidic bond, resulting in a syn conformation, while the cytosine retains its anti configuration [23, 24]. A high density of base sequences favoring Z-DNA is found near transcription start sites [25], where Z-DNA is stabilized by negative supercoiling of DNA [24, 26]. Z-DNA is induced by a set of binding proteins near promoter regions, which boosts the transcription of downstream genes [27]. Z-DNA is highly immunogenic, and antibodies against it [28–30] are used to find locations prone to Z-DNA conformations. The current view is that Z-DNA formation plays a role in gene expression, regulation and recombination [24, 27, 31–37].

Since the right-handed form is DNA's dominant duplex conformation, research has focused on the microscopic mechanisms behind the B-Z DNA transition and controversial models have ensued [38]. Proposed transition mechanisms include: base-pair opening before base-pair plane and phosphate backbone angle rotation within the core of the helix [22]; successive flipping of base-pair planes, without any disruption of the WC pairs [39]; models with intermediate structure [40–44], such as one with two A-DNA-like intermediates [41]; extrusion of bases, as observed in the crystal structure of a B-Z junction [44], followed by propagation and reformation of the pairs. Recent molecular

dynamics (MD) simulations indicate that the transition is governed by a complex free energy landscape which allows for the coexistence of several competing mechanisms so that the transition is better described in terms of a reaction path ensemble [45].

After the discovery of Z-DNA, it was found that the right-handed A-RNA double helix made of CG repeats may also be transformed into a left-handed double helix or Z-RNA [46–50] under conditions of high ionic strength or high pressure [51]. The first detailed structure of Z-RNA of natural sequence, r$(CGCGCG)_2$, was described in an NMR study at high ionic strength in 2004 [50]. However, in contrast to Z-DNA, considerably less is known about Z-RNA and what role it may play in terms of biological functions. Recent experiments – based on binding to the RNA-editing enzyme ADAR1 – have probed the structure of Z-RNA under physiological ionic strength conditions, and provided some evidence that there may even be more than one type of Z-RNA present, either *in vitro* or *in vivo* [52].

An important structural determinant of nucleic acids is that they are polyanionic in nature, and water and counterions are crucial for their stability. In solution, counterions surround the nucleic acid structures and neutralize the nucleic acid anionic phosphates. In addition, they can establish water-mediated contacts and less frequent direct contacts with the electronegative groups. These counterions affect both the structure and stability of the nucleic acid conformations, and therefore their biological function. Specifically, counterions can help regulate genome packing [6, 7], ribozyme activity [8], RNA folding [9–11], and aid in mediating DNA-protein interactions [12]. In fact, due to the highly charged nature of DNA and RNA, it is unlikely that these could be packaged into their compact cellular forms in the absence of counterions [6]. Ions also perform an important role in the transition between the right-handed forms at low salt concentrations, and the left-handed forms at high salt concentrations. In solution, the cations move close to the nucleic acid molecule, finding their way into the major and minor grooves, and backbone oxygens. Given the dynamical nature of the system, the cations generally become localized for relatively short periods of time before drifting away and being replaced by other cations from the solution. Similarly, the mobile anions are likely to be excluded from the near nucleic acid region due to electrostatic repulsion. However, nucleotide electropositive edges have been shown to exhibit specific anion binding sites that also turn out to be good locations for the binding of the negatively charged aspartic and glutamic amino acids and negatively charged groups of other ligands [53].

Considering the importance of the ion distribution in stabilizing nucleic acid structure, it is not surprising that this issue has received intense scrutiny over the past three decades. Although the ions are mobile, some of them can be localized long enough (especially divalent cations) to show up as bound ions in X-ray diffraction studies [54–65], although different atomic resolution crystal structures of equal sequences may differ on the presence of bound ions [66, 67]. Additional improvements have become possible through a combination of anomalous small-angle X-ray scattering (ASAXS) [68–70] and atomic emission spectroscopy (AES) [71]. With these techniques, it has been possible to count explicitly the number of ions in a given region, and thereby provide information as to their time-averaged distributions. NMR studies have also been used to study nucleic acid structure and surrounding ions, especially when precision is improved by the addition of residual dipolar couplings. Thus, for instance, NMR studies have found bound monovalent ions in either the major or minor groove of DNA [56, 61, 72–77].

However, it is still difficult to obtain true spatial resolution and dynamical information with these experimental techniques. This is where classical MD simulations are extremely useful, as they make it possible to explicitly track the motion of ions around the nucleic acids in order to quantify their locations and effect on structure. In fact, MD simulation of the ion atmosphere around nucleic acids have been around for decades [53, 74, 78–93], especially since the correct treatment of electrostatics [94–98] led to stable, reliable trajectories [95]. With some exceptions, the majority of the work has focused on B-DNA. Only recently have more systematic studies on A-RNA emerged [99–103]. In addition, much effort in the refinement of nucleic acid fields has taken place in the last decade [93,104–106], leading to more accurate results in the relatively long-time simulations that are required for the equilibration of the ion distribution.

In Chapter 2, we report on a large-scale MD study of the ion distribution around individual (5'-CGCGCGCGCGCG-3')$_2$ dodecamers in solution in B-DNA, A-RNA, Z-DNA and Z-RNA forms. The duplexes are immersed in rather large water boxes to account for the fact that ionic distribution functions need to be calculated to approximately 30 Å. The study involves 20 simulations lasting 120 ns each. We chose this sequence because we wanted to carry out a comparative study of ion distribution, and the CG sequence is crucial for the left-handed forms. A fair comparison between the different structures needs the same sequence, as previous simulations have shown that the ion distributions exhibit sequence-dependent features. From a chemical point of view, a comparison between RNA

and DNA structures only involves the presence or absence of the 2'-OH group in the sugar (and avoids the extra T/U change associated with AT sequences). The regularity of the sequence also allows for much more "clear-cut" results and conclusions about distribution around CG pairs. Salt effects in this sequence have been studied in 2000, via 2.5 ns long simulations involving the distribution of the K^+ ion [84]; and more recently in simulations that studied the effect of force fields, water model and salt concentration on the structure of A-RNA [99, 101] (these studies did not report results on the ionic distribution itself). Other than that, most of the recent simulations on the ion atmosphere around right-handed nucleic acid duplexes employ sequences that are relatively rich in A and T or U nucleotides. Although *in vitro* the transition from the right-handed forms to the left-handed forms can be triggered with the addition of salt at high concentrations [50], the Z-forms also exist under physiological ionic strength [52]. Simulations allow us to stabilize the left-handed forms and then to slowly increase the concentration of salt in order to discern how ion binding is linked to the structure of the duplex.

In terms of ions, we investigated the distribution of the monovalent Na^+ and K^+, and divalent Mg^{2+} ions around these structures, with various concentration values of their chloride salts. The cations most frequently found around nucleic acids are K^+ (approximately 0.14 M inside the cell) and Mg^{2+}, while the Na^+ ion is most frequently found in extracellular fluids [90]. Since most studies of monovalent ions around nucleic acids involve the Na^+ ion, it is of interest to study the differences in binding between Na^+ and K^+. In this work, we carry out a comparative study of the distributions of each of these cations around each of the four possible duplexes, and discuss in detail the sources of the various, important observed differences.

1.2 A new fast laser melting method and its application on B-DNA and Z-DNA

Recently, free electron lasers have been used for melting biomolecular complexes [107–111]. Kawasaki and coworkers [109–111] have developed a mid-infrared free electron laser with highly specific oscillation characteristics having a high photon density, a picosecond pulse structure, and a range of tunable frequencies. To date, such a laser pulse with frequencies tuned to the amide I bands has been used in experiments to dissociate amyloid-like fibrils of lyzozyme into its native forms [109], convert insulin fibrils into soluble monomers [110],

and dissociate a short human thyroid hormone peptide [111]. A simulated laser pulse has also been used in atomistic molecular dynamics simulations to dissociate amyloid fibrils [112], and to study the break-up of a peptide-based nanotube [113]. Aside from investigating the dissociation of selected biomolecules, in this work we propose that the application of a simulated laser pulse could provide for new computational opportunities to probe the relative response and, quite possibly, the relative stability of similar structures, such as nucleic acids with the same sequence but with different secondary structures.

A crucial aspect for the understanding of nucleic acid structures and their function is the relative stability of different competing structures that are involved in cellular processes. The discernment of the relative structural stabilities of different nucleic acid structures, both *in vitro* and *in silico*, is rather challenging. A time honored technique to address this issue is to carry out a thermal denaturation experiments [114], where a sample is heated beyond the melting point. This causes a conformational transition in the molecule that is measured by recording a specific variable as a function of temperature. Typical experiments include the recording of the absorbance at 260 nm in a UV-visible spectrophotometer [115]; circular dichroism spectra [116, 117]; fluorescence spectra [118]; Raman signals [119, 120]; or NMR measurements [121, 122].

Thermal denaturing experiments can also be carried out computationally, but fully atomistic melting simulations are extremely expensive. Instead, simplified models have been introduced for estimating the melting temperature of DNA duplexes, based on empirical or statistical thermodynamics models [123–139] that predict the stability of nucleic acid secondary structure, including RNA [127, 140–153]. Unfortunately, these models are still empirical and can fail to capture subtle sequence-dependent effects, or unusual conformations different from A and B duplexes, or conformations with noncanonical or mismatched base pairs, etc. A state-of-the-art web-based tool for predicting fluorescent high-resolution DNA melting curves and denaturation profiles of PCR products is the software package uMELT [154]. This tool, however, can only be applied to standard DNA duplexes. Since secondary structures other than B-DNA can play crucial roles in processes such as gene expression, extending the computational predictive capacity to these unusual structures is a desirable goal.

In Chapter 3, we use fast melting by an infrared laser pulse of B-DNA and Z-DNA duplexes with sequence d(5'-CGCGCGCGCGCG-3')$_2$ in order to compare the melting and healing response of the duplexes and to explore whether this response reflects their

relative stability. The left-handed, double-helix Z-DNA with two antiparallel chains joined by Watson-Crick (WC) base pairs was first crystallized in 1979 [22]. Z-DNA is favored by sequences that alternate purines and pyrimidines, mainly CG or GC. The left-handed helix is thinner and more rigid than B-DNA, and while the pyrimidine-purine base pairs are in *anti-anti* conformation in B-DNA, the guanine rotates around its glycosidic bond in Z-DNA, resulting in an *anti-syn* configuration [23, 24]. As a consequence of these rotations, the phosphate groups become closer in Z-DNA; the sugar-phosphate backbone displays the characteristic zig-zag pattern that gave rise to the name Z-DNA; and the CpG and GpC steps stack differently, causing a dinucleotide step to be the repeating unit in Z-DNA. A high density of base sequences favoring Z-DNA is found near transcription start sites [25], where Z-DNA is stabilized by negative supercoiling of DNA [23, 24, 26]. In the cell, Z-DNA is induced by a set of binding proteins near promoter regions, which boosts the transcription of downstream genes [27]. Z-DNA is highly immunogenic, and antibodies against it [28–30] are used to find locations prone to Z-DNA conformations. The current view is that Z-DNA formation plays a role in gene expression, regulation and recombination [24, 27, 31–37].

B-DNA is more stable than Z-DNA under physiological ionic strength and pH conditions. While in the cell Z-DNA can be stabilized by the negative supercoiling [23, 24, 26] that, for instance, arises from the relocation of the RNA polymerase during transcription, and through binding with highly specific Z-DNA binding proteins such as ADAR1, E3L and PKZ [155–158], *in vitro* inducers and stabilizers are needed for Z-DNA, such as high ionic concentrations, base chemical modifications, organic solvents, divalent cations and transition metal complexes, small molecular complexes, etc. (these are reviewed in Ref. [159]). Computationally, the determination of the relative stability of B-DNA and Z-DNA is not trivial. Existing predictive tools do not have provisions for this alternative structure, and therefore one has to resort to very long time simulations or careful theoretical and statistical modeling. Considerable insight has been obtained by studying the microscopics behind the B-Z DNA transition [22, 38–44, 160–163], and the ion distribution around these nucleic acid structures [164]. Recent molecular dynamics (MD) simulations indicate that the transition is governed by a complex free energy landscape which allows for the coexistence of several competing mechanisms so that the transition is better described in terms of a reaction path ensemble [163].

We choose B-DNA and Z-DNA duplexes with sequence d(5'-CGCGCGCGCGCG-3')$_2$

to test their behavior when exposed to fast melting by an infrared laser pulse. In addition to the intrinsic interest of the melting and healing process, we propose that this technique could be used as a relatively inexpensive tool to determine relative stability in fully solvated, atomically accurate nucleic acid structures. Naturally, a laser pulse implies a nonequilibrium process and the results cannot be used to construct an equilibrium melting curve, and therefore the ensuing free energy estimates cannot be obtained either. The perturbation, however, can map out the responses of the structures for the entire range of applied fields, covering extremely small perturbations all the way to complete melting. The response can be measured by a wide range of variables including normalized handedness, base stacking, Watson-Crick hydrogen bonds, etc. A systematic trend in all these variables (in this case that Z-DNA is more "melted" than B-DNA) for all strengths of the field is interpreted as being indicative of the greater stability of B-DNA under the solvation conditions in the simulations. A traditional equilibrium molecular dynamics simulation not only is orders of magnitude more expensive but also faces additional challenges. Since B-DNA is more prone to fray at the ends than Z-DNA, the melting of short B-DNA oligomers is more susceptible to length effects than Z-DNA, a difficulty that can be partially bypassed with the laser melting technique. In addition, Z-DNA is believed to be stabilized by higher temperatures [40, 165–167], which complicates the interpretation of the traditional melting experiments.

1.3 Structure and Dynamics of DNA and RNA Double Helices of Trinucleotide Repeats: CAG, CCG and CGG

Trinucleotide repeats (TRs) belong to the family of simple sequence repeats (SSRs), that comprises all sequences with core motifs of 1 to 6 (and even 12) nucleotides that are repeated up to 30 times (and more for pathological cases) [168]. SSRs exhibit "dynamic mutations" that do not follow Mendelian inheritance (which asserts that mutations in a single gene are stably transmitted between generations). In the 1990's scientists discovered that inherited neurological disorders known as "anticipation diseases", where the age of the onset of the disease decreased and its severity increased, were caused by the intergenerational expansion of SSRs [169–172]. After a certain threshold in the

length of the repeated sequence, the probability of further expansion and the severity of the disease increase with the length of the repeat. To date, approximately 30 DNA expandable SSR diseases have been identified and the list is expected to grow [14, 15]. In particular, the dynamic mutations in human genes associated with TRs cause severe neurodegenerative and neuromuscular disorders, known as Trinucleotide (or Triplet) Repeat Expansion Diseases (TREDs) [170, 173–175]. The expansion is believed to be primarily caused by some sort of slippage during DNA replication, repair, recombination or transcription [14, 15, 172, 176–180]. Cell toxicity and death have been linked to the atypical conformation and functional changes of the transcripts and, when TRs are present in exons, of the translated proteins [15–21, 181–184]. The expanded RNA transcripts exhibit secondary structures that sequester regulatory proteins and cause abnormal nuclear foci [185–188]. Contributing to the complexity of the pathological mechanisms, there is also evidence that antisense transcripts of the expansion, i.e., expanded repeats resulting from the bidirectional transcription of the DNA TR expansions, can also form nuclear RNA foci that contribute to toxicity, and that both sense and antisense expansions can trigger protein translation in the absence of the start ATG codon, giving rise to the unconventional repeat-associated non-ATG (RAN) translation [189].

CAG/GAC Trinucleotide Repeats

Of all the TRs, CAG repeats give rise to the largest group of neurodegenerative diseases. CAG repeats in the 5'-UTR of the gene PPP2R2B cause spinocerebellar ataxia type 12 (SCA12), while CAG repeats in the exon part of various genes cause other nine late-onset, progressive neurodegenerative disorders, including Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA) and several spinocerebellar ataxia (SCAs). These disorders are also known as polyglutamine (polyQ) diseases [190], because although CAG repeats could likely encode three different amino acid repeats depending on the reading frame (codons CAG, AGC and GCA would code for polyQ, polyS and polyA, respectively), the CAG expansions in these genes only lead to polyQ expansions. These polyQ diseases, like other TREDs, are caused by expansions greater than a given threshold [190]. For instance, in HD, the normal polyQ (or CAG repeat) length is 10-34 repeats, and pathological lengths are 36-250 repeats. Although each disease has a different pathology, they all share a common feature: the formation of polyQ aggregates [191], where the mature fibrils display cross-β

conformations [192–198]; and the eventual neuronal death.

Interestingly, after the discovery of the CAG repeats and their relation to neurological disease, it was found that the GAC trinucleotide is also involved in a completely different class of diseases from the known TREDs. These diseases are caused by a *very small change* in the repeat number, and therefore do not qualify as TREDs. In particular, the human gene for cartilage oligomeric matrix protein exhibits a $(GAC)_5$ repeat. Expansion by one repeat causes multiple epiphyseal dysplasia, while expansion by two repeats or, alternatively, deletion by one repeat causes pseudoachondroplasia [199]. The structure of the various duplexes seem to strongly depend on the pH of the solution and the ionic strength [200]. While the CAG trinucleotide leads to expansion, the GAC trinucleotide does not (except for at most two extra repeats).

Although the mechanisms underlying TREDs are believed to be extremely complex, simple and robust trends beyond the repeat threshold have been identified, such as the correlation between the repeat length and the probability of further expansion and increased severity of the disease. Another important breakthrough has been the recognition that stable atypical DNA secondary structure in the expanded repeats is "a common and causative factor for expansion in human disease" [13]. In addition, mutant transcripts also contribute to the pathogenesis of TREDs through toxic RNA gain-of-function [15–21]. In this mechanism, the RNA TRs sequester proteins that are generally involved in pre-mRNA splicing and regulation. Thus, a first step towards the understanding of these diseases involves the structural characterization of the atypical DNA and RNA structures. Since there is experimental consensus that the most typical DNA and RNA TR secondary structures, at least in the initial stages of expansion, are hairpins whose stem lengths can wildly vary [21, 201–203], a characterization of the mismatched helical duplexes forming the stems provides a foundation towards a structural understanding of the TR atypical secondary structures.

At present, little is known about the atomic structure and associated dynamics of trinucleotide CAG and GAC repeats. To date, experimental investigations have only considered CAG repeats in RNA; there are no experimental studies with atomic resolution of GAC repeats – not for DNA or RNA; and, perhaps most importantly, there are no experimental atomic resolution experiments of CAG repeats in DNA. Given that the expansions that characterize TRs originate at the DNA level, a structural understanding of these repeats at the atomic level in DNA is particularly important. Also, as described

above, GAC repeats and CAG repeats behave in radically different ways in a biological context, and teasing out the structural differences between these two repeats both in the RNA and DNA context may help in the elucidation of their different behavior with respect to expansion diseases.

Here we briefly review the experimental results for CAG repeats in RNA. The X-ray RNA-CAG duplex crystal structures include the following sequences: the sequence r(5'-GG-$(CAG)_2$-CC)$_2$ [204], and the sequence r(5'-UUGGGC-$(CAG)_3$-GUCC)$_2$ [205, 206]. This last sequence was also analyzed via NMR [206]. The first study found that the duplexes favor the A-RNA form and that the A-A non-canonical pairs are in the anti-anti conformation. In the second sequence, both anti-anti and syn-anti A-A conformations were observed: the A-A pairs in the internal CAG always displayed the anti-anti conformation, while one [206] or two [205] of the terminal A-A pairs displayed the anti-syn conformation. These results are in general agreement with the complementary molecular dynamics (MD) simulations [205]. Thus, while the anti-anti conformation (with fluctuations) for an internal A-A pair in a TR is common to the three studies, the nature of the anti-syn conformations is not clearly established. This is because these conformations occur in the terminal A-A pairs of r(5'-UUGGGC-$(CAG)_3$-GUCC)$_2$, where the A-A pairs are flanked by CC/GG steps. Using high level *ab initio* calculations, it has been shown that CC/GG steps are the least stable of the ten dinucleotide steps, with well-separated energies [207] from the other dinucleotide steps. Since these steps are never present in a genuine $(CAG)_n$ TR (which only exhibits GpC steps), it is clear that their presence could bias the conformation of the adjacent A-A pairs.

In addition to these RNA-CAG studies there is one MD study for CAG repeats in DNA [208]. This study uses a sequence that is more relevant to the expanded disease, mainly d$(CAG)_6$. According to the conclusions of this study, the A-A mismatch in DNA behaves in exactly the opposite way than its RNA counterpart: it disfavors the anti-anti and the anti-(+syn) conformations and adopts the (-syn)-(-syn) conformations, resulting in a local Z-form around the mismatch [208]. These results are intriguing and raise questions as to the true nature of the A-A mismatches in DNA-CAG.

In Chapter 4, we present a unified and comparative description of the nucleic acid duplexes for both DNA and RNA for both CAG and GAC trinucleotide repeats based on MD simulations and free energy calculations. A review of the field of MD simulations of nucleic acids is beyond the scope of this work, and the reader is referred to the

authoritative reviews presented in Refs. [93, 209, 210]. Out of the four possible DNA/RNA CAG/GAC cases, there is experimental data only for RNA-CAG. We therefore begin by making the connection with this experimental data through an explicit investigation of a specific sequence employed in these studies and then move on to a four-trinucleotide repeat duplex. After that we consider the other three cases –specifically RNA-GAC, DNA-CAG and DNA-GAC. In particular, we present results corresponding to both free energy calculations and regular 1 μs MD simulations of a single mismatch duplexes (5'-CCG-CAG-CGG-3')$_2$ and (5'-GGC-GAC-GCC-3')$_2$ both for RNA and DNA, and for regular 1 μs MD simulations of four-trinucleotide repeat duplexes (5'-(CAG)$_4$-3')$_2$ and (5'-(GAC)$_4$-3')$_2$. For each of the four duplexes, the free energy calculations involve two maps, each computed with a different pair of collective variables. The eight resulting free energy maps allow us to identify and rank the minima corresponding to the different A-A mismatch conformations. We also identify mechanisms of transition of the A-A mismatches towards the global free energy minimum, and link these mechanisms to paths over the free energy maps. We complete the work with a characterization of the neutralizing Na^+ ion distributions around the mismatches. Strictly speaking, the non-canonical A-A pairs in RNA are not "mismatches" since RNA is not necessarily self-complementary. However, since we are considering both DNA and RNA in their duplex form, we will call these non-canonical base pairs mismatches for simplicity.

CCG,CGG Trinucleotide Repeats and an associated e-motif structure

In Chapter 5 and Chapter 6 we are interested in CGG and CCG TRs, which are overexpressed in the exons of the human genome: CGG TRs are found in the 5'-untranslated region (5'-UTR) of the fragile X mental retardation gene (FMR1) [211], while CCGs are found both in the 5'-UTR and translated regions of more than one gene. The normal range of the CGG TRs in the population is 5-54, with the last ten repeats increasing the probability of disease in descendants [212, 213]. TRs of 55-200 CGGs constitute premutations associated with fragile X-associated tremor ataxia syndrome (FXTAS) in males [214] and premature ovarian failure in females [215]. TRs longer than 200 CGG cause the inherited fragile X mental retardation syndrome [216]. CCG TRs are related to three TREDs: the longest expansion occurs in the FRM2 gene giving rise to chromosome X-linked mental retardation (FRAXE) [217], and they also seem to play a role in Huntington's disease [218],

and myotonic dystrophy type 1 [219].

For the atypical DNA and RNA structures, various experimental methods *in vitro*, such as CD, UV absorbance, NMR, electrophoretic mobility assay, and chemical or enzymatic digestion [220], show a general trend to formation of duplexes and hairpins, depending on the sequence length and environment conditions. Among these secondary structures, those formed by CGG expansions seem to be the most stable.

Crystallographic studies for short RNA duplexes provide valuable atomic detail. For the CGG expansion, two crystallographic studies using unmodified sequences 5'-G-$(CGG)_2$-C-3' (PDB ID 3R1C, Ref. [221]) and 5'-UU-GGGC-$(CGG)_3$-GUCC-3' (PDB ID 3JS2, Ref. [222]) found that the RNA helices have the A-form, with some variations, with the G-G pairs in a typical anti-syn conformation, with two hydrogen bonds between the Watson-Crick edge of G_{anti} and the Hoogsteen edge of G_{syn}. For the CCG sequence, there is one crystallographic RNA duplex with an unmodified sequence 5'-G-$(CCG)_2$-C-3' (PDB ID 4E59, Ref. [223]), and one solution NMR DNA duplex 5'-$(CCG)_2$-3' (PDB ID 1NOQ, Ref. [224, 225]). The C-rich structures are less conclusive because they involve only two repeats, which results in the slipping of one strand with respect to the other. In the RNA crystal structure, this dislocation and the stacking of the oligomers along the c-axis in the crystal results in a single C·C pair effectively surrounded by four C-G Watson-Crick pairs (with two overhanging C's). Thus, it is not clear whether this structural environment for the single "mismatch" can reproduce the one that would occur in the cell for longer $(CCG)_n$ sequences, where each C·C pair may (or may not) be surrounded by only two Watson-Crick pairs. The C·C pair surrounded by four Watson-Crick C-G base pairs as shown in the RNA crystal might be overconstrained with respect to that in a real CCG expansion.

Another important issue when considering possible TR conformations is the nature of the Watson-Crick pairs that surround the mismatches [226, 227]: sequences of the form 5'-$(CGG)_n$-3' and 5'-$(CCG)_n$-3' (without slipping) exhibit GpC steps between the Watson-Crick base pairs, while sequences of the form 5'-$(GGC)_n$-3' and 5'-$(GCC)_n$-3' (without slipping) exhibit CpG steps between the Watson-Crick base pairs. The two RNA G-rich crystal structures [221, 222] involve GpC steps; terminal mismatches in 5'-UU-GGGC-$(CGG)_3$-GUCC-3' in Ref. [222] are surrounded by CC/GG steps, not present in a $(CGG)_n$ expansion. Indeed, with the use of high level *ab initio* calculations, it has been shown that CC/GG steps are the least stable of the ten dinucleotide steps, with

well-separated energies [207] from the other dinucleotide steps. The slipping of strands with respect to each other in the (CCG) sequences results in GpC steps for the RNA crystal [223] and in CpG steps for the DNA NMR structure [224] (as opposed to the GpC steps that would result if the DNA strands were paired at the ends).

The work presented here is part of our effort to achieve a unified and comparative description of the nucleic acid duplexes obtained from SSRs for both DNA and RNA, considering all the possible reading frames that result in CpG or GpC steps between the Watson-Crick base pairs. Our previous work includes a characterization of the four helical duplexes obtained from the CAG (GpC steps) and GAC (CpG steps) TRs for both RNA and DNA [228]; and of the twelve helical duplexes derived from the (GGGGCC) hexanucleotide repeat (HR) expansion in the C9ORF72 gene, and its associated antisense (GGCCCC) expansion [229]. CAG TRs are known to cause ten late-onset progressive neurodegenerative diseases, including spinocerebellar ataxia type 12 (SCA12), Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA) and several other spinocerebellar ataxia (SCA) diseases [190]. On the other hand, GAC repeats behave quite differently: expansion by one repeat in the human gene for cartilage oligomeric matrix protein, which exhibits a $(GAC)_5$ repeat, causes multiple epiphyseal dysplasia, while expansion by two repeats or, alternatively, deletion by one repeat causes pseudoachondroplasia [199]. A (GGGGCC) HR expansion in the first intron of the C9ORF72 gene has been shown to be the major cause behind frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) [230, 231]. While the unaffected population carries fewer than 20 repeats (generally no more than a couple), large expansions greater than 70 repeats and usually encompassing 250-1600 repeats have been found in C9FTD and ALS patients. The twelve duplexes that we studied result from the three different reading frames in sense and antisense HRs for both DNA and RNA. These duplexes display atypical structures relevant not only for a molecular level understanding of these diseases but also for enlarging the repertoire of nucleic-acid structural motifs.

In Chapter 5, we present results for molecular dynamics (MD) simulations and free energy calculations for both CCG and GGC trinucleotide repeats, with either CpG or GpC steps, for both RNA and DNA. This results in eight different non-equivalent helical duplexes. We compare results with the one case, G-rich RNA with GpC steps, which is well characterized experimentally. The good agreement with the experimental structures

helps validate our results for the other seven cases. In addition to the free energy maps, we identify mechanisms of transition of the mismatches towards the global free energy minimum, and link these mechanisms to paths over the free energy maps. We complete the work with a characterization of the neutralizing Na^+ ion distributions around the mismatches. Strictly speaking, the non-canonical C·C and G·G pairs in RNA are not "mismatches" since RNA is not necessarily self-complementary. However, since we are considering both DNA and RNA in their duplex form, we will call these non-canonical base pairs mismatches for simplicity.

In the DNA duplex, the slipping of the strands leaves the two 5'-C terminal unpaired, and a single central C·C mismatch surrounded by two Watson-Crick pairs. This gives rise to the "e-motif", where the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the $i-2$ residue (i.e., towards the 5' direction in each strand). In Chapter 6, we carried out the study on e-motif structures for TRs as well as haxanucleotide repeats. This e-motif was seen in a solution NMR DNA antiparallel duplex where each strand consisted of two repeats, 5'-$(CCG)_2$-3' (PDB ID 1NOQ). By contrast, this e-motif has not been observed in RNA. In the only crystallographic structure available for RNA with unmodified CCG sequences, the antiparallel RNA duplex 5'-G-$(CCG)_2$-C-3' (PDB ID 4E59, Ref. [223]), the two short strands also slip with respect to each other. This dislocation and the stacking of the oligomers along the c-axis in the crystal results in a single C·C pair effectively surrounded by four C-G Watson-Crick pairs (with two overhanging C's). However, the C·C pair remains inside the helix, and no e-motif is observed.

Remarkably, since the initial publication of the NMR DNA 5'-$(CCG)_2$-3'duplex results (1995), there has been no other direct structural observation of the e-motif, reflecting the difficulty of experimental observation of flexible DNA duplexes, made probably more flexible by the presence of the mismatches. However, there has been indirect observations that support the presence of e-motifs in DNA homoduplexes and hairpins of various lengths. The two most important results in this direction were obtained by chemical modification of the bases followed by subsequent cleavage. These studies also provided indirect evidence to the proposition that $d(GCC)_n$ homoduplexes or hairpin stems exhibit an *extended* e-motif formed by consecutive extrahelical C·C mismatches. Finally, notice that the GCC alignment in these duplexes is different from the CCG alignment in the NMR DNA duplex. However, since the short strands slip with respect to each other in the

two-repeat duplex, the NMR structure also exhibits CpG steps between the Watson-Crick pairs.

In Chapter 6 we present results from molecular dynamics simulations that provide a detailed structural and dynamical characterization of the e-motif. We first encountered an e-motif in our study of the hexanucleotide repeats behind ALS and C9FTD diseases [229]. Here we extend this study and add the C-rich trinucleotide repeats in order to determine what sequences give rise to the e-motif, how stable it is and what are the mechanisms of the transition from internal C·C mismatch to e-motif.

As the author of this thesis, Feng Pan did the main work of production, analysis and part of writing for the papers included in Chapter 2, Chapter 4, Chapter 5 and Chapter 6. For the paper included in Chapter 3 Feng Pan did part of the production, analysis and writing.

References

[1] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737, 1953.

[2] P. Svoboda and A. Cara. Hairpin RNA: A secondary structure of primary importance. *Cellular and Molecular Life Sciences*, 63:901–908, 2006.

[3] David W. Staple and Samuel E. Butcher. Pseudoknots: RNA Structures with Diverse Functions . *CPLOS Biol.*, 3:6, 2005.

[4] C. R. Woese, S. Winkers, and R. R. Gutell. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci.*, 87:87, 1990.

[5] R. T. Batey, S. D. Gilbert, and R. K. Montange. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, 432:7015, 2004.

[6] D. C. Rau, B. Lee, and V. A. Parsegian. Measurement of the Repulsive Force between Polyelectrolyte Molecules in Ionic Solution: Hydration Forces between Parallel DNA Double Helices. *P. Natl. Acad. Sci. USA*, 81:2621–2625, 1984.

[7] C. Knobler and W. Gelbart. Physical Chemistry of DNA Viruses. *Ann. Rev. Phys. Chem.*, 60:367–383, 2009.

[8] M. Fedor. Comparative Enzymology and Structural Biology of RNA Self-Cleavage. *Ann. Rev. Biophys.*, 38:271–299, 2009.

[9] D. E. Draper, D. Grilley, and A. M. Soto. Ions and RNA Folding. *Annu. Rev. Bioph. Biom.*, 34:221–243, 2005.

[10] E. Koculi, C. Hyeon, D. Thirumalai, and S. A. Woodson. Charge Density of Divalent Metal Cations Determines RNA Stability. *J. Am. Chem. Soc.*, 129:2676–2682, 2007.

[11] P. Li, J. Vieregg, and I. Tinoco. How RNA Unfolds and Refolds. *Ann. Rev. Biophys.*, 77:77–100, 2008.

[12] A. MacKerell Jr. and L. Nilsson. Molecular Dynamics Simulations of Nucleic Acid-Protein Complexes. *Curr. Opin. Struct. Bio.*, 18:194–199, 2008.

[13] CT McMurray. DNA secondary structure: A common and causative factor for expansion in human disease. *Proc. Natl. Acad. Sci. USA*, 96:1823–1825, 1999.

[14] CE Pearson, KN Edamura, and JD Cleary. Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.*, 6:729–742, 2005.

[15] Mirkin, S. Expandable DNA repeats and human disease. *Nature*, 447:932, 2007.

[16] Laura P. W. Ranum and Thomas A. Cooper. RNA-mediated neuromuscular disorders. *Ann. Rev. of Neuroscience*, 6:259–277, 2006.

[17] Ling-Bo Li and Nancy M. Bonini. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends in Neurosciences*, 33:292–298, 2010.

[18] P Jin, DC Zarnescu, FP Zhang, CE Pearson, JC Lucchesi, K Moses, and ST Warren. RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in Drosophila. *Neuron*, 39:739–747, 2003.

[19] H Jiang, A Mankodi, MS Swanson, RT Moxley, and CA Thornton. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum. Mol. Genet.*, 13:3079–3088, 2004.

[20] R. Daughters, D. Tuttle, W. Gao, Y. Ikeda, M. Moseley, T. Ebner, M. Swanson, and L. Ranum. RNA Gain-of-Function in Spinocerebellar Ataxia Type 8. *PLoS Genet.*, 5:e1000600, 2009.

[21] W. Krzyzosiak, K. Sobczak, M. Wojciechowska, A. Fiszer, A. Mykowska, and P. Kozlowski. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nuc. Acids Res.*, 40:11–26, 2012.

[22] A H Wang, G J Quigley, F J Kolpak, J L Crawford, J H van Boom, G van der Marel, and A Rich. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, 282:680–686, 1979.

[23] A. Nordheim and A. Rich. The sequence $(dC - dA)_n.(dG - dT)_n$ forms left-handed Z-DNA in negatively supercoiled plasmids. *Proc. Natl. Acad. Sci.*, 80:1821, 1983.

[24] A. Rich, A. Nordheim, and A. H. Wang. The chemistry and biology of left-handed Z-DNA. *Ann. Rev. Phys. Chem.*, 53:791–846, 1984.

[25] G. Schroth, P. Chou, and P. J. Ho. Mapping Z-DNA in the human genome – computer-aided mapping revelas a nonrandom distribution of potential Z-DNA forming sequences in human genes. *J. Biol. Chem.*, 267:11846, 1992.

[26] L. F. Liu and J. C. Wang. Supercoiling of the DNA Template during Transcription. *Proc. Natl. Acad. Sci.*, 84:7024–7027, 1987.

[27] D. Oh, Y. Kim, and A. Rich. Z-DNA binding proteins can act as potent effectors of gene expression in vivo. *Proc. Natl. Acad. Sci.*, 99:16666, 2002.

[28] H. J. Lipps. Antibodies against Z-DNA react with the macronucleus but not the micronucleus of the hypotrichous ciliate stylonychia mytilus. *Cell*, 32:435–441, 1983.

[29] F. Lancillotti, M.C. Lopez, C. Alonso, and B.D. Stollar. Locations of Z-DNA in polytene chromosomes. *J. Cell Biol.*, 100:1759, 1985.

[30] A. Herbert and A. Rich. Left-handed Z-DNA: structure and function. *Genetica*, 106:37, 1999.

[31] E.B. Kmiec, K.J. Angelides, and W.K. Holloman. Left-handed DNA and the synaptic pairing reaction promoted by ustilago rec1 protein. *Cell*, 40:139, 1985.

[32] J.A. Blaho and R.D. Wells. Left-handed Z-DNA binding by the reca protein of escherichia-coli. *J. Biol. Chem.*, 262:6082, 1987.

[33] A. Jaworski, W.T. Hsieh, J.A. Blaho, J.E. Larson, and R.D. Wells. Left-handed DNA in vivo. *Science*, 238:773, 1987.

[34] T. Schwartz, M.A. Rould, K. Lowenhaupt, A. Herbert, and A. Rich. Crystal structure of the Z alpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, 284:1841, 1999.

[35] A.E. Vinogradov. DNA helix: the importance of being GC-rich. *Nucleic Acids Res.*, 31:1838, 2003.

[36] P.C. Champ, S. Maurice, J.M. Vargason, T. Champ, and P.S. Ho. Distributions of Z-DNA and nuclear factor I in human chromosome 22; a model for coupled transcriptional regulation. *Nucleic Acids Res.*, 32:6501, 2004.

[37] G. Wang, L. Christensen, and K. Vasquez. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci.*, 103:2677, 2006.

[38] Miguel A. Fuertes, Victoria Cepeda, Carlos Alonso, and José M. Pérez. Molecular mechanisms for the B-Z transition in the example of poly[d(G-C)·d(G-C)] polymers. a critical review. *Chem. Rev.*, 106(6):2045–2064, 2006.

[39] S.C. Harvey. DNA structural dynamics - longitudinal breathing as a possible mechanism for the B-reversible-Z transition. *Nucleic Acids Res.*, 11:4867, 1983.

[40] S. Goto. Characterization of intermediate conformational states in the B↔Z transition of $poly(dG-dC)\cdot poly(dG-dC)$. *Biopolymers*, 23:2211, 1984.

[41] W. Saenger and U. Hienemann. Raison d'etre and structural model for the B-Z transition of poly d(G-C)⋆poly d(G-C). *FEBS Lett.*, 257:223, 1989.

[42] A.T. Ansevin and A.H. Wang. Evidence for a new Z-type left-handed DNA helix - properties of Z(WC)-DNA. *Nucleic Acids Res.*, 18:6119, 1990.

[43] W. Lim and Y. Feng. The stretched intermediate model of B-Z DNA transition. *Biophys. J.*, 88:1593, 2005.

[44] S. C. Ha, K. Lowenhaupt, A. Rich, Y. G. Kim, and K. K. Kim. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature*, 437:1183, 2005.

[45] M. Moradi, V. Babin, C. Roland, and C. Sagui. Reaction path ensemble of the b-z-dna transition: a comprehensive atomistic study. *Nucleic Acids Res.*, 41:33–43, 2012.

[46] K. Hall, P. Cruz, I. Tinoco, T. Jovin, and J. van de Sande. Z-RNA - a left-handed RNA double helix. *Nature*, 311:584, 1984.

[47] R. W. Adamiak, A. Galat, and B. Skalski. Salt- and Solvent-dependent Conformational Transitions of Ribo-CGCGCG Duplex. *Biochim. Biophys. Acta*, 825:345–352, 1985.

[48] I. Tinoco Jr., P. Cruz, P. Davis, K. Hall, C. C. Hardin, R. A. Mathies, J. D. Puglisi, M. A. Trulson, W. C. Johnson Jr., and T. Neilson. Z-RNA: a left-handed double helix. In P. H. van Knippenberg and C. W. Hilbers, editors, *Structure and Dynamics of RNA*, volume 110 of *NATO ASI Series A*, pages 55–69. Plenum Press, New York, NY, 1986.

[49] M. Popenda, E. Biala, J. Milecki, and R. W. Adamiak. Solution structure of RNA duplexes containing alternating CG base pairs: NMR study of $r(CGCGCG)_2$

and $2' - O - Me(CGCGCG)_2$ under low salt conditions. *Nucleic Acids Res.*, 25:4589–4598, 1997.

[50] M. Popenda, J. Milecki, and R. W. Adamiak. High salt solution structure of a left-handed RNA double helix. *Nucleic Acids Res.*, 32:4044, 2004.

[51] A. Krzyzaniak, J. Barciszewski, J. P. Furste, R. Bald, V. A. Erdmann, P. Salanski, and J. Jurczak. A-Z-RNA Conformational Changes Effected by High Pressure. *Int. J. Biol. Macromol.*, 16:159–162, 1994.

[52] D. Placido, B. Brown, Ky. Lowenhaupt, A. Rich, and A. Athanasiadis. A left-handed RNA double helix bound by the Z-alpha domain of the RNA-editing enzyme ADAR1. *Structure*, 15:395, 2007.

[53] P. Auffinger, L. Bielecki, and E. Westhof. Anion Binding to Nucleic Acids. *Structure*, 12:379–388, 2004.

[54] G. G. Prive, K. Yanagi, and R. E. Dickerson. Structure of the B-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J. Mol. Biol.*, 217:177 – 199, 1991.

[55] X. Shui, C. C. Sines, L. McFail-Isom, D. VanDerveer, and L. D. Williams. Structure of the Potassium Form of CGCGAATTCGCG: DNA Deformation by Electrostatic Collapse around Inorganic Cations. *Biochemistry*, 37:16877–16887, 1998.

[56] X. Shui, L. McFail-Isom, G. Hu, and L. D. Williams. The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*, 37:8341 – 8355, 1998.

[57] V. Tereshko, G. Minasov, and M. Egli. A 'Hydrat-ion' Spine in a B-DNA Minor Groove. *J. Am. Chem. Soc.*, 121:3590–3595, 1999.

[58] L. McFail-Isom, C. C. Sines, and L. D. Williams. DNA Structure: Cations in Charge? *Curr. Opin. Struct. Bio.*, 9:298–304, 1999.

[59] E. Ennifar, M. Yusupov, P. Walter, R. Marquet, B. Ehresmann, C. Ehresmann, and P. Dumas. The Crystal Structure of the Dimerization Initiation Site of Genomic HIV-1 RNA Reveals an Extended Duplex with Two Adenine Bulges. *Struct. Fold. Des.*, 7:1439–1449, 1999.

[60] H. Robinson, Y. G. Gao, R. Sanishvili, A. Joachimiak, and A. H. J. Wang. Hexahydrated Magnesium Ions Bind in the Deep Major Groove and at the Outer Mouth of A-form Nucleic Acid Duplexes. *Nucleic Acids Res.*, 28:1760–1766, 2000.

[61] S. B. Howerton, C. C. Sines, D. VanDerveer, and L. D. Williams. Locating Monovalent Cations in the Grooves of B-DNA. *Biochemistry*, 40:10023–10031, 2001.

[62] J. A. Subirana and M. Soler-Lopez. Cations as Hydrogen Bond Donors: A View of Electrostatic Interactions in DNA. *Annu. Rev. Bioph. Biom.*, 32:27–45, 2003.

[63] T. K. Chiu and R. E. Dickerson. 1 Angstrom crystal structure of B-DNA real sequence specific and groove specific bending of DNA by magnesium and calcium. *J. Mol. Biol.*, 301:915, 2000.

[64] E. Ennifar, P. Walter, and P. Dumas. A Crystallographic Study of the Binding of 13 Metal Ions to Two Related RNA Duplexes. *Nucleic Acids Res.*, 31:2671–2682, 2003.

[65] Y. Timsit and S. Bombard. The 1.3 Angstrom Resolution Structure of the RNA Tridecamer r(GCGUUUGAAACGC): Metal ion Binding Correlates with Base Unstacking and Groove Contraction. *RNA-A Publication of the RNA Society*, 13:2098–2107, 2007.

[66] V. Tereshko, G. Minasov, and M. Egli. The Dickerson-Drew B-DNA Dodecamer Revisited at Atomic Resolution. *J. Am. Chem. Soc.*, 121:470–471, 1999.

[67] T. K. Chiu, M. Kaczor-Grzeskowiak, and R. E. Dickerson. Absence of Minor Groove Monovalent Cations in the Crosslinked Dodecamer CGCGAATTCGCG. *J. Mol. Biol.*, 292:589–608, 1999.

[68] K. Andresen, R. Das, H. Y. Park, H. Smith, L. W. Kwok, J. S. Lamb, E. J. Kirkland, D. Herschlag, K. D. Finkelstein, and L. Pollack. Spatial Distribution of Competing Ions around DNA in Solution. *Phys. Rev. Lett.*, 93:248103, 2004.

[69] V. Chu, Y. Bai, J. Lipfert, D. Herschlag, and S. Doniach. A Repulsive Field: Advances in the Electrostatics of the Ion Atmosphere. *Curr. Opin. Chem. Bio.*, 12:619–625, 2008.

[70] G. C. L. Wong and L. Pollack. Electrostatics of Strongly Charged Biological Polymers: Ion-Mediated Interactions and Self-Organization in Nucleic Acids and Proteins. *Ann. Rev. Phys. Chem.*, 61:171–189, 2010.

[71] Y. Bai, M. Greenfeld, K. Travers, V. Chu, J. Lipfert, S. Doniach, and D. Herschlag. Quantitative and Comprehensive Decomposition of the Ion Atmosphere around Nucleic Acids. *J. Am. Chem. Soc.*, 129:14981–14988, 2007.

[72] N. V. Hud and J. Feigon. Localization of Divalent Metal Ions in the Minor Groove of DNA A-tracts. *J. Am. Chem. Soc.*, 119:5756–5757, 1997.

[73] N. V. Hud, V. Sklenar, and J. Feigon. Localization of Ammonium Ions in the Minor Groove of DNA Duplexes in Solution and the Origin of DNA A-tract Bending. *J. Mol. Biol.*, 286:651–660, 1999.

[74] A. Bonvin. Localisation and Dynamics of Sodium Counterions around DNA in Solution from Molecular Dynamics Simulation. *Eur. Biophys. J. Biophy.*, 29:57–60, 2000.

[75] V. P. Denisov and B. Halle. Sequence-specific Binding of Counterions to B-DNA. *P. Natl. Acad. Sci. USA*, 97:629–633, 2000.

[76] F. C. Marincola, V. P. Denisov, and B. Halle. Competitive Na+ and Rb+ Binding in the Minor Groove of DNA. *J. Am. Chem. Soc.*, 126:6739–6750, 2004.

[77] T. Maehigashi, C. Hsiao, KK. Woods, T. Moulaei, N. V. Hud, and L. D. Williams. B-DNA Structure is Intrinsically Polymorphic: Even at the Level of Base Pair Positions. *Nucleic Acids Res.*, 40:3714–3722, 2012.

[78] G. S. Manning. Molecular Theory of Polyelectrolyte Solutions With Applications to Electrostatic Properties of Polynucleotides. *Q. Rev. Biophys.*, 11:179–246, 1978.

[79] R. J. Bacquet and P. J. Rossky. Ionic Distributions and Competitive Association on DNA Mixed Salt Solutions. *J. Phys. Chem.*, 92:3604–3612, 1988.

[80] D. M. York, T. Darden, D. Deerfield, and L. G. Pedersen. The interaction of NA(I), CA(II), and MG(II) metal-ions with duplex DNA – a theoretical modeling study. *Intl. J. Quant. Chem.*, 19:145–166, 1992.

[81] S. W. W. Chen and B. Honig. Monovalent and divalent salt effects on electrostatic free energies defined by the nonlinear Poisson-Boltzmann equation: Application to DNA binding reaction. *J. Phys. Chem. B*, 101:9113–9118, 1997.

[82] M. A. Young, B. Jayaram, and D. L. Beveridge. Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: fractional occupancy of electronegative pockets. *J. Am. Chem. Soc.*, 119:59–69, 1997.

[83] M. Feig and B. M. Pettitt. Sodium and Chlorine Ions as Part of the DNA Solvation Shell. *Biophys. J.*, 77:1769–1781, 1999.

[84] P. Auffinger and E. Westhof. Water and Ion Binding around RNA and DNA (C,G) Oligomers. *J. Mol. Biol.*, 300:1113–1131, 2000.

[85] M. Orozco, A. Pérez, A. Noy, and F. J. Luque. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.*, 32:350–364, 2003.

[86] M. Rueda, E. Cubero, C. A. Laughton, and M. Orozco. Exploring the Counterion Atmosphere around DNA: What can be Learned from Molecular Dynamics Simulations? *Biophys. J.*, 87:800–811, 2004.

[87] S. Y. Ponomarev, K. M. Thayer, and D. L. Beveridge. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci.*, 101:14771–14775, 2004.

[88] P. Várnai and Krystyna Zakrzewska. DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.*, 320:4269–4280, 2004.

[89] C. H. Taubes, U. Mohanty, and S. Chu. Ion atmosphere around nucleic acid. *J. Phys. Chem. B*, 109:21267–21272, 2005.

[90] P. Auffinger and Y. Hashem. Nucleic Acid Solvation: From Outside to Insight. *Curr. Opin. Struct. Bio.*, 17:325–333, 2007.

[91] A. E. Garcia and D. Paschek. Simulation of the Pressure and Temperature folding/unfolding Equilibrium of a Small RNA Hairpin. *J. Am. Chem. Soc.*, 130:815, 2008.

[92] Jejoong Yoo and Aleksei Aksimentiev. Competitive Binding of Cations to Duplex DNA Revealed through Molecular Dynamics Simulations. *J. Phys. Chem. B*, 116:12946–12954, 2012.

[93] A. Perez, F. Javier Luque, and M. Orozco. Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.*, 45:196–205, 2012.

[94] G. Andres Cisneros, Mikko Karttunen, Pengyu Ren, and Celeste Sagui. Classical Electrostatics for Biomolecular Simulations. *Chem. Rev.*, 114(1):779–814, 2014.

[95] D. M. York, T. A. Darden, and L. G. Pedersen. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *J. Chem. Phys.*, 99:8345 – 8348, 1993.

[96] T. A. Darden, D. M. York, and L. G. Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089 – 10092, 1993.

[97] D. M. York, W. Yang, H. Lee, T. A. Darden, and L. G. Pedersen. Towards the accurate modeling of DNA: the importance of long-range electrostatics. *J. Am. Chem. Soc.*, 117:5001 – 5002, 1995.

[98] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577 – 8593, 1995.

[99] I. Besseova, M. Otyepka, K. Reblova, and J. Sponer. Dependence of A-RNA Simulations on the Choice of the Force Field and Salt Strength. *Phys. Chem. Chem. Phys.*, 11:10701–10711, 2009.

[100] S. Kirmizialtin and R. Elber. Computational Exploration of Mobile Ion Distributions Around RNA Duplex. *J. Phys. Chem. B*, 114:8207–8220, 2010.

[101] I. Besseova, P. Barnas, P. Kuhrova, P. Kosinova, M. Otyepka, and J. Sponer. Simulations of A-RNA Duplexes. The Effect of Sequence, Solute Force Field, Water Model, and Salt Concentration. *J. Phys. Chem. B*, 116:9899–9916, 2012.

[102] S. Kirmizialtin, S. A. Pabit, S. P. Meisburger, L. Pollack, and R. Elber. RNA and Its Ionic Cloud: Solution Scattering Experiments and Atomically Detailed Simulations. *Biophys. J.*, 102:819–828, 2012.

[103] S. Kirmizialtin, A. R. Silalahi, R. Elber, and M. O. Fenley. The Ionic Atmosphere around A-RNA: Poisson-Boltzmann and Molecular Dynamics Simulations. *Biophys. J.*, 102:829–838, 2012.

[104] A. Perez, I. Marchan, 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.*, 92:3817–3829, 2007.

[105] P. Banas, D. Hollas, M. Zgarbva, P. Jurecka, M. Orozco, T. E. Cheatham, J. Sponer, and M. Otyepka. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.*, 6:3836–3849, 2010.

[106] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham, and P. Jurecka. 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.

[107] D. Ozawa, H. Yagi, T. Ban, A. Kameda, T. Kawakami, H. Naiki, and Y. Goto. Destruction of amyloid fibrils of a beta-microglobin fragment by laser beam irradiation. *J. Biol. Chem.*, 284:1009, 2009.

[108] H. Yagi, D. Ozawa, K. Sakuri, T. Kawakami, H. Kuyama, O. Nishimura, T. Shimanouchi, R. Kuboi, H. Naiki, and Y. Goto. Laser-induced propagation and destruction of amyloid beta fibrils. *J. Biol. Chem.*, 285:19660, 2010.

[109] T. Kawasaki, J. Fujioka, T. Imai, and K. Tsukiyama. Effect of mid-infrared free-electron laser irradiation on refolding of amyloid-like fibrils of Lysozyme into native form. *The Protein Journal*, 31:710, 2012.

[110] T. Kawasaki, J. Fujioka, T. Imai, K. Torigoe, and K. Tsukiyama. Mid-infrared free-electron laser tuned to the amide I band for converting insoluble amyloid-like protein fibrils into the soluble monomeric form. *Lasers in Medical Science*, 29:1701, 2014.

[111] T. Kawasaki, T. Yaji, T. Imai, T. Ohta, and K. Tsukiyama. Synchrotron-infrared microscopy analysis of amyloid fibrils irradiated by mid-infrared free-electron laser. *Am. J. of Anal. Chem.*, 5:384, 2014.

[112] V. H. Man, Ph. Derreumaux, M. S. Li, C. Roland, C. Sagui, and Ph. Nguyen. Picosecond dissociation of amyloid fibrils with infrared laser: a nonequilibrium simulation study. *J. Chem. Phys.*, submitted:, 2015.

[113] V.H. Man, P.M. Truong, Ph. Derreumaux, M.S. Li, C. Roland, C. Sagui, and Ph. Nguyen. Picosecond melting of peptide nanotube with infrared laser: a nonequilibrium study. *Nano Lett.*, submitted:, 2015.

[114] J.-F. Mergny and L. Lacroix. Analysis of Thermal Melting Curves. *Oligonucleotides*, 13(6):515–537, 2003.

[115] R. Thomas. The denaturation of DNA. *Gene*, 135(1-2):77–9, 1993.

[116] J.Z. Jin, K. J. Breslauer, R. A. Jones, and B. L. Gaffney. Tetraplex formation of a guanine-containing nonameric DNA fragment. *Science*, 250(4980):543–6, 1990.

[117] C. Hardin, E. Henderson, T. Watson, and J. K. Prosser. Monovalent cation induced structural transitions in telomeric DNAs: G-DNA folding intermediates. *Biochemistry*, 30(18):4460–4472, 1991.

[118] J.-L. Mergny and J.-C. Maurizot. Fluorescence Resonance Energy Transfer as a Probe for G-Quartet Formation by a Telomeric Repeat. *Chembiochem*, 2(2):124–32, 2001.

[119] J. G. Duguid, V. A. Bloomfield, J. M. Benevides, and G. J. Thomas. DNA Melting Investigated by Differential Scanning Calorimetry and Raman Spectroscopy. *Biophys J.*, 71(6):3350–60, 1996.

[120] L. Movileanu, J. M. Benevides, and G. J. Thomas. Determination of Base and Backbone Contributions to the Thermodynamics of Premelting and Melting Transitions in B DNA. *Nucleic Acids Res.*, 30(17):3767–77, 2002.

[121] J.-L. Mergny, A.-T. Phan, and L. Lacroix. Following G-quartet Formation by UV-spectroscopy. *FEBS Lett.*, 435(1):74–8, 1998.

[122] P. Cahen, M. Luhmer, C. Fontaine, C. Morat, J. Reisse, and K. Bartik. Study by (23)Na-NMR, (1)H-NMR, and ultraviolet spectroscopy of the thermal stability of an 11-basepair oligonucleotide. *Biophys J.*, 78(2):1059–1069, 2000.

[123] D. Poland and H. A. Scheraga. Phase Transitions in One Dimension and the Helix–Coil Transition in Polyamino Acids. *J. Chem. Phys.*, 45(5):1456–6, 1966.

[124] D. Poland. Recursion relation generation of probability profiles for specific sequence macromolecules with long range correlations. *Biopolymers*, 13(9):1859–1871, 1974.

[125] M. Fixman and J. J. Freire. Theory of DNA melting curves. *Biopolymers*, 16(12):2693–2704, 1977.

[126] M. Y. Azbel. Phase transitions in DNA. *Phys. Rev. A*, 20:1671, 1979.

[127] K. J. Breslauer, R. Frank, H. Bloecker, and L. A. Marky. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA*, 83(11):3746–3750, 1986.

[128] M. Peyrard and A. R. Bishop. Statistical mechanics of a nonlinear model for DNA denaturation. *Phys. Rev. Lett.*, 62(23):2755–2758, 1989.

[129] T. Dauxois, M. Peyrard, and A. R. Bishop. Entropy-driven DNA denaturation. *Phys. Rev. E*, 47(1):R44–R47, 1993.

[130] T. Dauxois, M. Peyrard, and A. R. Bishop. Dynamics and thermodynamics of a nonlinear model for DNA denaturation. *Phys. Rev. E*, 47(1):684–95, 1993.

[131] T. Dauxois and M. Peyrard. Entropy-driven transition in a one-dimensional system. *Phys. Rev. E*, 51(5):4027–4040, 1995.

[132] N. Sugimoto, S.-I. Nakano, M. Yoneyama, and K.-I. Honda. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, 24(22):4501–4505, 1996.

[133] J. Santa Lucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95(4):1460–1465, 1997.

[134] R. Owczarzy, P. M. Vallone, F.J. Gallo, T. M. Paner, M. J. Lane, and A. S. Benight. Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers*, 44(3):217–239, 1998.

[135] A. Campa and A. Giansanti. Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains. *Phys. Rev. E*, 58(3-B):3585–3588, 1998.

[136] T. V. Chalikian, J. Volker, G. E. Plum, and K. J. Breslauer. A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques. *Proc. Natl. Acad. Sci. USA*, 96(14):7853–7857, 1999.

[137] V. Ivanov, Y. Zeng, and G. Zocchi. Statistical mechanics of base stacking and pairing in DNA melting. *Phys. Rev. E*, 70(5):051907, 2004.

[138] D. Poland. DNA melting profiles from a matrix method. *Biopolymers*, 73(2):216–228, 2004.

[139] C. Richard and A. J. Guttmann. Poland-Scheraga Models and the DNA Denaturation Transition. *J. Stat. Phys.*, 115(3-4):925–947, 2004.

[140] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA*, 83(24):9373–9377, 1986.

[141] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci.*, 86(20):7706–10, 1989.

[142] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.

[143] K. B. Hall and L. W. W. McLaughlin. Thermodynamic and structural properties of pentamer DNA.DNA, RNA.RNA, and DNA.RNA duplexes of identical sequence. *Biochemistry*, 30(44):10606–13, 1991.

[144] L. Ratmeyer, R. Vinayak, Y. Zhong, G. Zon, and W. D. Wilson. Sequence specific thermodynamic and structural properties for DNA.RNA duplexes. *Biochemistry*, 33(17):5298–5304, 1994.

[145] N. Sugimoto, K.-I. Honda, and M. Sasaki. Application of the thermodynamic parameters of DNA stability prediction to double-helix formation of deoxyribooligonucleotides . *Nucleosides and Nucleotides*, 13(6–7):1311–1317, 1994.

[146] N. Sugimoto, M. Katoh, S.-I. Nakano, T. Ohmichi, and M. Sasaki. RNA/DNA hybrid duplexes with identical nearest-neighbor base-pairs have identical stability. *FEBS Lett.*, 354(1):74–78, 1994.

[147] M. J. Doktycz, M. Morris, S. J. Dormady, K. L. Beattie, and B. Jacobson. Optical melting of 128 octamer DNA duplexes. Effects of base pair location and nearest neighbors on thermal stability. *J. Biol. Chem.*, 270(15):8439–45, 1995.

[148] J. Santa Lucia, H. T. Allawi, and P. A Seneviratne. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 35(11):3555–3562, 1996.

[149] J. Santa Lucia and D. H. Turner. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, 44(3):309–319, 1997.

[150] J. Santa Lucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95(4):1460–5, 1998.

[151] D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt, and D. H. Turner. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, 5(11):1458–1469, 1999.

[152] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–940, 1999.

[153] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

[154] Z. Dwight, R. Palais, and C. T. Wittwer. uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics*, 27(7):1019–1020, 2010.

[155] A. Herbert, M. Schade, K. Lowenhaupt, J. Alfken, Th. Schwartz, L. S. Shlyakhtenko, Y. L. Lyubchenko, and A. Rich. The Zalpha domain from human ADAR1 binds to the Z-DNA conformer of many different sequences. *Nucleic Acids Res.*, 26(15):3486, 1998.

[156] J. D. Kahmann, D. A. Wecking, V. Putter, K. Lowenhaupt, Y.-G. Kim, P. Schmieder, H. Oschkinat, A. Rich, and M. Schade. The solution structure of the n-terminal domain of e3l shows a tyrosine conformation that may explain its reduced affinity to z-dna in vitro. *Proc. Natl. Acad. Sci.*, 101(9):2712–2717, 2004.

[157] S. C. Ha, J. Choi, H.-Y. Hwang, A. Rich, Y.-G. Kim, and K. K. Kim. The structures of non-cg-repeat z-dnas co-crystallized with the z-dna-binding domain, hz alpha(adar1). *Nucleic Acids Res.*, 37(2):629–637, 2009.

[158] C.-X. Wu, S.-J. Wang, G. Lin, and C.-Y. Hu. The Zalpha domain of PKZ from Carassius auratus can bind to d(GC)(n) in negative supercoils. *Fish Shellfish Immunol.*, 28(5):783, 2010.

[159] L. Yang, S. Wang, T. Tian, and X. Zhou. Advancements in Z-DNA: development of inducers and stabilizers for B to Z transition. *Curr. Med. Chem.*, 19(4):557, 2012.

[160] R. Elber, A. Ghosh, and A. Cardenas. Long time dynamics of complex systems. *Acc. Chem. Res.*, 35:396, 2002.

[161] M. A. Kastenholz, T. U. Schwartz, and P. H. Hünenberger. The transition between the B and Z conformations of DNA investigated by targeted molecular dynamics simulations with explicit solvation. *Biophys. J.*, 91(8):2976 – 2990, 2006.

[162] J. Lee, Y.G. Kim, K.K. Kim, and C. Seok. Transition between B-DNA and Z-DNA: free energy landscape for the B-Z junction propagation. *J. Phys. Chem. B*, 114:9872, 2010.

[163] M. Moradi, V. Babin, C. Roland, and C. Sagui. Reaction path ensemble of the b-z-dna transition: a comprehensive atomistic study. *Nucleic Acids Res.*, 41:33–43, 2013.

[164] F. Pan and C. Roland and C. Sagui. Ion distribution around left- and right-handed DNA and RNA duplexes: a comparative study. *Nucl. Acids Res.*, 42:13981–96, 2014.

[165] J. H. van de Sande and T. M. Jovin. Z* DNA, the left-handed helical form of poly[d(G-C)] in MgCl2-ethanol, is biologically active. *EMBO J.*, 1(1):115–120, 1982.

[166] D. J. Patel, S. A. Kozlowski, A. Nordheim, and A. Rich. Right-handed and left-handed DNA: studies of B- and Z-DNA by using proton nuclear Overhauser effect and P NMR. *Proc. Natl. Acad. Sci. USA*, 79(5):1413–1417, 1982.

[167] K. B. Roy and H. T. Miles. A thermally driven interconversion of B and Z-DNA. *Biochem. Biophys. Res. Commun.*, 115(1):100–105, 1983.

[168] H Ellegren. Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.*, 5:435–445, 2004.

[169] I Oberle, F. Rouseau, D. Heitz, D. Devys, S. Zengerling, and J.L. Mandel. Molecular-basis of the fragile-X syndrome and diagnostic applications. *Am. J. of Human Genet.*, 49:76, 1991.

[170] P Giunti, MG Sweeney, M Spadaro, C Jodice, A Novelletto, P Malaspina, M Frontali, and Harding AE. The trinucleotide repeat expansion on chromosome 6p (sca1) in autosomal dominant cerebellar ataxias. *Brain*, 117:645–649, 1994.

[171] V Campuzano, L Montermini, MD Molto, L Pianese, M Cossee, F Cavalcanti, E Monros, F Rodius, F Duclos, A Monticelli, F Zara, J Canizares, H Koutnikova, SI Bidichandani, C Gellera, A Brice, P Trouillas, G DeMichele, A Filla, R DeFrutos, F Palau, PI Patel, S DiDonato, JL Mandel, S Cocozza, M Koenig, and M Pandolfo. Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, 271:1423–1427, 1996.

[172] Sergei M. Mirkin. DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. in Struct. Biol.*, 16:351–358, 2006.

[173] Wells, R.D. and Warren, S. *Genetic instabilities and neurological diseases*. Academic Press, San Diego, CA, Elsevier, 1998.

[174] Orr, H. and Zoghbi, H. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, 30:575, 2007.

[175] CE Pearson and RR Sinden. Slipped strand DNA (S-DNA and SI-DNA), trinucleotide repeat instability and mismatch repair: A short review. In Sarma, RH and Sarma, MH, editor, *Structure, Motion, Interaction and Expression of Biological Macromolecules, Vol 2*, pages 191–207. US NIH, 1998. 10th Conversation in Biomolecular Stereodynamics Conference, SUNY Albany, JUN 17-21, 1997.

[176] RD Wells, R Dere, ML Hebert, M Napierala, and LS Son. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucl. Acids Res.*, 33:3785–3798, 2005.

[177] Jane C. Kim and Sergei M. Mirkin. The balancing act of DNA repeat expansions. *Curr. Opin. in Genet. & Devel.*, 23:280–288, 2013.

[178] JP Cleary, DM Walsh, JJ Hofmeister, GM Shankar, MA Kuskowski, DJ Selkoe, and KH Ashe. Natural oligomers of the amyloid-protein specifically disrupt cognitive function. *Nat. Neurosci.*, 8:79–84, 2005.

[179] Vincent Dion and John H. Wilson. Instability and chromatin structure of expanded trinucleotide repeats. *Trends in Genet.*, 25:288–297, 2009.

[180] Cynthia T. McMurray. Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair*, 7:1121–1134, 2008.

[181] V Campuzano, L Montermini, Y Lutz, L Cova, C Hindelang, S Jiralerspong, Y Trottier, SJ Kish, B Faucheux, P Trouillas, FJ Authier, A Durr, JL Mandel, A Vescovi, M Pandolfo, and M Koenig. Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum. Mol. Genet.*, 6:1771–1780, 1997.

[182] E. Kim, M. Napierala, and S. Dent. Hyperexpansion of GAA repeats affects post-initiation steps of FXN transcription in Friedreich's ataxia. *Nucl. Acids Res.* , 39:8366–8377, 2011.

[183] D. Kumari, R. Biacsi, and K. Usdin. Repeat expansion in intron 1 of the Frataxin gene reduces transcription initiation in Friedreich ataxia. *Faseb J.*, 25:895, 2011. Experimental Biology Meeting 2011, Washington, DC, APR 09-13, 2011.

[184] T. Punga and M. Buehler. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *Embo Mol. Medicine*, 2:120–129, 2010.

[185] Jason R. O'Rourke and Maurice S. Swanson. Mechanisms of RNA-mediated Disease. *J. Biol. Chem.*, 284(12):7419–7423, 2009.

[186] Peter K. Todd and Henry L. Paulson. RNA-mediated neurodegeneration in repeat expansion disorders. *Ann. Neurol.*, 67(3):291–300, 2010.

[187] Gloria V. Echeverria and Thomas A. Cooper. RNA-binding proteins in microsatellite expansion disorders: mediators of RNA toxicity. *Brain Res.*, 1462:100–111, 2012.

[188] Marzena Wojciechowska and Wlodzimierz J. Krzyzosiak. Cellular toxicity of expanded RNA repeats: focus on RNA foci. *Hum. Mol. Genet.*, 20(19):3811–3821, 2011.

[189] Tao Zu, Brian Gibbens, Noelle S. Doty, Mario Gomes-Pereira, Aline Huguet, Matthew D. Stone, Jamie Margolis, Mark Peterson, Todd W. Markowski, Melissa A. C. Ingram, Zhenhong Nan, Colleen Forster, Walter C. Low, Benedikt Schoser, Nikunj V. Somia, H. Brent Clark, Stephen Schmechel, Peter B. Bitterman, Genevieve Gourdon, Maurica S. Swanson, Melinda Moseley, and Laura P. W. Ranum. Non-atg–initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci USA*, 108(1):260–265, 2011.

[190] Huda Y. Zoghbi and Harry T. Orr. Glutamine repeats and neurodegeneration. *Annual Review of Neuroscience*, 23(1):217–247, 2000.

[191] Stephen W. Davies, Mark Turmaine, Barbara A. Cozens, Marian DiFiglia, Alan H. Sharp, Christopher A. Ross, Eberhard Scherzinger, Erich E. Wanker, Laura Mangiarini, and Gillian P. Bates. Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the hd mutation. *Cell*, 90:537–548, 1997.

[192] Pawel Sikorski and Edward Atkins. New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils. *Biomacromolecules*, 6(1):425–432, 2005.

[193] Deepak Sharma, Leonid M. Shinchuk, Hideyo Inouye, Ronald Wetzel, and Daniel A. Kirschner. Polyglutamine homopolymers having 8-45 residues form slablike beta-crystallite assemblies. *Proteins*, 61(2):398–411, 2005.

[194] Robert Schneider, Miria C. Schumacher, Henrik Mueller, Deepak Nand, Volker Klaukien, Henrike Heise, Dietmar Riedel, Gerhard Wolf, Elmar Behrmann, Stefan Raunser, Ralf Seidel, Martin Engelhard, and Marc Baldus. Structural characterization of polyglutamine fibrils by solid-state NMR spectroscopy. *J. Mol. Biol.*, 412(1):121–136, 2011.

[195] Lauren E. Buchanan, Joshua K. Carr, Aaron M. Fluitt, Andrew J. Hoganson, Sean D. Moran, Juan J. de Pablo, James L. Skinner, and Martin T. Zanni. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. *Proc. Natl. Acad. Sci.*, 111(11):5796–801, 2014.

[196] Karunakar Kar, Cody L. Hoop, Kenneth W. Drombosky, Matthew A. Baker, Ravindra Kodali, Irene Arduini, Patrick C.A. van der Wel, W. Seth Horne, and Ronald Wetzel. β-hairpin-mediated nucleation of polyglutamine amyloid formation. *J. Mol. Biol.*, 425(7):1183–1197, 2013.

[197] Viet Hoang Man, Christopher Roland, and Celeste Sagui. Structural determinants of polyglutamine protofibrils and crystallites. *ACS Chem. Neurosci.*, 6(4):632–645, 2015.

[198] Y. Zhang, V.H. Man, C Roland, and C. Sagui. Amyloid properties of asparagine and glutamine in prion-like proteins. *ACS Chem. Neurosci.*, 7:576–587, 2016.

[199] E. Delot, L. M. King, M. D. Briggs, W. R. Wilcox, and D. H. Cohn. Trinucleotide Expansion Mutations in the Cartilage Oligomeric Matrix Protein (Comp) Gene. *Hum Mol Genet*, 8:123–128, 1999.

[200] M. Vorlickova, I. Kejnovska, M. Tumova, and J. Kypr. Conformational properties of DNA fragments containing GAC trinucleotide repeats associated with skeletal displasias. *Eur. Biophys. J.*, 30:197–85, 2001.

[201] M Mitas, A Yu, J Dill, and IS Haworth. The trinucleotide repeat sequence D(CGG)(15) forms a heat-stable hairpin containing G(syn).G(anti) base-pairs. *Biochem.*, 34:12803–12811, 1995.

[202] AM Gacy, G Goellner, N Juranic, S Macura, and CT McMurray. Trinucleotide repeats that expand in human-disease form hairpin structures in-vitro . *Cell*, 81:533–540, 1995.

[203] J Petruska, N Arnheim, and MF Goodman. Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nuc. Acids Res.* , 24:1992–1998, 1996.

[204] Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak, and Wojciech Rypniewski. Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nuc. Acids Res.*, 38(22):8370–8376, 2010.

[205] Ilyas Yildirim, HaJeung Park, Matthew D. Disney, and George C. Schatz. A dynamic structural model of expanded rna cag repeats: A refined x-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *JACS*, 135(9):3528–38, 2013.

[206] Arpita Tawani and Amit Kumar. Structural Insights Reveal the Dynamics of the Repeating r(CAG) Transcript Found in Huntington's Disease (HD) and Spinocerebellar Ataxias (SCAs). *PLoS One*, 10(7):e0131788, 2015.

[207] Daniel Svozil, Pavel Hobza, and Jiří Šponer. Comparison of intrinsic stacking energies of ten unique dinucleotide steps in A-RNA and B-DNA duplexes. Can we determine correct order of stability by quantum-chemical calculations? *J. Phys. Chem. B*, 114(2):1191–203, 2010.

[208] Noorain Khan, Narendar Kolimi, and Thenmalarchelvi Rathinavelan. Twisting right to left: A...a mismatch in a cag trinucleotide repeat overexpansion provokes left-handed z-dna conformation. *PLoS Comput Biol.*, 11(4):e1004162, 2015.

[209] Thomas Cheatham III and David Case. Twenty five years of nucleic acid simulations. *Biopolymers*, 99:969, 2013.

[210] Jiri Sponer and Miroslav Krepl and Pavel Banas and Petra Kuhrova and Marie Zgarboba and Petr Jurecka and Marek Havrila and Michal Otyepka. How to understand atomistic molecular dynamics simulations of RNA and protein-RNA complexes? *WIREs RNA*, 8, 2017.

[211] Ying-Hui Fu, Derek P.A. Kuhl, Antonio Pizzuti, Maura Pieretti, James S. Sutcliffe, Stephen Richards, Annemieke J.M.H. Verkert, Jeanette J.A. Holden, Raymond G. Fenwick Jr., Stephen T. Warren, Ben A. Oostra, David L. Nelson, and C.Thomas Caskey. Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. *Cell*, 67(6):1047–1058, 1991.

[212] Nan Zhong, Weina Ju, James Pietrofesa, Daowen Wang, Carl Dobkin, and W. Ted Brown. Fragile X "gray zone" alleles: AGG patterns, expansion risks, and associated haplotypes. *Am J Med Genet.*, 64(2):261–5, 1996.

[213] C. Dombrowski, S. Lévesque, M. L. Morel, P. Rouillard, K. Morgan, and F. Rousseau. Premutation and intermediate-size FMR1 alleles in 10 572 males from the general population: loss of an AGG interruption is a late event in the generation of fragile X syndrome alleles. *Hum Mol Genet.*, 11(4):371–378, 2002.

[214] RJ Hagerman, M Leehey, W Heinrichs, F Tassone, R Wilson, J Hills, J Grigsby, B Gage, and PJ Hagerman. Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X. *Neurology*, 57(1):127–30, 2001.

[215] Stephanie L. Sherman. Premature Ovarian Failure among Fragile X Premutation Carriers: Parent-of-Origin Effect? *Am J Hum Genet.*, 67(1):11–3, 2000.

[216] I.A. Glass. X linked mental retardation. *J. Med. Genet.*, 28:361–371, 1991.

[217] Yanghong Gu, Ying Shen, Richard A. Gibbs, and David L. Nelson. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nat. Genet.*, 13(1):109–113, 1996.

[218] Baorong Zhang, Jun Tian, Yaping Yan, Xinzhen Yin, Guohua Zhao, Zhiying Wu, Weihong Gu, Kun Xia, and Beisha Tang. CCG polymorphisms in the huntingtin gene have no effect on the pathogenesis of patients with Huntington's disease in mainland Chinese families. *J. Neurol. Sci.*, 312(1-2):92–96, 2012.

[219] Claudia Braida, Rhoda K.A. Stefanatos, Berit Adam, Navdeep Mahajan, Hubert J.M. Smeets, Florence Niel, Cyril Goizet, Benoit Arveiler, Michel Koenig, Clotilde Lagier-Tourenne, Jean-Louis Mandel, Catharina G. Faber, Christine E.M. de Die-Smulders, Frank Spaans, and Darren G. Monckton. Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum. Mol. Genet.*, 19(1-2):1399–1412, 2010.

[220] M Mitas. Trinucleotide repeats associated with human disease. *Nuc. Acids Res.*, 25:2245–2253, 1997.

[221] Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak, and Wojciech Rypniewski. Crystal structures of CGG RNA repeats with implications for fragile X-associated tremor ataxia syndrome. *Nuc. Acids Res.*, 39:7308–7315, 2011.

[222] Amit Kumar, Pengfei Fang, Hajeung Park, Min Guo, Kendall W. Nettles, and Matthew D. Disney. A crystal structure of a model of the repeating r(cgg) transcript found in fragile syndrome. *Chembiochem.*, 12(14):2140–2142, Sep. 2011.

[223] Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak, and Wojciech Rypniewski. Crystallographic characterization of CCG repeats. *Nucl. Acids Res.*, 40:8155–8162, 2012.

[224] XL Gao, XN Huang, GK Smith, MX Zheng, and HY Liu. New antiparallel duplex motif of DNA CCG repeats that is stabilized by extrahelical basis symmetrically located in the minor-groove. *J. Am. Chem. Soc.*, 117:8883–8884, 1995.

[225] MX Zheng, XN Huang, GK Smith, XY Yang, and XL Gao. Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study. *J. of Mol. Biol.*, 264:323–336, 1996.

[226] JM Darlow and DRF Leach. Secondary structures in d(CGG) and d(CCG) repeat tracts. *J. Mol. Biol.*, 275:3–16, 1998.

[227] JM Darlow and DRF Leach. Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts in vivo. *J. Mol. Biol.*, 275:17–23, 1998.

[228] F. Pan and V. Man and C. Roland and C. Sagui. Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats. *Biophys. J.*, 113:19–36, 2017.

[229] Y. Zhang, C. Roland, and C. Sagui. Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases. *ACS Chemical Neuroscience*, 8:578–591, 2016.

[230] Mariely DeJesus-Hernandez, Ian R. Machkenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, NiCole A. Finch, Heather Flynn, Jennifer Adamson, Naomi Kouri, Aleksandra Wojtas, Pheth Sengdy, Ging-Yuek R. Hsiung, Anna Karydas, William W. Seeley, Keith A. Josephs, Giovanni Coppola, Daniel H. Geschwind, Zbigniew K. Wszolek, Howard Feldman, David S. Knopman, Ronald C. Petersen, Bruce L. Miller, and Dennis W. Dickson. Expanded ggggcc hexanucleotide repeat in noncoding region of c9orf72 causes chromosome 9p-linked ftd and als. *Neuron*, 72(2):245–256, Oct. 2011.

[231] Alan E. Renton, Elisa Majounie, Adrian Waite, Javier Simon-Sanchez, Sara Rollinson, J. Raphael Gibbs, Jennifer C. Schymick, Hannu Laaksovirta, John C. van Swieten, Liisa Myllykangas, Hannu Kalimo, Anders Paetau, Yevgeniya Abramzon, Anne M. Remes, Alice Kaganovich, Sanja W. Scholz, Jamie Duckworth, Jinhui Ding, Daniel W. Harmer, Dena G. Hernandez, Janel O. Johnson, Kin Mok, Mina Ryten, Danyah Trabzuni, and Rita J. Guerreiro. A hexanucleotide repeat expansion in c9orf72 is the cause of chromosome 9p21-linked als-ftd. *Neuron*, 72(2):257–268, Oct. 2011.

Chapter 2

Ion distributions around left- and right-handed DNA and RNA duplexes: a comparative study

Feng Pan, Christopher Roland and Celeste Sagui. (2014) Ion distributions around left- and right-handed DNA and RNA duplexes: a comparative study. *Nucleic Acids Research 42*, 13981-13996.

Published online 26 November 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 22* ***13981–13996***
doi: 10.1093/nar/gku1107

Ion distributions around left- and right-handed DNA and RNA duplexes: a comparative study

Feng Pan, Christopher Roland and Celeste Sagui*

Center for High Performance Simulations (CHiPS) and Department of Physics, North Carolina State University, Raleigh, NC 27695-8202, USA

Received September 03, 2014; Revised October 22, 2014; Accepted October 23, 2014

ABSTRACT

The ion atmosphere around nucleic acids is an integral part of their solvated structure. However, detailed aspects of the ionic distribution are difficult to probe experimentally, and comparative studies for different structures of the same sequence are almost non-existent. Here, we have used large-scale molecular dynamics simulations to perform a comparative study of the ion distribution around (5′-CGCGCGCGCGCG-3′)$_2$ dodecamers in solution in B-DNA, A-RNA, Z-DNA and Z-RNA forms. The CG sequence is very sensitive to ionic strength and it allows the comparison with the rare but important left-handed forms. The ions investigated include Na^+, K^+ and Mg^{2+}, with various concentrations of their chloride salts. Our results quantitatively describe the characteristics of the ionic distributions for different structures at varying ionic strengths, tracing these differences to nucleic acid structure and ion type. Several binding pockets with rather long ion residence times are described, both for the monovalent ions and for the hexahydrated $Mg[(H_2O)_6]^{2+}$ ion. The conformations of these binding pockets include direct binding through desolvated ion bridges in the GpC steps in B-DNA and A-RNA; direct binding to backbone oxygens; binding of $Mg[(H_2O)_6]^{2+}$ to distant phosphates, resulting in acute bending of A-RNA; tight 'ion traps' in Z-RNA between C-O2 and the C-O2′ atoms in GpC steps; and others.

INTRODUCTION

Helices formed by natural amino acids and nucleotides are predominantly right-handed, while left-handed forms, such as PPII helices in proteins and Z-DNA and Z-RNA duplexes, are relatively rare. A left-handed, double-helix DNA with two antiparallel chains joined by Watson-Crick (WC) base pairs was first revealed by a crystal structure of d(CGCGCG)$_2$ in 1979 (1). The term 'Z-DNA' was coined for this structure because the sugar-phosphate backbone displays a characteristic zig-zag pattern. Z-DNA is formed by dinucleotide repeats, and is a characteristic of sequences that alternate purines and pyrimidines, mainly CG or GC. These kinds of base pairs give rise to an anti–syn alternation, which is due to rotation of the guanine residue around its glycosidic bond, resulting in a syn conformation, while the cytosine retains its anti configuration (2,3). A high density of base sequences favoring Z-DNA is found near transcription start sites (4), where Z-DNA is stabilized by negative supercoiling of DNA (3,5). Z-DNA is induced by a set of binding proteins near promoter regions, which boosts the transcription of downstream genes (6). Z-DNA is highly immunogenic, and antibodies against it (7–9) are used to find locations prone to Z-DNA conformations. The current view is that Z-DNA formation plays a role in gene expression, regulation and recombination (3,6,10–16).

Since the right-handed form is DNA's dominant duplex conformation, research has focused on the microscopic mechanisms behind the B-Z DNA transition and controversial models have ensued (17). Proposed transition mechanisms include: base-pair opening before base-pair plane and phosphate backbone angle rotation within the core of the helix (1); successive flipping of base-pair planes, without any disruption of the WC pairs (18); models with intermediate structure (19–23), such as one with two A-DNA-like intermediates (20); extrusion of bases, as observed in the crystal structure of a B-Z junction (23), followed by propagation and reformation of the pairs. Recent molecular dynamics (MD) simulations indicate that the transition is governed by a complex free energy landscape which allows for the coexistence of several competing mechanisms so that the transition is better described in terms of a reaction path ensemble (24).

After the discovery of Z-DNA, it was found that the right-handed A-RNA double helix made of CG repeats may also be transformed into a left-handed double helix or Z-RNA (25–29) under conditions of high ionic strength or high pressure (30). The first detailed structure of Z-RNA of natural sequence, r(CGCGCG)$_2$, was described in a nuclear magnetic resonance (NMR) study at high ionic strength in 2004 (29). However, in contrast to Z-DNA, considerably

*To whom correspondence should be addressed. Tel: +919 515 3111; Fax: +919 513 4804; Email: sagui@ncsu.edu

less is known about Z-RNA and what role it may play in terms of biological functions. Recent experiments—based on binding to the RNA-editing enzyme ADAR1—have probed the structure of Z-RNA under physiological ionic strength conditions, and provided some evidence that there may even be more than one type of Z-RNA present, either *in vitro* or *in vivo* (31).

An important structural determinant of nucleic acids is that they are polyanionic in nature, and water and counterions are crucial for their stability. In solution, counterions surround the nucleic acid structures and neutralize the nucleic acid anionic phosphates. In addition, they can establish water-mediated contacts and less frequent direct contacts with the electronegative groups. These counterions affect both the structure and stability of the nucleic acid conformations, and therefore their biological function. Specifically, counterions can help regulate genome packing (32,33), ribozyme activity (34), RNA folding (35–37), and aid in mediating DNA–protein interactions (38). In fact, due to the highly charged nature of DNA and RNA, it is unlikely that these could be packaged into their compact cellular forms in the absence of counterions (32). Ions also perform an important role in the transition between the right-handed forms at low salt concentrations, and the left-handed forms at high salt concentrations. In solution, the cations move close to the nucleic acid molecule, finding their way into the major and minor grooves, and backbone oxygens. Given the dynamical nature of the system, the cations generally become localized for relatively short periods of time before drifting away and being replaced by other cations from the solution. Similarly, the mobile anions are likely to be excluded from the near nucleic acid region due to electrostatic repulsion. However, nucleotide electropositive edges have been shown to exhibit specific anion binding sites that also turn out to be good locations for the binding of the negatively charged aspartic and glutamic amino acids and negatively charged groups of other ligands (39).

Considering the importance of the ion distribution in stabilizing nucleic acid structure, it is not surprising that this issue has received intense scrutiny over the past three decades. Although the ions are mobile, some of them can be localized long enough (especially divalent cations) to show up as bound ions in X-ray diffraction studies (40–51), although different atomic resolution crystal structures of equal sequences may differ on the presence of bound ions (52,53). Additional improvements have become possible through a combination of anomalous small-angle X-ray scattering (54–56) and atomic emission spectroscopy (57). With these techniques, it has been possible to count explicitly the number of ions in a given region, and thereby provide information as to their time-averaged distributions. NMR studies have also been used to study nucleic acid structure and surrounding ions, especially when precision is improved by the addition of residual dipolar couplings. Thus, for instance, NMR studies have found bound monovalent ions in either the major or minor groove of DNA (42,47,58–63).

However, it is still difficult to obtain true spatial resolution and dynamical information with these experimental techniques. This is where classical MD simulations are extremely useful, as they make it possible to explicitly track the motion of ions around the nucleic acids in order to quantify their locations and effect on structure. In fact, MD simulation of the ion atmosphere around nucleic acids has been around for decades (39,64–79), especially since the correct treatment of electrostatics (80–84) led to stable, reliable trajectories (81). With some exceptions, the majority of the work has focused on B-DNA. Only recently have more systematic studies on A-RNA emerged (85–89). In addition, much effort in the refinement of nucleic acid fields has taken place in the last decade (79,90–92), leading to more accurate results in the relatively long-time simulations that are required for the equilibration of the ion distribution.

In this article, we report on a large-scale MD study of the ion distribution around individual $(5'\text{-CGCGCGCGCGCG-}3')_2$ dodecamers in solution in B-DNA, A-RNA, Z-DNA and Z-RNA forms. The duplexes are immersed in rather large water boxes to account for the fact that ionic distribution functions need to be calculated to ∼30 Å. The study involves 20 simulations lasting 120 ns each. We chose this sequence because we wanted to carry out a comparative study of ion distribution, and the CG sequence is crucial for the left-handed forms. A fair comparison between the different structures needs the same sequence, as previous simulations have shown that the ion distributions exhibit sequence-dependent features. From a chemical point of view, a comparison between RNA and DNA structures only involves the presence or absence of the 2′-OH group in the sugar (and avoids the extra T/U change associated with AT sequences). The regularity of the sequence also allows for much more 'clear-cut' results and conclusions about distribution around CG pairs. Salt effects in this sequence have been studied in 2000, via 2.5-ns long simulations involving the distribution of the K^+ ion (70); and more recently in simulations that studied the effect of force fields, water model and salt concentration on the structure of A-RNA (85,87) (these studies did not report results on the ionic distribution itself). Other than that, most of the recent simulations on the ion atmosphere around right-handed nucleic acid duplexes employ sequences that are relatively rich in A and T or U nucleotides. Although *in vitro* the transition from the right-handed forms to the left-handed forms can be triggered with the addition of salt at high concentrations (29), the Z-forms also exist under physiological ionic strength (31). Simulations allow us to stabilize the left-handed forms and then to slowly increase the concentration of salt in order to discern how ion binding is linked to the structure of the duplex.

In terms of ions, we investigated the distribution of the monovalent Na^+ and K^+, and divalent Mg^{2+} ions around these structures, with various concentration values of their chloride salts. The cations most frequently found around nucleic acids are K^+ (∼0.14 M inside the cell) and Mg^{2+}, while the Na^+ ion is most frequently found in extracellular fluids (76). Since most studies of monovalent ions around nucleic acids involve the Na^+ ion, it is of interest to study the differences in binding between Na^+ and K^+. In this work, we carry out a comparative study of the distributions of each of these cations around each of the four possible duplexes and discuss in detail the sources of the various, important observed differences.

MATERIALS AND METHODS

Much effort in the refinement of nucleic acid force fields has taken place in the last decade (79), including quite recent reparameterizations in the AMBER force field (90–92). In this work, large-scale MD simulations were used to explore the ion distribution around DNA and RNA sequences $(CG)_6$ (a shorthand for (5′-CGCGCGCGCGCG-3′)$_2$) in an explicit solvent environment. The simulations were carried out using the PMEMD module of the AMBER 12 (93) software package with the ff12SB force field with parameters ff99BSC0 (90) for DNA and ff99BSC0+χ_{OL3} (87,91,92) for RNA. The TIP3P model (94) was used for the water molecules. The duplexes were placed in cubic box of ∼82 Å side, filled with a suitable number of water molecules. Such a large box is necessary, since cylindrical distribution functions (CDFs) need to be calculated to a distance of at least 30 Å. Simulation details and the equilibration process, which took longer than 1.5 ns, are given in the Supporting Information (SI) associated with this paper. The equilibration process was followed by 120 ns of constant pressure and temperature production runs. A system of oligonucleotides, water and ions takes long to reach full equilibrium, and the motion of ions is the rate-determining step for convergence (72). Indeed, for a palindromic sequence of a DNA dodecamer, it has been shown that convergence of the ion distribution in each strand (so that it reflects the symmetry of the sequence) takes ∼100 ns (73) (although internal structural parameters can take shorter times, anywhere between 10 and 50 ns (87)). Thus, equilibrium data were collected from only the last 100 ns of these runs, since a minimum of 20 ns is required in order to stabilize the ion distribution.

Five simulations were carried out for each duplex with the following ions: (i) 22 neutralizing Na^+, no excess salt; (ii) 22 neutralizing Na^+ and 0.4-M NaCl; (iii) 22 neutralizing K^+ and 0.4-M KCl; (iv) 11 neutralizing Mg^{2+} and 0.2-M $MgCl_2$; and (v) 22 neutralizing Na^+ and 4.0-M NaCl. Here, the 'salt' molarity is used to indicate 'excess' salt (over the neutralizing ions). Thus, we will refer to the case (i) above as 'zero salt', although it has 0.06-M Na^+, due to the neutralizing ions. Ions parameters are given in the SI.

Our analysis was focused on calculating the distribution of the mobile ions around the nucleic acid structures. To that end, we calculated the diffusion coefficients for the ions, the cylindrical and radial distribution functions (RDFs), and carried out an analysis of the efficacy of different sites in localizing the ions. The definitions for these quantities are standard, and their details are given in the SI section.

In terms of an analysis of a potential binding site, a Na^+ (K^+) ion was considered to be 'bound' to or localized next to a specific atom on the duplex if its distance to that atom was less than 3 Å (3.5 Å) for direct binding, or less than 6 Å when mediated by intervening water molecules. The direct binding distance corresponds to a minimum in the RDF around the electronegative nucleic acid atoms, which separates the first solvation shell from the second solvation shell. Both base and backbone sites were considered when calculating the occupancies. Occupancy was defined as the percentage of time that at least one ion was bound to a given duplex atom during the data collection time of the simulation (last 100 ns). An ion can potentially contribute to the occupancy of more than one atom in the duplex. For B-DNA and A-RNA, C(O2,N1) and G(N2,N3,N9) belong to the minor groove and C(N4) and G(O6,N7) belong to the major groove (Figure 1). The situation is less clear for the left-handed forms. For instance, in Z-DNA the minor groove is clearly discernible, while the major groove becomes flat. Here, we will use the same atomic conventions for the major and minor grooves as for the right-handed forms, as this convention has been used previously in the literature for Z-DNA (1) and for Z-RNA (29). For the backbone, we consider the phosphate oxygens, OP1 and OP2, the phosphate esteric oxygens, O5′ and O3′, the sugar-ring oxygen, O4′, and for the RNA structures the 2′-OH oxygen, O2′ (we will refer to these collectively as the O′ oxygens).

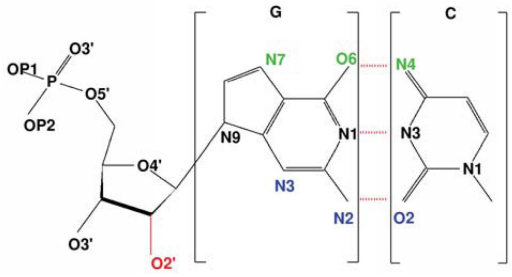


Figure 1. Atomic depiction of a CG Watson–Crick pair. Green and blue atoms represent the important major and minor groove atoms responsible for ion localization. The oxygen O2′ atom in RNA is marked in red. Hydrogen atoms are omitted for clarity.

The interaction between duplex sites and ions is, of course, a dynamical affair. Thus, an ion at a given binding site may drift away and then, after a period of time, be replaced by a new ion. Hence, to quantify how long an ion stays in a given binding site, we define a residence time τ_R by means of a standard time correlation function $C(t)_i$ for binding site i:

$$C(t)_i = \sum_{t_o} \sum_{t} p_i(t_o) p_i(t_o + t),$$

where $p_i(t)$ is unity if the site is occupied by an ion within 3 Å for Na^+ and 3.5 Å for K^+, and zero otherwise. Typically, this function takes the form of a decaying exponential and so $C(t)_i \sim \exp(-t/\tau_R)$ which in turn defines the residence time τ_R as a time constant. In our simulations, $C(t)_i$ were measured over time intervals of 10 ps to 1 ns, with an increasing step of 10 ps.

RESULTS

Cylindrical distribution functions

Figures 2 and 3 illustrate the CDFs for the different duplex structures at low and high salt concentrations, respectively. Figure 2 shows convergence of the CDFs, which occurs around 20 ns in the left-handed forms and after 40 ns in the right-handed forms. At zero-salt concentration (with only neutralizing Na^+; Figure 2), the peaks associated with the RNA structures are higher than those for the corre-

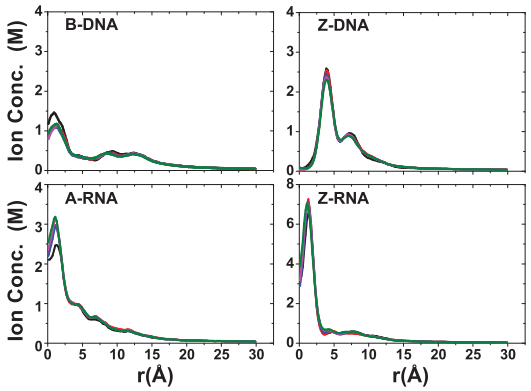


Figure 2. CDFs for Na^+ ions at zero salt. These plots also illustrate the convergence of these CDFs over time with different colors indicating ever increasing time intervals: black (20–30 ns), red (20–40 ns), blue (20–50 ns), magenta (20–60 ns) and green (20–70 ns).

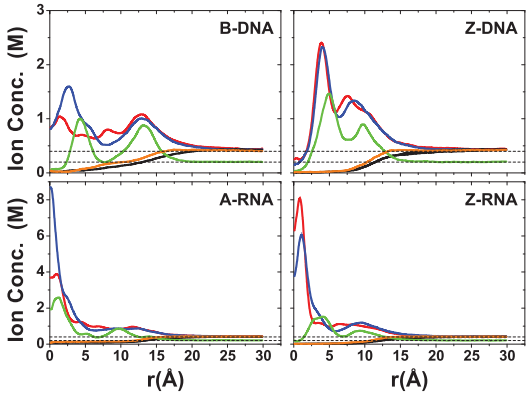


Figure 3. Converged CDFs at high salt concentrations. CDFs are shown for the three cations: Na^+ (red), K^+ (blue), Mg^{2+} (green); and negative ions: Cl^- (black) for NaCl and KCl, and Cl^- (orange) for $MgCl_2$. Results are under conditions of high salt concentration: 0.4 M for NaCl/KCl and 0.2 M for $MgCl_2$ (plus the corresponding neutralizing cations). Results for Cl^- ions for NaCl/KCl are virtually identical, and so these are not color differentiated.

sponding DNA structures, indicating that the number of ions localized around the RNA structures is larger than that for the DNA counterparts. In addition, the peaks associated with the left-handed structures are higher than those for the corresponding right-handed structures, pointing to higher localization of ions around the left-handed structures. The fall-off from the first peak differs considerably between the structures. For B-DNA, this is characterized by a set of minima and maxima between 7 and 14 Å, followed by a smooth decay to approximately zero (<0.05 M, which is the concentration of 22 Na^+ ions in the given box) at ∼28 Å. By contrast, Z-DNA has a single secondary maximum between 6 and 8 Å followed by a smooth decay to ∼0 at ∼23 Å. For A-RNA, there are a few, shallow oscillations after the initial peak, and it takes ∼26 Å to decay to zero. Likewise, a set of very shallow minima and maxima characterizes the Z-RNA fall-off, followed by a smooth decay to zero at ∼25 Å.

Turning to the results for high salt (22 Na^+ ions + 0.4-M NaCl; 22 K^+ ions + 0.4-M KCl; 11 Mg^{2+} ions + 0.2-M $MgCl_2$) in Figure 3, B-DNA now has a number of well discernible peaks associated with each of the three cations, Na^+, K^+ and Mg^{2+}. By comparison to the zero-salt distribution of Na^+ ions in Figure 2, there are now several lower peaks for Na^+ before the distribution smoothly decays to the bulk concentration value. The distributions for the three cations around B-DNA indicate the presence of two equally important 'binding shells' (more diffuse in the case of Na^+). The second binding shell is also present in Z-DNA, although its peak is considerably lower than the first peak. Distributions for Na^+ and K^+ ions are similar for Z-DNA, but the first peak is much more prominent for K^+ than Na^+ ions in B-DNA and A-RNA, while the first Na^+ peak is higher than the K^+ one for Z-RNA. In the RNA duplexes, the ions can become very close to the central axis. The peaks associated with the Mg^{2+} ions are shifted outward, in comparison to the peaks associated with the monovalent cations (except for A-RNA where the peaks for Na^+ and Mg^{2+} ions are centered at approximately the same distance).

The total ionic charge (both co-ions and counterions) accumulated as a function of radial distance from the helical axis is shown in Figure 4. The charge is normalized with respect to the total charge of the nucleic acid duplex (−22e). For any given distance before the asymptotic value, A-RNA is more screened than B-DNA. For regions close to the axis, A-RNA localizes more Na^+ and K^+ ions than both B-DNA and Z-DNA, and more Mg^{2+} ions than the three other forms. At intermediate distances, charge neutralization works better in the left-handed forms (which are comparable). The Mg^{2+} ion distribution is considerably more localized than the Na^+ ion distribution, which in turn is slightly more localized than the K^+ ion distribution.

Radial distribution functions

Figures 5–9 display RDFs of the Na^+ and Mg^{2+} ions with respect to different DNA/RNA atoms. Within the major groove, there are qualitative similarities for Na^+ binding between B-DNA and A-RNA in Figure 5. The first peak of the electronegative O6 atom occurs at 2.4 Å for both B-DNA and A-RNA. The bulky, hydrated Mg^{2+} (Figure 6) is displaced outward with binding distances with respect to O6 of 4.1 Å (B-DNA) and 4.3 Å (A-RNA). For the left-handed forms, differences in Na^+ binding are considerable. In particular, Z-RNA displays far stronger binding than Z-DNA, characterized by very strong O2 binding (at 2.3 Å), while Z-DNA has relatively little direct binding to O2 at 2.4 Å and a second more important, indirect binding peak at 4.6 Å. The distributions of Mg^{2+} for the Z forms seem to indicate mainly indirect binding and are qualitatively more similar.

Figure 7–9 show the RDFs between the ions and the oxygen atoms in the backbone. Figure 7 for the Na^+ ion shows a qualitative similarity between the B-DNA and A-RNA binding, with additional binding by the O2′ atom in RNA. The left-handed forms, on the other hand, show striking

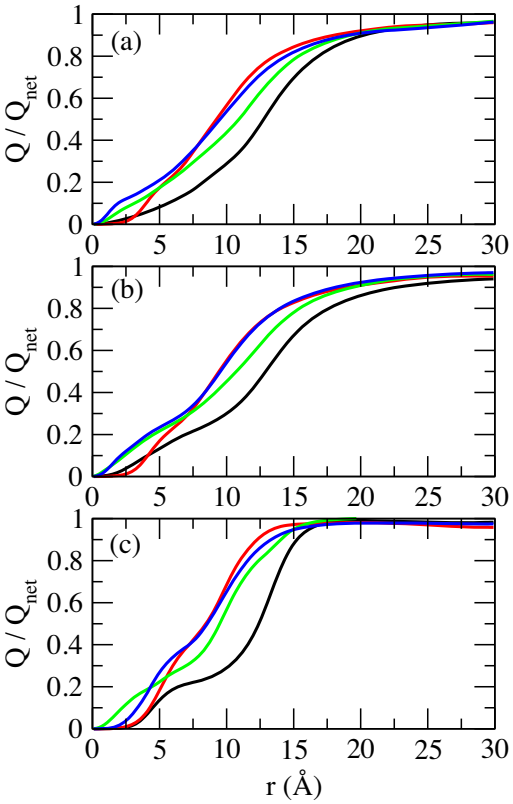


Figure 4. Accumulation of total ionic charge (counterions + co-ions) as a function of radial distance from the central axis of the duplex. The values are normalized with respect to the net charge of the duplex sequence (−22e). Different colors represent different structures: B-DNA (black); Z-DNA (red); A-RNA (green); Z-RNA (blue). (**a**) Na^+ at 0.4-M NaCl; (**b**) K^+ at 0.4-M KCl; (**c**) Mg^{2+} at 0.2-M $MgCl_2$.

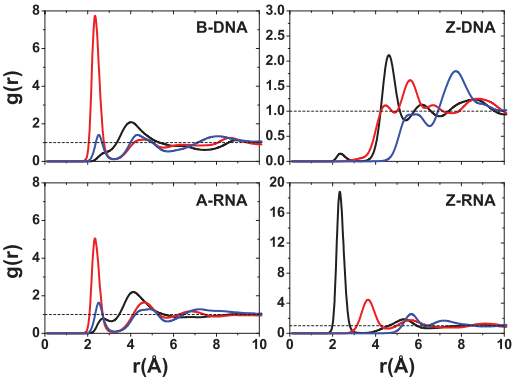


Figure 5. RDFs for Na^+ with respect to major or minor groove atoms at 0.4-M salt concentration. For B-DNA and A-RNA, the colors indicate RDFs with respect to atoms in the major groove: O6 (red), N7 (blue) and N4 (black). For Z-DNA and Z-RNA, the colors indicate RDFs with respect to atoms in the minor groove: O2 (black), N2 (red) and N3 (blue).

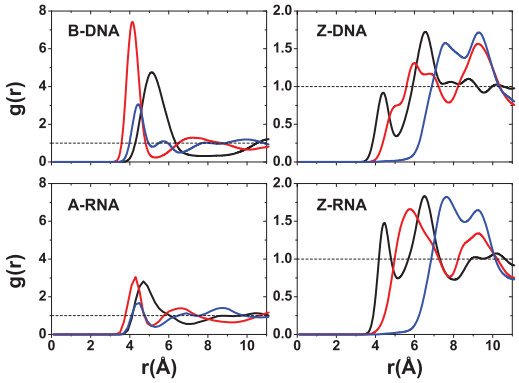


Figure 6. RDFs for Mg^{2+} with respect to major or minor groove atoms at 0.2-M salt concentration. For B-DNA and A-RNA, the colors indicate RDFs with respect to atoms in the major groove: O6 (red), N7 (blue) and N4 (black). For Z-DNA and Z-RNA, the colors indicate RDFs with respect to atoms in the minor groove: O2 (black), N2 (red) and N3 (blue).

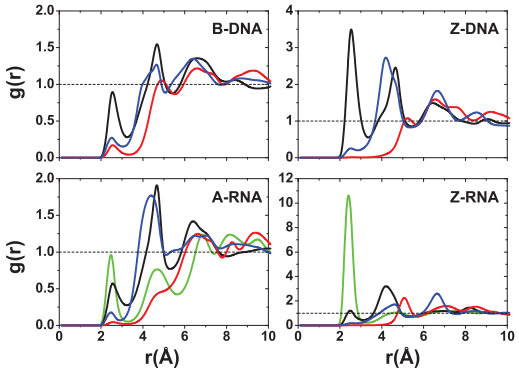


Figure 7. RDFs for Na^+ with respect to O′ backbone oxygens at 0.4-M salt concentration. Colors indicate: O2′ (green), O3′ (black), O4′ (red) and O5′ (blue).

differences. Z-DNA has a strong peak for O3′ at 2.6 Å and important secondary binding at 4.7 Å for O3′ and at 4.2 Å for O5′. Z-RNA has important indirect binding for O3′ at 4.2 Å but in addition it presents a dominant peak for O2′ at 2.4 Å. For Mg^{2+}, the distributions are relatively more similar (with stronger binding for the RNA forms; Figure 8) except, of course, for the binding to O2′ in the RNA forms. The RDFs for binding with the phosphate oxygens (OP1,OP2) are shown in Figure 9. For a given ion, the RDFs for the different structures all resemble each other closely, with stronger binding in the Z forms. For Na^+, OP1 binding is preferred over OP2, especially in the Z forms. Direct binding of Na^+ to the four duplexes takes place at ∼2.3 Å, and a weaker, secondary binding occurs at 4.5 Å. For Mg^{2+} ions, the binding is all indirect (alternatively, it is direct binding of the hexahydrated $Mg[(H_2O)_6]^{2+}$ ion). The RDFs for K^+ ions are given in the SI section. In general,

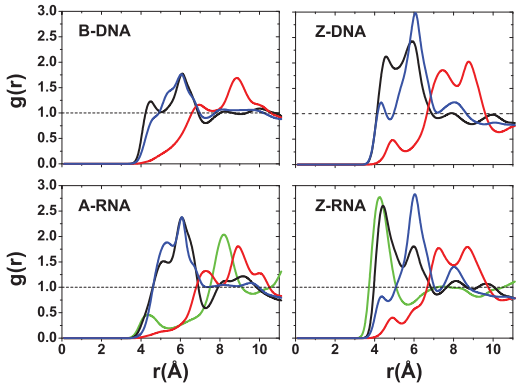


Figure 8. RDFs for Mg^{2+} with respect to O′ backbone oxygens at 0.2-M salt concentration. Colors indicate: O2′ (green), O3′ (black), O4′ (red) and O5′ (blue).

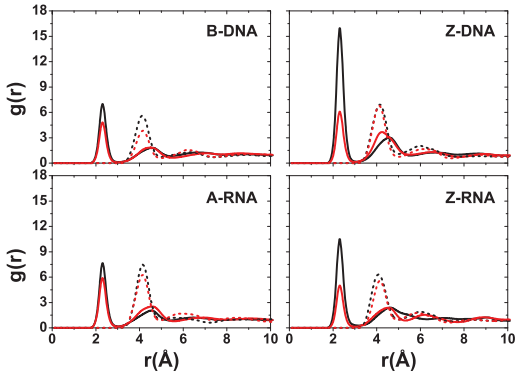


Figure 9. RDFs for Na^{+} (solid lines, 0.4-M salt concentration) and Mg^{2+} (dashed lines, 0.2-M salt concentration) with respect to phosphate oxygens. Colors indicate: OP1 (black) and OP2 (red).

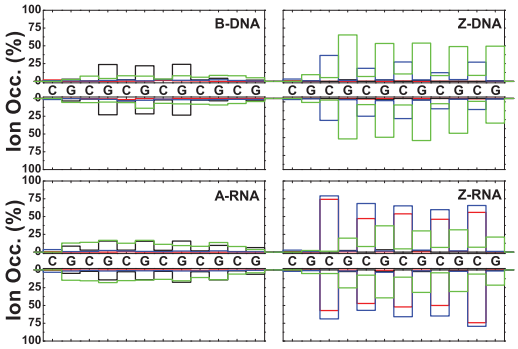


Figure 10. Na^{+} ion occupancies within 3 Å as a function of sequence for zero salt concentration. Colors represent: major groove (black), minor groove (red), O′ oxygen atoms on backbone (blue) and phosphate oxygens (green).

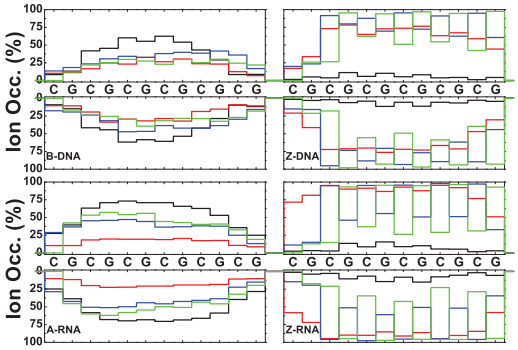


Figure 11. Na^{+} ion occupancies within 6 Å as a function of sequence for zero salt concentration. Colors represent: major groove (black), minor groove (red), O′ oxygen atoms on backbone (blue) and phosphate oxygens (green).

K^{+} seems to be more tightly bound than Na^{+} in the right-handed forms, and comparably in the Z forms (except for a large peak for binding of Na^{+} to O2 in Z-RNA).

Ion binding as a function of sequence

In addition to the RDFs, we have investigated ion binding to specific DNA/RNA atoms in a number of ways. Figures 10 and 11 plot the ion occupancy of each nucleotide in the sequence: the top part describes the 5′–3′ strand from left to right, while the lower part gives the occupancy of the other strand in reverse order (i.e. also in the 5′–3′ from left to right). This way of displaying the information should show mirror symmetry with respect to the horizontal axis upon ion distribution convergence. Figure 10 plots the occupancy results for Na^{+} ions close to the nucleotides (i.e. within 3 Å) for zero salt concentration (just neutralizing Na^{+} ions). Figure 11 shows similar results for ion occupancies within 6 Å, and therefore includes not only the so-called direct bindings but also the Na^{+} ion bindings mediated by approximately a single water molecule. In terms of the direct binding results, Figure 10 shows that for B-DNA most of the ion binding with ∼25% occupancy is associated with three central G nucleotides of the major groove, while the minor groove and backbone O binding are either very small or negligible. For A-RNA, the major groove binding becomes distributed among all the G nucleotides with occupancies of ∼15%; and the binding with the phosphate oxygens, in the 10–15% range, is distributed more or less uniformly along the backbone. By contrast, the phosphate oxygen binding in Z-DNA is strongly centered on the Gs with populations around 50–60%; binding to the O′ oxygens is also important, in the range 20–40%, and centered on the Cs; and binding to the major and minor grooves is negligible. Z-RNA also presents binding to the phosphate oxygens on the Gs and to the O′ oxygens on the Cs, but their relative importance is reversed compared to Z-DNA: with ranges 20–40% for the phosphate oxygens and ranges 60–80% for the O′ oxygens. Binding to the minor groove along

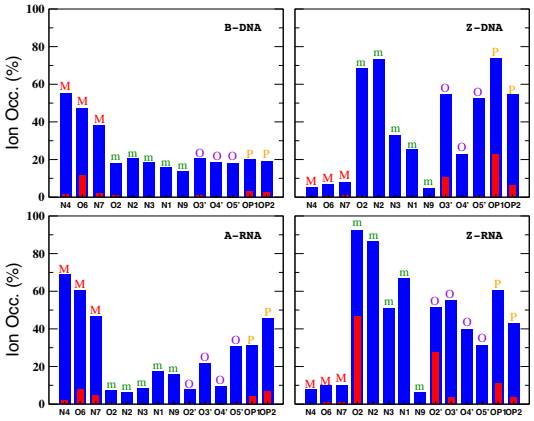


Figure 12. Na^+ ion occupancies with respect to specific atoms at zero salt concentration. Occupations for direct and indirect binding are indicated with red (3 Å) and blue (6 Å). The results represent average values with respect to the entire duplex. Letters on top of the bars represent different nucleic acid regions: M (major groove), m (minor groove), O (O′ oxygen atoms on backbone) and P (phosphate oxygens).

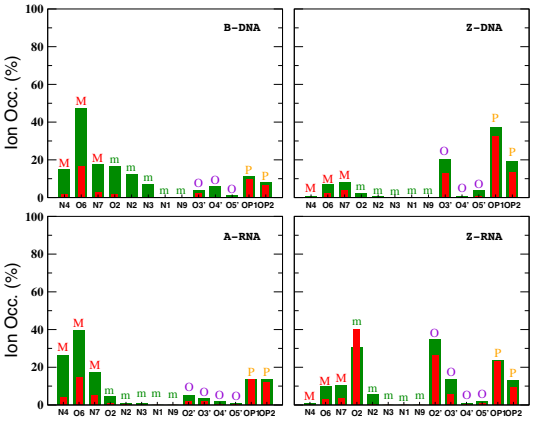


Figure 13. Ion occupancies with respect to specific atoms at 0.4-M salt concentration within the direct binding region. Colors represent ion type: Na^+ (red) and K^+ (green). The results represent average values with respect to the entire duplex. Letters on top of the bars represent different nucleic acid regions: M (major groove), m (minor groove), O (O′ oxygen atoms on backbone) and P (phosphate oxygens).

the Cs in Z-RNA is also very important, with ranges 50–70%. The net effect of including ion binding all the way to 6 Å is illustrated in Figure 11. Naturally, the occupancy becomes much larger and more uniform. Results for the high salt concentration (NaCl and KCl) and Mg^{2+} ions are presented in the SI section.

In addition to the different distribution functions already discussed, it is possible to further characterize the localization of ions due to particular atoms in the nucleic acid duplexes. Figure 12 gives bar plots for the Na^+ ion occupancies for direct (≤3 Å) and indirect (≤6 Å) binding with respect to specific atoms at zero salt. In agreement with the CDFs and RDFs presented before, there is similarity between the right-handed structures for the Na^+ ions, with the O6 atom contributing more to direct binding (the OP oxygens have a comparable contribution for A-RNA). The left-handed structures, on the other hand, differ considerably with respect to each other and with respect to their right-handed counterparts. Figure 12 shows that most direct binding in Z-DNA occurs at OP1 with secondary contributions from OP2 and O3′. By contrast, in Z-RNA there is a big contribution to binding by O2 (see also Figure 5), followed by O2′ and, to a lesser degree, the phosphate oxygens and O3′. Figure 13 compares direct binding for K^+ and Na^+ (for 0.4-M salt concentration). In general, K^+ exhibits larger ion occupancy, especially in the major groove for the right-handed forms (and part of the minor groove for B-DNA). The only exception to this observation occurs at atom O2 in Z-RNA, where Na^+ exhibits higher ion occupancy. Comparing direct binding for Na^+ in Figure 12 (zero salt) and Figure 13 (0.4-M salt concentration), one can see that there are only relatively small changes in the ion occupancies, indicating that these sites are primarily saturated at low concentrations.

Anion distribution

The aim of this work is to characterize the cation distribution around the four types of duplexes. However, we also checked whether the Cl^- anions had contacts with the nucleic acids. We found that, indeed, there were short-lived contacts. For instance, a contact close to C-N4 was seen in all four duplexes, but its occupation was extremely low. In addition, we found low-population bindings to the O2′ sugar groups for the RNA duplexes, and to G-N2 in B-DNA and A-RNA. So indeed, it is possible for the anion to intrude the first hydration shell of nucleic acids (39), but these events are too rare to be statistically significant. The CDFs in Figure 3 show that the behavior of the Cl^- anion distribution is similar in all cases: it is close to zero next to the helical axis with a smooth increase out to its bulk value. For the DNA structures, values of the Cl^- distribution start to rise before 10 Å, while for the RNA structures the values start to rise only after 10 Å: Cl^- anions generally are not allowed near the duplexes, and this repulsion is stronger for the RNA structures.

Structural interpretation of ion occupation

Table 1,3 gives the residence times for the three ions, and Table 2 gives the average distances of the monovalent ions to the nucleotide atoms. Figure 14 illustrates some typical binding pockets for the Na^+ ion in the different nucleic acid structures. Figure 14a and b gives a snapshot of the binding of the Na^+ ion by O6 in B-DNA and A-RNA. The ion is directly bound to the O6 atoms and also close to the N4 atoms of the major groove. This binding pocket is shown in more detail in Figure S8, where the hydrogen atoms covalently linked to the N4 atoms and four bound waters are depicted explicitly. In this conformation, the Na^+ ion is strongly bound by the G-O6 atoms, and given the geometry

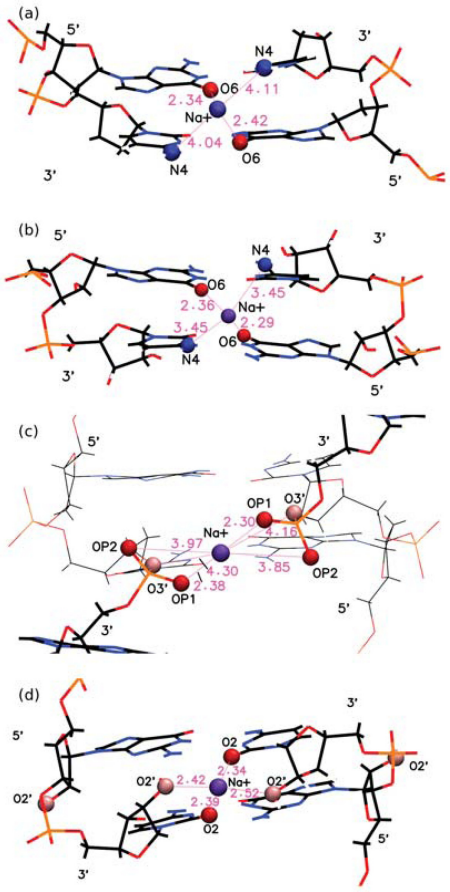


Figure 14. Atomistic details of some direct binding sites of Na^+ for different nucleic acid structures. (**a**) Binding to G-O6 (and proximity to C-N4) in B-DNA; (**b**) binding to G-O6 (and proximity to C-N4) in A-RNA; (**c**) binding to G-OP1, G-OP2 and C-O3′ in Z-DNA; (**d**) binding to C-O2 and C-O2′ in Z-RNA. The configurations shown are just a snapshot at a given time obtained from the MD simulations. Details of (a) are also provided in Figure S8.

of the GpC steps, the ion manages to position itself relatively close to the N4 atoms by avoiding the electropositive N4 hydrogen atoms, which point away from the ion. In the process, the ion loses two of the waters in its binding shell. In A-RNA, the distance to the N4 atoms decreases slightly, as shown in Table 2. This is particularly true for K^+, which shows a residence time of 2.4 ns for N4. In the four cases (B-DNA/A-RNA combined with Na^+/K^+), the configurations can be described as ion bridges, without intervening waters between the ions and the O6 and N4 atoms. In addition, K^+ exhibits long residence times for O2 and N2 in the minor groove of B-DNA. Figure 14c shows the direct binding of Na^+ to the phosphate oxygens in a CpG step in Z-DNA and Figure 14d shows the binding of Na^+ to O2 and O2′ in Z-RNA. In Z-DNA, the binding of Na^+ to G-OP1 is direct, with the longest residence time (∼2.3 ns) and a short binding distance of ∼2.4 Å. Instead, the OP1 in the C nucleotide points outside the helical core in Z-DNA, and therefore its residence time is only ∼0.43 ns. With respect to the OP2 atoms in Z-DNA, the residence time of Na^+ is ∼0.35 ns in C-OP2 and 0.65 ns in G-OP2. For the Z-RNA conformation shown in Figure 14d, the cation is localized by four atoms: two each of O2 and O2′ all situated in the C nucleotides. These atoms are relatively close to the cation, with distances in the 2.3–2.5 Å range. This illustrates how Z-RNA is much more efficient in localizing or trapping cations, and is reflected in the residence times given in Table 1. The times associated with Z-RNA O2 in C and O2′ in C are very high: 8.6 ns and 7.8 ns, respectively. These residence times are so much longer than the times associated with other binding sites that ions making their way into these positions in Z-RNA are effectively trapped there for a very long time. We have also examined the characteristics of the OP1 and OP2 binding for Z-RNA, which qualitatively resemble those of Z-DNA.

Residence times associated with K^+ are considerably longer than those for Na^+, except for Z-RNA where Na^+ has longer residence time for the O2 and O2′ in C. The binding distances for K^+ tend to be longer than those for Na^+, reflecting the larger size of the K^+ ion (for which we considered distances ≤3.5 Å as direct binding). An exception to this is seen in cytosine O3′ in Z-DNA which displays direct binding for K^+ (residence time of 2.1 ns) but not for Na^+. Smaller fluctuations in the binding distances shown in Table 2 correlate with stronger, direct binding, while larger fluctuations correlate with higher mobility and indirect binding.

Table 3 also gives the characteristic residence times for the hexahydrated $Mg[(H_2O)_6]^{2+}$ ions. Figure 15a shows $Mg[(H_2O)_6]^{2+}$ localization in the major groove of B-DNA, with the hexahydrated ion bound to G-O6 (with a very large residence time of 9.4 ns) and G-N7 (and also in proximity of C-N4) in the GpC steps. The same binding can be found in A-RNA (see Figure S9a), with residence times of 2.4 ns (O6) and 2.2 ns (N4). These two cases are quite similar to those described for Na^+ in Figure 14a and b and Figure S8. In addition, the ion can display long-time binding to the phosphate oxygens in A-RNA with equivalent residence times (2.1–2.3 ns). This occurs, for instance, in the bridge between distant phosphate groups as shown in Figure 15b and c. The latter results in high bending of the duplex, and in this way the $Mg[(H_2O)_6]^{2+}$ ion completely closes access of other ions to the middle of the duplex. Z-DNA shows strong localization at the phosphate oxygens, with residence times of 7 ns (G-OP1) and 5.6 ns (G-OP2). This can be seen in Figure S9b, where the OP oxygens are pointing outward, away from the core of the helix, and in Figure S9c, where the ion is in the minor groove of Z-DNA. Z-RNA displays several binding sites: G-OP1 (7.6 ns), G-OP2 (5 ns), C-O2 (1.9 ns), C-O2′ (6 ns) and C-O3′ (4.9 ns). Figure 15d shows binding in the minor groove of Z-RNA.

Table 1. Residence time (in ns) of Na^+ and K^+ for different atoms in the nucleic acid structures

	$Na^+(K^+)$ residence times (ns)									
B-DNA	N4 0.83(1.74)	O2 –(1.82)	O6 1.01(2.03)	N2 –(1.84)	N3 –(1.48)	N7 0.25(1.37)	O3′ 0.38(0.41)	O4′ –(1.74)	OP1 0.39(0.86)	OP2 0.35(0.80)
Z-DNA	O2 –(0.30)	O6 –(0.62)	N7 –(0.68)	O3′(C) 1.25(2.12)	O3′(G) –(0.51)	OP1(C) 0.43(1.18)	OP1(G) 2.26(4.75)	OP2(C) 0.35(0.86)	OP2(G) 0.65(1.84)	
A-RNA	N4 0.43(2.45)	O2 –(0.38)	O6 0.54(1.47)	N7 0.35(0.85)	O2′ –(0.52)	OP1 0.42(1.58)	OP2 0.47(1.56)			
Z-RNA	O2 7.83(2.79)	O6 –(0.91)	O2′(C) 8.56(5.42)	O2′(G) –(0.71)	O3′(C) 1.10(1.43)	O3′(G) –(0.67)	OP1(C) 0.43(0.99)	OP1(G) 1.21(1.92)	OP2(C) 0.36(0.76)	OP2(G) 0.60(1.82)

These results are for 0.4-M salt concentration.

Table 2. Distance from ions to specific nucleic acid atoms for direct binding

	$Na^+(K^+)$ distances (Å)					
B-DNA	O6(G6) $2.38 \pm 0.13(2.78 \pm 0.17)$	O6(G18) $2.40 \pm 0.14(2.80 \pm 0.19)$	N4(C7) $3.72 \pm 0.50(3.87 \pm 0.56)$	N4(C19) $3.58 \pm 0.54(4.06 \pm 0.61)$		
Z-DNA	OP1(G8) $2.39 \pm 0.15(2.82 \pm 0.20)$	OP1(G20) $2.36 \pm 0.13(2.80 \pm 0.19)$	OP2(G8) $3.86 \pm 0.36(4.11 \pm 0.39)$	OP2(G20) $4.12 \pm 0.30(4.13 \pm 0.41)$	O3′(C7) $3.96 \pm 0.26(3.39 \pm 0.54)$	O3′(C19) $3.97 \pm 0.27(3.37 \pm 0.52)$
A-RNA	O6(G6) $2.36 \pm 0.11(2.79 \pm 0.17)$	O6(G18) $2.39 \pm 0.14(2.78 \pm 0.18)$	N4(C7) $3.13 \pm 0.52(3.54 \pm 0.36)$	N4(C19) $3.59 \pm 0.38(3.66 \pm 0.40)$		
Z-RNA	O2(C7) $2.39 \pm 0.13(2.83 \pm 0.21)$	O2(C19) $2.39 \pm 0.12(2.79 \pm 0.18)$	O2′(C7) $2.43 \pm 0.13(2.80 \pm 0.15)$	O2′(C19) $2.44 \pm 0.13(2.77 \pm 0.14)$		

The labels in the parenthesis refer to the residues to which the atoms belong. Usually four to six atoms act to form a single ion trap in the middle part of a duplex, as shown in Figure 14. The distances are averages and are based on the last 10 ns of MD simulations.

Table 3. Residence time (in ns) of Mg^{2+} for different atoms in the nucleic acid structures

	Mg^{2+} residence times (ns)													
B-DNA	N4 2.04	O6 9.43	N7 2.21	O3′ 0.79	O5′ 0.52	OP1 1.20	OP2 0.93							
Z-DNA	O2 1.08	O6 0.57	N2 1.04	N7 0.59	O3′(C) 2.61	O3′(G) 0.56	O5′(C) 1.37	O5′(G) 1.15	OP1(C) 2.15	OP1(G) 7.00	OP2(C) 1.24	OP2(G) 5.62		
A-RNA	N4 2.20	O6 2.42	N7 1.13	O2′ 0.42	O3′ 0.75	O5′ 0.82	OP1 2.25	OP2 2.10						
Z-RNA	N4 0.71	O2 1.89	O6 0.90	N2 1.02	N7 0.96	O2′(C) 6.04	O2′(G) 0.57	O3′(C) 4.93	O3′(G) 0.48	O5′(C) 0.96	O5′(G) 0.95	OP1(C) 1.73	OP1(G) 7.61	OP2(C) 0.85	OP2(G) 4.99

These results are for 0.2-M salt concentration, using a cutoff of 5 Å.

Nucleic acid structure as a function of salt concentration

We have checked whether the structure of the duplexes changes as a function of salt concentration. We consider the case for Na^+ ion with 0-M, 0.4-M and 4.0-M excess NaCl salt. All the structures show typical alternating, periodic features in the various parameters that reflect the regular, alternating sequence pattern. Once the left-handed structures are equilibrated, they show no measurable sensitivity (within the statistical errors) to the salt concentration. B-DNA shows no sensitivity up to 0.4-M NaCl. However, for 4.0 M, some localized changes at base pairs 7–9 can probably be attributed to the high salt concentration. In particular, the sugar pucker of base G8 in B-DNA changes from its predominantly C2′-endo conformation to C1′-endo and some C2′-exo (Figure S10a). This affects other parameters around this base, such as the step twist shown in Figure S10b and the glycosidic angle, that changes from -112° at 0 M and 0.4 M to -58° at 4.0 M, effectively switching from an *anti* conformation to a 'syn' conformation. These local changes may indicate the onset of an instability due to the high salt concentration. On the other hand, A-RNA is the structure that shows most sensitivity to salt concentration globally. This is quite apparent in helical and step parameters. Two such examples are given in Figure S10c and d. Although the figures shown in Figure S10 are only averaged over the last 10 ns of the simulations, they give a good idea of the general trends. Inclination, step roll, helical and step twist and x-displacement all increase with salt concentration while helical and step rise, and propeller decrease with salt concentration. In other words, the structure becomes more compact and more A-like, an observation that has been reported before (85,87). For instance, we measure the average helical inclination and roll step parameter (during the last 10 ns) as $\simeq 10^\circ$ and $\simeq 6^\circ$, respectively, for zero excess salt, and $\sim 17^\circ$ and $\simeq 10^\circ$, respectively, for 4.0-M NaCl.

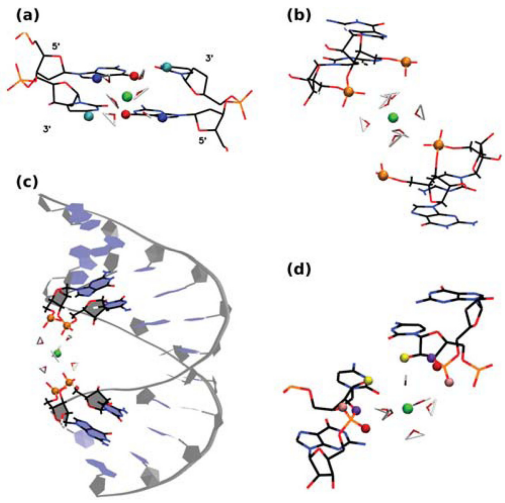


Figure 15. Atomistic details of binding sites of hexahydrated Mg^{2+} (green) for different nucleic acid structures. (**a**) Binding to G-O6 (red) and G-N7 (blue) (C-N4 is shown in cyan) in B-DNA; (**b**) binding to phosphate oxygens (orange) in A-RNA; (**c**) overall view of (b) along the duplex; (**d**) binding to G-OP1 (red), G-OP2 (pink), C-O2′ (yellow) and C-O3′ (violet) in Z-RNA in a similar arrangement as that shown in Figure 14c. The configurations shown are just a snapshot at a given time obtained from the MD simulations.

DISCUSSION AND CONCLUSIONS

Role of the ion type

With MD simulations, we have calculated the distribution of the K^+, Na^+ and Mg^{2+} cations (and the Cl^- anions) around right-handed B-DNA and A-RNA, and left-handed Z-DNA and Z-RNA. As previously pointed out (76), differences in binding between the Na^+ and K^+ cations are important because while K^+ cations dominate in the intracellular fluids (at ∼0.14 M), most studies have been carried out with the Na^+ cation, which is mainly present in extracellular fluids. In fact, K^+ cations are important for activating RNA systems, and for ribosome structure that can otherwise unfold in their absence (76,95,96). In solution, positive cations attract water molecules that form a 'solvation shell' around them. Large angle X-ray scattering and double difference infrared spectroscopy have determined (97) the bond distance between the cation and the first-shell water oxygen as 2.43 Å for Na^+ and 2.81 Å for K^+. This bond distance in Mg^{2+} is ∼2.00–2.15 Å (98). For the same average distance, we measure 2.44 Å for Na^+, 2.89 Å for K^+ and 2.00 Å for Mg^{2+}, which provides for a good validation of the ion parameters being used.

The cations are naturally attracted to the electronegative sites around the nucleic acid duplexes, and both direct and water-mediated binding are seen. Many of the results can be understood by the fact that the ionic radius of each K^+, Na^+ and Mg^{2+} decreases in that order, while the strength of the first solvation shell around a metal cation decreases with its radius and increases with its charge. Thus, Mg^{2+} maintains its first solvation shell during the entire length of the simulations, and effectively acts as a hexahydrated cation, $Mg[(H_2O)_6]^{2+}$. On the other hand, both monovalent cations can become partially dehydrated, but K^+ (having a larger radius than Na^+) can shed some of its first solvation waters with more ease and therefore penetrate deeper into smaller nucleic acid pockets. In fact, the hydration enthalpy of K^+ has been measured experimentally to be ∼17 kcal/mol higher than that of Na^+ (99,100). The Na^+ ion can crystallize with full hydration shell with large low-symmetry counterions, while the inability of K^+ to form well-defined hydrated structures in the solid state is a sign of its weak hydration (97).

Divalent ions are known to play an important role in the folding of biomolecules, especially RNA systems, and MD simulations have contributed to elucidate the role of cations in oligonucleotide systems. (35,57,101–107,102–107). In particular, Mg^{2+} cations are the most important divalent ions for the formation of RNA structures and their functional role in the cell. The fact that Mg^{2+} is very resistant to dehydration is reflected in crystallography, where Mg^{2+} in high-resolution crystal structures were observed to be fully hydrated (40,49). It has been seen that in crystallography, where concentrations are higher than physiological levels, $Mg[(H_2O)_6]^{2+}$ can bind to sites that they would not normally do under physiological conditions, where they would wander off and be replaced by partially dehydrated monovalent cations (76). From the point of view of the simulations, it is known that accurate descriptions of divalent ions such as Mg^{2+} require the use of polarizable force fields (108), and the first simulations of DNA in explicit solvent with a fully polarizable force field (109–111) (applied to all atoms) showed that even a very simple representation of the atomic polarizability can improve the accuracy of the simulations. In spite of this, the parameters used in the present simulations (see SI) are known to give good qualitative insights into ion binding sites.

We have also looked at the distribution of the Cl^- anions and found that there were short-lived contacts, including contacts close to C-N4 in all four duplexes, to the O2′ sugar groups for the RNA duplexes, and to G-N2 in B-DNA and A-RNA, which confirms that it is possible for the anion to intrude upon the first hydration shell of nucleic acids (39). However, these events are quite rare and are not statistically significant. The CDFs (Figure 3) show that Cl^- anions generally are not allowed near the duplexes, particularly in the RNA structures. It is, however, possible that our simulations underestimate the contacts due to the absence of polarization. Studies on ion interface solvation (112–117) show that halide anions (all but fluoride) prefer surface solvation versus bulk solvation, and thus—in absence of other competing forces, such as those due to the presence of other ions—the halide ion would naturally migrate, for instance, to a biomolecule–water interface. The driving force for surface solvation is the polarizability of water. Thus, the inclusion of polarization in the force field could result in larger probabilities for closer contacts between co-ions and nucleic acids.

We have also tracked the diffusion constants for the different ions in the presence of different duplexes, and the results are presented in the SI.

Comparison between different ions for B-DNA and A-RNA

Results for our sequence, $(5'\text{-CGCGCGCGCGCG-}3')_2$, show both similarities to and differences from other recent extensive simulations of ion distributions (74,73,78,86); the differences mainly due to the fact that the sequences employed in these other simulations are relatively rich in A and T or U nucleotides.

At zero excess salt, most Na^+ direct binding to B-DNA occurs in the major groove, especially in the three central Gs in the GpC steps. For A-RNA, the Na^+ direct binding is spread more evenly among all the Gs in the GpC steps of the major groove and is comparable to the binding to phosphate oxygens in the backbone. Direct binding to the major groove is saturated so that when the excess salt content increases to 0.4 M, direct binding to the major groove barely increases (direct binding to the phosphates increases slightly). For both duplexes and both salt concentrations, direct binding of Na^+ in the minor groove is negligible. By contrast, more favorable binding of Na^+ in minor grooves has been found in A-tract DNA or ApT steps both in NMR studies (61,62) and simulations (73,78). Naturally, the population of water-mediated Na^+ contacts at 6 Å is much larger for each nucleotide than that for direct binding. Still, the occupation of the major groove in these water-mediated contacts is larger than the occupation of the minor groove, especially for A-RNA. For K^+, direct binding is qualitatively similar to that of Na^+, except that population in the major groove associated with the Gs in GpC steps is considerably increased, and there is a non-negligible population in the minor groove for B-DNA. On the other hand, minor groove occupation is almost non-existent for A-RNA (Figure 13). Preferred binding of K^+ in the major groove has been observed experimentally (47,63).

Other general differences between B-DNA and A-RNA can be observed in the CDFs and RDFs presented in Figures 2,3,5–8. In B-DNA water-mediated binding of cations is more evenly distributed: while the majority resides in the major groove with occupancy varying between 40 and 56% for Na^+ (Figure 12), the minor groove and backbone oxygens have non-negligible, comparable occupations around 20%. In A-RNA even water-mediated cations are more localized: occupancy in the major groove reaches up to 70%, it is very small for the minor groove, and rises again for the phosphate oxygens (Figures 12 and 13). Thus A-RNA exhibits a stronger localization of counterions, which is reflected in the higher peaks and faster decay in the CDFs in Figures 2 and 3. In A-RNA most of the cation density lies in the major groove along the cylindrical axis of the helix, a consequence of the structure of the helix whose base-pair planes are not perpendicular to the helix axis (as in B-DNA) but tilted toward the axis, thus leaving the core of the helix more exposed. This penetration of the ions into the open core of the helix is 'complete' for K^+ at high salt (Figure 3), but even $Mg[(H_2O)_6]^{2+}$ is very close to the helix axis (here calculated as the global z-axis with 3DNA, as in SCHNAaP (118), defined by vectors that are a combination of C1′ and G-N9/C-N4 atoms along the same strand, as developed by Rosenberg *et al.* (119)). Interestingly, a comparison with the Na^+ CDFs for A-RNA for a different sequence with high content of A and U nucleotides (86) shows important sequence-dependent effects. For instance, the CDF for Na^+ at 0.4-M NaCl in Figure 3, (86), shows several 'binding regions' with the maximum peak at ∼10 Å. For our CG sequence, the peak near the origin completely dominates the distribution of Na^+ (an effect that is even larger for K^+). This can be understood in terms of the perfect regularity of the pure CG sequence where the GpC steps strongly encourage binding to the major groove.

Particularly noticeable in our systems is the direct binding of the monovalent ions by G-O6 in the GpC steps, which has been observed before (70,74). A K^+ 'ion bridge' of the monovalent cation joining the two electronegative G-O6 on different strands (Figure 14a,b) was observed both in DNA and RNA (70). In our results, this ion bridge also extends to the C-N4 atom in the corresponding GpC steps for both Na^+ and K^+, and for both B-DNA and A-RNA. Binding to C-N7 is also important (Figures 12 and 13), but with shorter residence times. This can be explained by the geometry of the steps. A cation between G-O6 and C-N7 will also be attracted by the backbone oxygens, and therefore its motion will be more diffusive, less localized. The absence of binding to the CpG steps was attributed to the protruding of the two electropositive cytosine amino groups in the CpG steps (70), which is consistent with our observations. Finally, with respect to the backbone oxygens, there is direct binding to O3′ and water-mediated binding to both O3′ and O5′. Naturally, A-RNA also shows binding to O2′, although this is not as strong as for O3′ and O5′ (Figure 7). Both phosphate oxygens have comparable binding in both forms, B-DNA and A-RNA, and for both Na^+ and K^+ (Figures 9 and 13). Residence times are considerably longer for K^+ than for Na^+. Long residence times near the phosphates have been also observed in some recent B-DNA simulations (74).

The Mg^{2+} cation has quite different binding properties from its monovalent counterparts, due to its tightly bound first solvation shell. In the CDFs at 0.2-M excess salt concentration in Figure 3, we observe two clearly defined binding regions for both B-DNA and A-RNA (and a very small intermediate peak for A-RNA). In agreement with the RDFs and CDFs, Figure 4 shows that the distribution of Mg^{2+} cations is more localized than those for Na^+/K^+ for all four duplexes. This strong localization of Mg^{2+} ions around A-RNA was observed recently (86), where it occurs not only in the presence of $MgCl_2$ but also in mixtures with NaCl salt. The RDFs in Figure 6 show strong $Mg[(H_2O)_6]^{2+}$ binding in the major groove to G-O6 at 4.1 Å (B-DNA) and 4.3 Å (A-RNA), with a much higher peak in B-DNA, and a second important peak further away for C-N4. An example of this binding geometry is given in Figure 15 a and Figure S8, where the residence times for G-O6 are 9.4 ns for B-DNA and 2.4 ns for A-RNA. The closest backbone oxygen bindings occur at O3′ and O5′ with a more distant peak for O2′ in A-RNA. Binding to the two phosphate oxygens is very strong, with the first peak strongly centered at 4.1 Å. An interesting case occurs when the hexahydrated ion binds phosphate pairs belonging to distant base pairs as shown for A-RNA in Figure 15b and c resulting in acute bending of the duplex and the prevention of access of other ions to the middle of the duplex. Binding to the minor groove is negligible, considerably less than that of Na^+ or K^+ at 6 Å. Binding of $Mg[(H_2O)_6]^{2+}$ to the guanines

in the major grooves has been measured previously (86) and has been seen experimentally (45,50,120,121).

With respect to the dependence of the right-handed forms on the salt concentration, we found that B-DNA is less sensitive than A-RNA. There are no structural changes in B-DNA at 0.4-M NaCl and only at the high concentration of 4.0-M NaCl excess salt, we see what can be interpreted as the start of a salt-induced instability in base G8. This is manifested as a flipping of G8 from the 'anti' to the 'syn' conformation. By contrast, A-RNA is quite sensitive to salt concentration, exhibiting global changes even at 0.4-M NaCl. The nature of these changes is in complete agreement with previous findings (85,87). Mainly, the structure becomes more compact and switches more to the A-form as salt concentration increases. This is seen in an increase of inclination, step roll, helical and step twist, and x-displacement with salt concentration, accompanied by a decrease of helical and step rise, and propeller. These trends are coupled, as an increase in helical inclination leads to a larger base pair roll (which follows the same pattern as Figure S9d), narrowing of the major groove and to a reduction of helical rise. It is not clear whether this is an effect of the force field (which at present is the most accurate one for the description of nucleic acids (87)), but it has been observed (87) that a different water model (SPC/E) can result in an even more compact A-RNA structure. Unfortunately, the existing experimental data includes a wide span of compactness for A-RNA duplexes, which precludes determining whether this trend toward compactness with increasing NaCl salt is real or not. The trend would seem counter-intuitive as a high increase in salt is expected to lead to a change in handedness, and thus one would expect a decrease of both helical and step twist, an increase of rise, etc. with salt concentration. Experimentally, however, the transition has been achieved under different conditions, such as chemical modification of the bases (122); high pressure (30) or different salt conditions, such as 6.0-M $NaClO_4$ (29).

Comparison for different ion distributions between the left-handed and the right-handed forms

There are similarities but also strong differences in ion distributions around the left-handed forms. These differences are measured not only when comparing to the right-handed forms but also when comparing between Z-DNA and Z-RNA.

Since ion distribution is a function of the duplex structure, we briefly review Z-DNA and Z-RNA structures. Both duplexes, formed by antiparallel strands with WC base-pairing, are characterized by the typical zig-zag pattern of the sugar-phosphate backbone, and a dinucleotide repeat unit. In Z-DNA the major and minor grooves are very similar in width, while in Z-RNA the base pairs are closer to the helix axis (with smaller *x*- and *y*-displacements) and both a deep, narrow minor groove and the major groove are well-defined. In addition to handedness, other major differences with the right-handed forms include the glycosyl angle: it is 'anti' for both G and C in the right-handed duplexes, but in the left-handed forms it is 'syn' for G and 'anti' for C. Both handedness and glycosyl angles have been used as highly discriminating order parameters for the description of the structural transition between B-DNA and Z-DNA, leading to a complex free energy landscape which allows for the co-existence of several competing mechanisms (24). The predominant sugar pucker (C2′-endo in B-DNA; C3′-endo in A-RNA) is C2′-endo for C and C3′-endo for G in the Z forms. Twist angles of 33° to 38° for B-DNA and 29° to 34° for A-RNA also change drastically. The Z forms have higher negative values for the GpC steps, and smaller values for the CpG steps. Both forms have intra-strand stacking for GpC steps and inter-strand stacking for CpG steps, as observed experimentally in Z-DNA (1), and in Z-RNA at low ionic strengths (31). In Z-RNA, the C-O2′ groups are deeply buried in the narrow minor groove, while the G-O2′ and the phosphate oxygens reside on the outer helix surface. We have maintained the atomic definition of major and minor grooves (as shown in Figure 1) to follow the convention in the literature for Z-DNA (1) and for Z-RNA (29), although from a geometrical point of view, the role of these grooves appears to be inverted for the left-handed forms.

CDFs in Figure 2,3 show that: (i) the ion distributions converge faster in the left-handed forms than in the right-handed forms; (ii) the maximum peaks are closer to the helical axis in the right-handed forms than in the left-handed forms; (iii) the relative height of the maximum peaks depends on the cation type. For Na^+, first and second binding shells (when present) are higher in the Z forms than in the right-handed forms and, for each handedness, first peaks are higher for the RNA duplexes than for the DNA duplexes. While the K^+ first peaks are considerably higher than those for the other two cations in the right-handed forms, they are equal to or lower than those of Na^+ in the left-handed forms. Figure 4 shows that for any given distance before the asymptotic value, A-RNA is more screened than B-DNA. For regions close to the axis, A-RNA localizes more Na^+ and K^+ ions than both B-DNA and Z-DNA, and more Mg^{2+} ions than the three other forms. At intermediate distances, charge neutralization works better in the left-handed forms (which are comparable). For all duplexes, the Mg^{2+} ion distribution is considerably more localized than the Na^+ ion distribution, which in turn is slightly more localized than the K^+ ion distribution.

Binding to major and minor grooves changes dramatically when handedness changes (the role of the minor groove in the Z forms being closer to the role of the major groove in the right-handed forms). RDFs in Figure 5–9 show that (i) Na^+ binding in the major groove for both B-DNA and A-RNA is similar, driven mainly by binding to G-O6, with localization also close to C-N4, followed by C-N7 in the GpC steps. (ii) Na^+ binding in the minor groove of Z-DNA and Z-RNA is very different. Z-RNA has a very large peak for O2 at 2.3 Å and a second peak for N2 at 3.7 Å. Z-DNA has almost no binding at these short distances, with the first non-negligible peak for O2 at 4.6 Å. (iii) While Mg^{2+} binding in the major groove is stronger for B-DNA than A-RNA, its binding in the minor groove of Z-RNA is stronger than in Z-DNA. Binding to C-O2 shows two clearly defined bindings shells at 4.4 and 6.5 Å, which correspond to a direct binding to $Mg[(H_2O)_6]^{2+}$ and binding with one intermediate water molecule to $Mg[(H_2O)_6]^{2+}$. (iv) The O2′ in RNA naturally provides a source of binding differences between DNA and RNA. For Na^+ the patterns

of binding to O3′ and O5′ (and to less extent O4′) are relatively similar between B-DNA and A-RNA, with binding to O2′ less important than to O3′ and O5′. Instead, an important binding peak to O3′ at 2.5 Å in Z-DNA becomes minor in Z-RNA, while binding to O2′ at the same distance dominates. Similarly, for Mg^{2+} the patterns of binding to the O′ oxygens are similar between B-DNA and A-RNA, except for O2′ in A-RNA, which is not dominant. For Mg^{2+} in Z-RNA, the first binding peak to O3′ increases in Z-RNA with respect to Z-DNA, and the O2′ peak at 4.2 Å is also very important. (v) Binding of Na^+ to the OP2 oxygens is similar for the four duplexes, while binding to OP1 increases for the Z forms. For Mg^{2+} , these bindings are comparable for the four duplexes.

Figure 10–13 show sequence-specific features. Both for zero and high salt concentration, Na^+ direct binding to the major groove is small (K^+ exhibits slightly more binding in the major groove) for both Z-DNA and Z-RNA. Indirect binding to the major groove increases the population (centered at the G's). The left-handed forms have distinct patterns of strong direct binding to the phosphate oxygens in the Gs (larger in Z-DNA than Z-RNA) and to the O′ oxygens in the Cs (larger in Z-RNA due to binding to O2′). With respect to the minor groove, direct binding of Na^+ and K^+ is quite different for Z-DNA and Z-RNA: in Z-DNA the ions do not bind directly to the minor groove, but they do so, with large occupation numbers, through intermediate waters; in Z-RNA direct binding to the minor groove is quite high, mainly due to C-O2. Mg^{2+} binding at 6 Å follows a similar pattern: not much binding in the major groove, with slightly more binding in the minor groove (but certainly not as high as for Na^+ at 6 Å), preferential binding to the phosphate oxygens in the Gs, and to the O′ oxygens in the Cs.

In Z-DNA, the longest residence times for monovalent ions occur for G-OP1 (2.3 ns for Na^+ and 4.7 ns for K^+) with a second longest residence time at C-O3′ . Thus, in the 5′-3′ direction the ion binds to the O3′ of the sugar ring of a cytosine and to the two phosphates G-OP1 and G-OP2 immediately following C-O3′. This, with the equivalent set of atoms one CpG step ahead in the opposite strand, forms a pocket, as illustrated in Figure 14 c for Na^+ in Z-DNA. Table 2 indicates that for Na^+, G-OP1 is a site of direct binding while G-OP2 and C-O3′ are sites of indirect binding. On the other hand, K^+ finds itself in a tighter pocket, with direct binding to both G-OP1 and C-O3′, and indirect binding to G-OP2, which explains the longer residence times for K^+ associated with these positions. Z-RNA seems to provide the best 'ion trap' with direct binding to the C-O2 and the C-O2′ atoms in a GpC step in one strand and the same set of atoms belonging to the same GpC step in the opposite strand (see example in Figure 14d). Bound to four RNA atoms, the ion only retains two of its first solvation shell waters. This ion bridge has been observed experimentally at physiological ionic strengths (31) . These Z-RNA atoms exhibit the largest residence times for monovalent ions observed in our simulations: 7.8 ns (2.8 ns) for Na^+ (K^+) in C-O2 and 8.6 ns (5.4 ns) for Na^+ (K^+) in C-O2′. Finally, the hexahydrated $Mg[(H_2O)_6]^{2+}$ ion also finds strong binding pockets in both Z-DNA and Z-RNA, specially when the phosphate oxygens are involved. Typical binding pockets are shown in Figure 15d and Figure S9b and c. Long residence times in Z-DNA include 7 ns for G-OP1 and 5.6 ns for G-OP2. Long residence times in Z-RNA include 6.0 ns in C-O2′, 4.9 ns in C-O3′, 7.6 ns in G-OP1 and 5.0 ns in G-OP2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank the NC State HPC Center for extensive computational support.

FUNDING

National Science Foundation [NSF-1021883, NSF-1148144]. Funding for open access charge: National Science Foundation [NSF-1021883, NSF-1148144].
Conflict of interest statement. None declared.

REFERENCES

1. Wang,A.H., Quigley,G.J., Kolpak,F.J., Crawford,J.L., van Boom,J.H., van der Marel,G. and Rich,A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
2. Nordheim,A. and Rich,A. (1983) The sequence $(dC - dA)_n.(dG - dT)_n$ forms left-handed Z-DNA in negatively supercoiled plasmids. *Proc. Natl Acad. Sci. U.S.A.*, **80**, 1821.
3. Rich,A., Nordheim,A. and Wang,A.H. (1984) The chemistry and biology of left-handed Z-DNA. *Ann. Rev. Phys. Chem.*, **53**, 791–846.
4. Schroth,G., Chou,P. and Ho,P.J. (1992) Mapping Z-DNA in the human genome–computer-aided mapping revelas a nonrandom distribution of potential Z-DNA forming sequences in human genes. *J. Biol. Chem.*, **267**, 11846.
5. Liu,L.F. and Wang,J.C. (1987) Supercoiling of the DNA template during transcription. *Proc. Natl Acad. Sci. U.S.A.*, **84**, 7024–7027.
6. Oh,D., Kim,Y. and Rich,A. (2002) Z-DNA binding proteins can act as potent effectors of gene expression in vivo. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 16666.
7. Lipps,H.J. (1983) Antibodies against Z-DNA react with the macronucleus but not the micronucleus of the hypotrichous ciliate stylonychia mytilus. *Cell*, **32**, 435–441.
8. Lancillotti,F., Lopez,M., Alonso,C. and Stollar,B. (1985) Locations of Z-DNA in polytene chromosomes. *J. Cell Biol.*, **100**, 1759.
9. Herbert,A. and Rich,A. (1999) Left-handed Z-DNA: structure and function. *Genetica*, **106**, 37.
10. Kmiec,E., Angelides,K. and Holloman,W. (1985) Left-handed DNA and the synaptic pairing reaction promoted by ustilago rec1 protein. *Cell*, **40**, 139.
11. Blaho,J. and Wells,R. (1987) Left-handed Z-DNA binding by the reca protein of escherichia-coli. *J. Biol. Chem.*, **262**, 6082.
12. Jaworski,A., Hsieh,W., Blaho,J., Larson,J. and Wells,R. (1987) Left-handed DNA in vivo. *Science*, **238**, 773.
13. Schwartz,T., Rould,M., Lowenhaupt,K., Herbert,A. and Rich,A. (1999) Crystal structure of the Z alpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, **284**, 1841.
14. Vinogradov,A. (2003) DNA helix: the importance of being GC-rich. *Nucleic Acids Res.*, **31**, 1838.
15. Champ,P., Maurice,S., Vargason,J., Champ,T. and Ho,P. (2004) Distributions of Z-DNA and nuclear factor I in human chromosome 22; a model for coupled transcriptional regulation. *Nucleic Acids Res.*, **32**, 6501.
16. Wang,G., Christensen,L. and Vasquez,K. (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl Acad. Sci. U.S.A.*, **103**, 2677.

17. Fuertes,M.A., Cepeda,V., Alonso,C. and Pérez,J.M. (2006) Molecular mechanisms for the B-Z transition in the example of poly[d(G-C)·d(G-C)] polymers. A critical review. *Chem. Rev.*, **106**, 2045–2064.
18. Harvey,S. (1983) DNA structural dynamics—longitudinal breathing as a possible mechanism for the B-reversible-Z transition. *Nucleic Acids Res.*, **11**, 4867.
19. Goto,S. (1984) Characterization of intermediate conformational states in the B↔Z transition of $poly(dG-dC)\cdot poly(dG-dC)$. *Biopolymers*, **23**, 2211.
20. Saenger,W. and Hienemann,U. (1989) Raison d'etre and structural model for the B-Z transition of poly d(G-C)⋆poly d(G-C). *FEBS Lett.*, **257**, 223.
21. Ansevin,A. and Wang,A. (1990) Evidence for a new Z-type left-handed DNA helix - properties of Z(WC)-DNA. *Nucleic Acids Res.*, **18**, 6119.
22. Lim,W. and Feng,Y. (2005) The stretched intermediate model of B-Z DNA transition. *Biophys. J.*, **88**, 1593.
23. Ha,S.C., Lowenhaupt,K., Rich,A., Kim,Y.G. and Kim,K.K. (2005) Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature*, **437**, 1183.
24. Moradi,M., Babin,V., Roland,C. and Sagui,C. (2012) Reaction path ensemble of the B-Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Res.*, **41**, 33–43.
25. Hall,K., Cruz,P., Tinoco,I., Jovin,T. and van de Sande,J. (1984) Z-RNA—a left-handed RNA double helix. *Nature*, **311**, 584.
26. Adamiak,R.W., Galat,A. and Skalski,B. (1985) Salt- and solvent-dependent conformational transitions of ribo-CGCGCG duplex. *Biochim. Biophys. Acta*, **825**, 345–352.
27. Tinoco,I. Jr., Cruz,P., Davis,P., Hall,K., Hardin,C.C., Mathies,R.A., Puglisi,J.D., Trulson,M.A. Jr., Johnson,W.C. and Neilson,T. (1986) Z-RNA: a left-handed double helix. In: van Knippenberg,PH and Hilbers,CW (eds). *Structure and Dynamics of RNA*. Vol. 110 of NATO ASI Series A, Plenum Press, NY, pp. 55–69.
28. Popenda,M., Biala,E., Milecki,J. and Adamiak,R.W. (1997) Solution structure of RNA duplexes containing alternating CG base pairs: NMR study of $r(CGCGCG)_2$ and $2'-O-Me(CGCGCG)_2$ under low salt conditions. *Nucleic Acids Res.*, **25**, 4589–4598.
29. Popenda,M., Milecki,J. and Adamiak,R.W. (2004) High salt solution structure of a left-handed RNA double helix. *Nucleic Acids Res.*, **32**, 4044.
30. Krzyzaniak,A., Barciszewski,J., Furste,J.P., Bald,R., Erdmann,V.A., Salanski,P. and Jurczak,J. (1994) A-Z-RNA conformational changes effected by high pressure. *Int. J. Biol. Macromol.*, **16**, 159–162.
31. Placido,D., Brown,B., Lowenhaupt,K., Rich,A. and Athanasiadis,A. (2007) A left-handed RNA double helix bound by the Z-alpha domain of the RNA-editing enzyme ADAR1. *Structure*, **15**, 395.
32. Rau,D.C., Lee,B. and Parsegian,V.A. (1984) Measurement of the repulsive force between polyelectrolyte molecules in ionic solution: hydration forces between parallel DNA double helices. *Proc. Natl Acad. Sci. U.S.A.*, **81**, 2621–2625.
33. Knobler,C. and Gelbart,W. (2009) Physical chemistry of DNA viruses. *Annu. Rev. Phys. Chem.*, **60**, 367–383.
34. Fedor,M. (2009) Comparative enzymology and structural biology of RNA self-cleavage. *Annu. Rev. Biophys.*, **38**, 271–299.
35. Draper,D.E., Grilley,D. and Soto,A.M. (2005) Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 221–243.
36. Koculi,E., Hyeon,C., Thirumalai,D. and Woodson,S.A. (2007) Charge density of divalent metal cations determines RNA stability. *J. Am. Chem. Soc.*, **129**, 2676–2682.
37. Li,P., Vieregg,J. and Tinoco,I. (2008) How RNA unfolds and refolds. *Annu. Rev. Biophys.*, **77**, 77–100.
38. MacKerrell,A. Jr. and Nilsson,L. (2008) Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, **18**, 194–199.
39. Auffinger,P., Bielecki,L. and Westhof,E. (2004) Anion binding to nucleic acids. *Structure*, **12**, 379–388.
40. Prive,G.G., Yanagi,K. and Dickerson,R.E. (1991) Structure of the B-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J. Mol. Biol.*, **217**, 177–199.
41. Shui,X., Sines,C.C., McFail-Isom,L., Van Derveer,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.
42. Shui,X., McFail-Isom,L., Hu,G. and Williams,L.D. (1998) The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*, **37**, 8341–8355.
43. Tereshko,V., Minasov,G. and Egli,M. (1999) A 'hydration' spine in a B-DNA minor groove. *J. Am. Chem. Soc.*, **121**, 3590–3595.
44. McFail-Isom,L., Sines,C.C. and Williams,L.D. (1999) DNA structure: cations in charge?. *Curr. Opin. Struct. Biol.*, **9**, 298–304.
45. Ennifar,E., Yusupov,M., Walter,P., Marquet,R., Ehresmann,B., Ehresmann,C. and Dumas,P. (1999) The crystal structure of the dimerization initiation site of genomic HIV-1 RNA reveals an extended duplex with two adenine bulges. *Struct. Fold. Des.*, **7**, 1439–1449.
46. Robinson,H., Gao,Y.G., Sanishvili,R., Joachimiak,A. and Wang,A. H.J. (2000) Hexahydrated magnesium ions bind in the deep major groove and at the outer mouth of A-form nucleic acid duplexes. *Nucleic Acids Res.*, **28**, 1760–1766.
47. Howerton,S.B., Sines,C.C., VanDerveer,D. and Williams,L.D. (2001) Locating monovalent cations in the grooves of B-DNA. *Biochemistry*, **40**, 10023–10031.
48. Subirana,J.A. and Soler-Lopez,M. (2003) Cations as hydrogen bond donors: a view of electrostatic interactions in DNA. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 27–45.
49. Chiu,T.K. and Dickerson,R.E. (2000) 1 angstrom crystal structure of B-DNA real sequence specific and groove specific bending of DNA by magnesium and calcium. *J. Mol. Biol.*, **301**, 915.
50. Ennifar,E., Walter,P. and Dumas,P. (2003) A crystallographic study of the binding of 13 metal ions to two related RNA duplexes. *Nucleic Acids Res.*, **31**, 2671–2682.
51. Timsit,Y. and Bombard,S. (2007) The 1.3 angstrom resolution structure of the RNA tridecamer r(GCGUUUGAAACGC): metal ion binding correlates with base unstacking and groove contraction. *RNA*, **13**, 2098–2107.
52. Tereshko,V., Minasov,G. and Egli,M. (1999) The Dickerson-Drew B-DNA dodecamer revisited at atomic resolution. *J. Am. Chem. Soc.*, **121**, 470–471.
53. Chiu,T.K., Kaczor-Grzeskowiak,M. and Dickerson,R.E. (1999) Absence of minor groove monovalent cations in the crosslinked dodecamer CGCGAATTCGCG. *J. Mol. Biol.*, **292**, 589–608.
54. Andresen,K., Das,R., Park,H.Y., Smith,H., Kwok,L.W., Lamb,J.S., Kirkland,E.J., Herschlag,D., Finkelstein,K.D. and Pollack,L. (2004) Spatial distribution of competing ions around DNA in solution. *Phys. Rev. Lett.*, **93**, 248103.
55. Chu,V., Bai,Y., Lipfert,J., Herschlag,D. and Doniach,S. (2008) A repulsive field: advances in the electrostatics of the ion atmosphere. *Curr. Opin. Chem. Biol.*, **12**, 619–625.
56. Wong,G. C.L. and Pollack,L. (2010) Electrostatics of strongly charged biological polymers: ion-mediated interactions and self-organization in nucleic acids and proteins. *Annu. Rev. Phys. Chem.*, **61**, 171–189.
57. Bai,Y., Greenfeld,M., Travers,K., Chu,V., Lipfert,J., Doniach,S. and Herschlag,D. (2007) Quantitative and comprehensive decomposition of the ion atmosphere around nucleic acids. *J. Am. Chem. Soc.*, **129**, 14981–14988.
58. Hud,N.V. and Feigon,J. (1997) Localization of divalent metal ions in the minor groove of DNA A-tracts. *J. Am. Chem. Soc.*, **119**, 5756–5757.
59. Hud,N.V., Sklenar,V. and Feigon,J. (1999) Localization of ammonium ions in the minor groove of DNA duplexes in solution and the origin of DNA A-tract bending. *J. Mol. Biol.*, **286**, 651–660.
60. Bonvin,A. (2000) Localisation and dynamics of sodium counterions around DNA in solution from molecular dynamics simulation. *Eur. Biophys. J.*, **29**, 57–60.
61. Denisov,V.P. and Halle,B. (2000) Sequence-specific binding of counterions to B-DNA. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 629–633.
62. Marincola,F.C., Denisov,V.P. and Halle,B. (2004) Competitive Na+ and Rb+ binding in the minor groove of DNA. *J. Am. Chem. Soc.*, **126**, 6739–6750.
63. Maehigashi,T., Hsiao,C., Woods,K., Moulaei,T., Hud,N.V. and Williams,L.D. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, **40**, 3714–3722.

64. Manning,G.S. (1978) Molecular theory of polyelectrolyte solutions with applications to electrostatic properties of polynucleotides. *Q. Rev. Biophys.*, **11**, 179–246.
65. Bacquet,R.J. and Rossky,P.J. (1988) Ionic distributions and competitive association on DNA mixed salt solutions. *J. Phys. Chem.*, **92**, 3604–3612.
66. York,D.M., Darden,T., Deerfield,D. and Pedersen,L.G. (1992) The interaction of Na(I), Ca(II), and Mg(II) metal-ions with duplex DNA—a theoretical modeling study. *Intl. J. Quant. Chem.*, **19**, 145–166.
67. Chen,S.W.W. and Honig,B. (1997) Monovalent and divalent salt effects on electrostatic free energies defined by the nonlinear Poisson-Boltzmann equation: application to DNA binding reaction. *J. Phys. Chem. B*, **101**, 9113–9118.
68. Young,M.A., Jayaram,B. and Beveridge,D.L. (1997) Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: fractional occupancy of electronegative pockets. *J. Am. Chem. Soc.*, **119**, 59–69.
69. Feig,M. and Pettitt,B.M. (1999) Sodium and chlorine ions as part of the DNA solvation shell. *Biophys. J.*, **77**, 1769–1781.
70. Auffinger,P. and Westhof,E. (2000) Water and ion binding around RNA and DNA (C,G) oligomers. *J. Mol. Biol.*, **300**, 1113–1131.
71. 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.
72. Rueda,M., Cubero,E., Laughton,C.A. and Orozco,M. (2004) Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations?. *Biophys. J.*, **87**, 800–811.
73. Ponomarev,S.Y., Thayer,K.M. and Beveridge,D.L. (2004) Ion motions in molecular dynamics simulations on DNA. *Proc. Natl Acad. Sci. U.S.A.*, **101**, 14771–14775.
74. Várnai,P. and Zakrzewska,K. (2004) DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.*, **320**, 4269–4280.
75. Taubes,C.H., Mohanty,U. and Chu,S. (2005) Ion atmosphere around nucleic acid. *J. Phys. Chem. B*, **109**, 21267–21272.
76. Auffinger,P. and Hashem,Y. (2007) Nucleic acid solvation: from outside to insight. *Curr. Opin. Struct. Biol.*, **17**, 325–333.
77. Garcia,A.E. and Paschek,D. (2008) Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin. *J. Am. Chem. Soc.*, **130**, 815.
78. Yoo,J. and Aksimentiev,A. (2012) Competitive binding of cations to duplex DNA revealed through molecular dynamics simulations. *J. Phys. Chem. B*, **116**, 12946–12954.
79. Perez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.*, **45**, 196–205.
80. Cisneros,G.A., Karttunen,M., Ren,P. and Sagui,C. (2014) Classical electrostatics for biomolecular simulations. *Chem. Rev.*, **114**, 779–814.
81. York,D.M., Darden,T.A. and Pedersen,L.G. (1993) The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the Ewald and truncated list methods. *J. Chem. Phys.*, **99**, 8345–8348.
82. Darden,T.A., York,D.M. and Pedersen,L.G. (1993) Particle mesh Ewald: an N log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
83. York,D.M., Yang,W., Lee,H., Darden,T.A. and Pedersen,L.G. (1995) Towards the accurate modeling of DNA: the importance of long-range electrostatics. *J. Am. Chem. Soc.*, **117**, 5001–5002.
84. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577–8593.
85. Besseova,I., Otyepka,M., Reblova,K. and Sponer,J. (2009) Dependence of A-RNA simulations on the choice of the force field and salt strength. *Phys. Chem. Chem. Phys.*, **11**, 10701–10711.
86. Kirmizialtin,S. and Elber,R. (2010) Computational exploration of mobile ion distributions around RNA duplex. *J. Phys. Chem. B*, **114**, 8207–8220.
87. Besseova,I., Barnas,P., Kuhrova,P., Kosinova,P., Otyepka,M. and Sponer,J. (2012) Simulations of A-RNA duplexes. The effect of sequence, solute force field, water model, and salt concentration. *J. Phys. Chem. B*, **116**, 9899–9916.
88. Kirmizialtin,S., Pabit,S.A., Meisburger,S.P., Pollack,L. and Elber,R. (2012) RNA and its ionic cloud: solution scattering experiments and atomically detailed simulations. *Biophys. J.*, **102**, 819–828.
89. Kirmizialtin,S., Silalahi,A.R., Elber,R. and Fenley,M.O. (2012) The ionic atmosphere around A-RNA: Poisson-Boltzmann and molecular dynamics simulations. *Biophys. J.*, **102**, 829–838.
90. Perez,A., Marchan,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.*, **92**, 3817–3829.
91. Banas,P., Hollas,D., Zgarbva,M., Jurecka,P., Orozco,M., Cheatham,T.E., Sponer,J. and Otyepka,M. (2010) Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.*, **6**, 3836–3849.
92. Zgarbova,M., Otyepka,M., Sponer,J., Mladek,A., Banas,P., Cheatham,T.E. and Jurecka,P. (2011) Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.*, **7**, 2886–2902.
93. Case,D.A., Darden,T.A., Cheatham,T.E. III, Simmerling,C.L., Wang,J., Duke,R.E., Luo,R., Walker,R.C., Zhang,W., Merz,K.M. *et al.* (2012) *AMBER 12*. University of California, San Francisco, CA.
94. Jorgensen,W.L., Chandrasekhar,J., Madura,J. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
95. Naslund,P.H. and Hultin,T. (1971) Structural and functional defects in mammalian ribosomes after potassium deficiency. *Biochim. Biophys. Acta*, **254**, 104–116.
96. Ennis,H.L. and Artman,M. (1972) Ribosome size distribution in extracts of potassium-depleted Escherichia coli. *Biochem. Biophys. Res. Commun.*, **48**, 161–168.
97. Mahler,J. and Persson,I. (2012) A study of the hydration of the alkali metal ions in aqueous solution. *Inorg. Chem.*, **51**, 425–438.
98. Ohtaki,H. and Radnai,T. (1993) Structure and dynamics of hydrated ions. *Chem. Rev.*, **93**, 1157.
99. Tissandier,M.D., Cowen,K.A., Feng,W.Y., Gundlach,E., Cohen,M.H., Earhart,A.D., Coe,J.V. and Tuttle,T.R. Jr. (1998) The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster-ion solvation data. *J. Phys. Chem. A*, **102**, 7787–7794.
100. Schmid,R., Miah,A.M. and Sapunov,V.N. (2000) A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii). *Phys. Chem. Chem. Phys.*, **2**, 97–102.
101. Thirumalai,D., Lee,N., Woodson,S.A. and Klimov,D.K. (2001) Early events in RNA folding. *Annu. Rev. Phys. Chem.*, **52**, 751–762.
102. Auffinger,P., Bielecki,L. and Westhof,E. (2003) The Mg2+ binding sites of the 5S rRNA loop E motif as investigated by molecular dynamics simulations. *Chem. Biol.*, **10**, 551–561.
103. Das,R., Kwok,L.W., Millett,I.S., Bai,Y., Mills,T.T., Jacob,J., Maskel,G.S., Seifert,S., Mochrie,S. G.J., Thiyagarajan,P. *et al.* (2003) The fastest global events in RNA folding: electrostatic relaxation and tertiary collapse of the tetrahymena ribozyme. *J. Mol. Biol.*, **332**, 311–319.
104. Auffinger,P., Bielecki,L. and Westhof,E. (2004) Symmetric K+ and Mg2+ ion-binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J. Mol. Biol.*, **335**, 555–571.
105. Woodson,S.A. (2005) Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr. Opin. Chem. Biol.*, **9**, 104–109.
106. Das,R., Travers,K.J., Bai,Y. and Herschlag,D. (2005) Determining the Mg^{2+} stoichiometry for folding an RNA metal ion core. *J. Am. Chem. Soc.*, **127**, 8272–8273.
107. Grilley,D., Soto,A.M. and Draper,D.E. (2006) Mg^{2+}-RNA interaction free energies and their relationship to the folding of RNA tertiary structures. *Proc. Natl Acad. Sci. U.S.A.*, **103**, 14003–14008.
108. Jiao,D., King,C., Grossfield,A., Darden,T. and Ren,P. (2006) Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential. *J. Phys. Chem. B*, **110**, 18553–18559.
109. Baucom,J., Transue,T., Fuentes-Cabrera,M.A., Krahn,J.M., Darden,T. and Sagui,C. (2004) Molecular dynamics simulations of the d(CCAACGTTGG)$_2$ decamer in crystal environment:

comparison of atomic point-charge, extra-point and polarizable force fields. *J. Chem. Phys.*, **121**, 6998–7008.
110. Babin,V., Baucom,J., Darden,T.A. and Sagui,C. (2006) Molecular dynamics simulations of DNA with polarizable force fields: convergence of an ideal B-DNA structure to the crystallographic structure. *J. Phys. Chem. B*, **110**, 11571–11581.
111. Babin,V., Baucom,J., Darden,T.A. and Sagui,C. (2006) Molecular dynamics simulations of polarizable DNA in crystal environment. *Int. J. Quantum Chem.*, **106**, 3260–3269.
112. Perera,L. and Berkowitz,M.L. (1991) Many-body effects in molecular-dynamics simulations of $Na^+(H_2O)_n$ and $Cl^-(H_2O)_n$ clusters. *J. Chem. Phys.*, **95**, 1954–1963.
113. Markovich,G., Giniger,R., Levin,M. and Cheshnovsky,O. (1991) Photoelectron-spectroscopy of iodine anion solvated in water clusters. *J. Chem. Phys.*, **95**, 9416–9419.
114. Perera,L. and Berkowitz,M.L. (1992) Structure and dynamics of $Cl^-(H_2O)_{20}$ clusters: the efect of the polarizability and the charge of the ion. *J. Chem. Phys.*, **96**, 8288–8294.
115. Perera,L. and Berkowitz,M.L. (1993) Stabilization energies of Cl^-, Br^-, and I^- ions in water clusters. *J. Chem. Phys.*, **99**, 4222–4224.
116. Perera,L. and Berkowitz,M.L. (1993) Many-body effects in molecular-dynamics simulations of $Na^+(H_2O)_n$ and $Cl^-(H_2O)_n$ clusters (VOL 95, PG 1954, 1991). *J. Chem. Phys.*, **99**, 4236–4237.
117. Herce,D.H., Perera,L., Darden,T.A. and Sagui,C. (2004) Surface solvation for an ion in a water cluster. *J. Chem. Phys.*, **122**, 024513.
118. Lu,X.J., Hassan,M.A.E. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *J. Mol. Biol.*, **273**, 668–680.
119. Rosenberg,J.M., Seeman,N.C., Day,R.O. and Rich,A. (1976) RNA double helices generated from crystal-structures of double helical dinucleoside phosphates. *Biochem. Biophys. Res. Commun.*, **69**, 979–987.
120. Froystein,N., Davis,J., Reid,B. and Sletten,E. (1993) Sequence-selective metal-ion binding to DNA oligonucleotides. *Acta Chem. Scand.*, **47**, 649–657.
121. Cate,J., Hanna,R. and Doudna,J. (1997) A magnesium ion core at the heart of a ribozyme domain. *Nat. Struct. Biol.*, **4**, 553–558.
122. Uesugi,S., Ohkuko,M., Urata,H., Ikehara,M., Kobayshi,Y. and Kyogoki,Y. (1984) Ribooligonucleotides r(C-G-C-G) analogues containing 8-substituted guanosine residues form left-handed duplexes with Z-form-like structures. *Am. Chem. Soc.*, **106**, 3675–3676.

Chapter 3

Comparative melting and healing of B-DNA and Z-DNA by an infrared laser pulse

Viet Hoang Man, Feng Pan, Celeste Sagui and Christopher Roland (2016) Comparative melting and healing of B-DNA and Z-DNA by an infrared laser pulse. *The Journal of Chemical Physics. 144*, 145101.

THE JOURNAL OF CHEMICAL PHYSICS **144**, 145101 (2016)

Comparative melting and healing of B-DNA and Z-DNA by an infrared laser pulse

Viet Hoang Man, Feng Pan, Celeste Sagui,[a] and Christopher Roland[b]
Department of Physics, North Carolina State University, Raleigh, North Carolina 27695-8202, USA

(Received 12 January 2016; accepted 21 March 2016; published online 12 April 2016)

We explore the use of a fast laser melting simulation approach combined with atomistic molecular dynamics simulations in order to determine the melting and healing responses of B-DNA and Z-DNA dodecamers with the same d(5′-CGCGCGCGCGCG-3′)$_2$ sequence. The frequency of the laser pulse is specifically tuned to disrupt Watson-Crick hydrogen bonds, thus inducing melting of the DNA duplexes. Subsequently, the structures relax and partially refold, depending on the field strength. In addition to the inherent interest of the nonequilibrium melting process, we propose that fast melting by an infrared laser pulse could be used as a technique for a fast comparison of relative stabilities of same-sequence oligonucleotides with different secondary structures with full atomistic detail of the structures and solvent. This could be particularly useful for nonstandard secondary structures involving non-canonical base pairs, mismatches, etc. © *2016 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4945340]

I. INTRODUCTION

Recently, free electron lasers have been used for melting biomolecular complexes.[1–5] Kawasaki and co-workers[3–5] have developed a mid-infrared free electron laser with highly specific oscillation characteristics having a high photon density, a picosecond pulse structure, and a range of tunable frequencies. To date, such a laser pulse with frequencies tuned to the amide I bands has been used in experiments to dissociate amyloid-like fibrils of lyzozyme into its native forms,[3] convert insulin fibrils into soluble monomers,[4] and dissociate a short human thyroid hormone peptide.[5] A simulated laser pulse has also been used in atomistic molecular dynamics simulations to dissociate amyloid fibrils[6] and viruses[7] and to study the break-up of a peptide-based nanotube.[8] Aside from investigating the dissociation of selected biomolecules, in this work we propose that the application of a simulated laser pulse could provide for new computational opportunities to probe the relative response and, quite possibly, the relative stability of similar structures, such as nucleic acids with the same sequence but with different secondary structures.

A crucial aspect for the understanding of nucleic acid structures and their function is the relative stability of different competing structures that are involved in cellular processes. The discernment of the relative structural stabilities of different nucleic acid structures, both *in vitro* and *in silico*, is rather challenging. A time honored technique to address this issue is to carry out a thermal denaturation experiments,[9] where a sample is heated beyond the melting point. This causes a conformational transition in the molecule that is measured by recording a specific variable as a function of temperature. Typical experiments include the recording of the absorbance at 260 nm in a UV-visible spectrophotometer;[10] circular dichroism spectra;[11,12] fluorescence spectra;[13] Raman signals;[14,15] or NMR measurements.[16,17]

Thermal denaturing experiments can also be carried out computationally, but fully atomistic melting simulations are extremely expensive. Instead, simplified models have been introduced for estimating the melting temperature of DNA duplexes, based on empirical or statistical thermodynamics models[18–34] that predict the stability of nucleic acid secondary structure, including RNA.[22,35–48] Unfortunately, these models are still empirical and can fail to capture subtle sequence-dependent effects, or unusual conformations different from A and B duplexes, or conformations with noncanonical or mismatched base pairs, etc. A state-of-the-art web-based tool for predicting fluorescent high-resolution DNA melting curves and denaturation profiles of PCR products is the software package uMELT.[49] This tool, however, can only be applied to standard DNA duplexes. Since secondary structures other than B-DNA can play crucial roles in processes such as gene expression, extending the computational predictive capacity to these unusual structures is a desirable goal.

In this work, we use fast melting by an infrared laser pulse of B-DNA and Z-DNA duplexes with sequence d(5′-CGCGCGCGCGCG-3′)$_2$ in order to compare the melting and healing response of the duplexes and to explore whether this response reflects their relative stability. The left-handed, double-helix Z-DNA with two antiparallel chains joined by Watson-Crick (WC) base pairs was first crystallized in 1979.[50] Z-DNA is favored by sequences that alternate purines and pyrimidines, mainly CG or GC. The left-handed helix is thinner and more rigid than B-DNA, and while the pyrimidine-purine base pairs are in *anti-anti* conformation in B-DNA, the guanine rotates around its glycosidic bond in Z-DNA, resulting in an *anti-syn* configuration.[51,52] As a consequence of these rotations, the phosphate groups become closer in Z-DNA; the sugar-phosphate backbone displays the characteristic zig-zag pattern that gave rise to the name Z-DNA; and the CpG

[a]Email: sagui@ncsu.edu
[b]Email: cmroland@ncsu.edu

and GpC steps stack differently, causing a dinucleotide step to be the repeating unit in Z-DNA. A high density of base sequences favoring Z-DNA is found near transcription start sites,[53] where Z-DNA is stabilized by negative supercoiling of DNA.[51,52,54] In the cell, Z-DNA is induced by a set of binding proteins near promoter regions, which boosts the transcription of downstream genes.[55] Z-DNA is highly immunogenic, and antibodies against it[56–58] are used to find locations prone to Z-DNA conformations. The current view is that Z-DNA formation plays a role in gene expression, regulation, and recombination.[52,55,59–65]

B-DNA is more stable than Z-DNA under physiological ionic strength and pH conditions. While in the cell Z-DNA can be stabilized by the negative supercoiling[51,52,54] that, for instance, arises from the relocation of the RNA polymerase during transcription, and through binding with highly specific Z-DNA binding proteins such as ADAR1, E3L, and PKZ,[66–69] *in vitro* inducers and stabilizers are needed for Z-DNA, such as high ionic concentrations, base chemical modifications, organic solvents, divalent cations and transition metal complexes, and small molecular complexes. (these are reviewed in Ref. 70). Computationally, the determination of the relative stability of B-DNA and Z-DNA is not trivial. Existing predictive tools do not have provisions for this alternative structure, and therefore one has to resort to very long time simulations or careful theoretical and statistical modeling. Considerable insight has been obtained by studying the microscopics behind the B-Z DNA transition,[50,71–81] and the ion distribution around these nucleic acid structures.[82] Recent molecular dynamics (MD) simulations indicate that the transition is governed by a complex free energy landscape which allows for the coexistence of several competing mechanisms so that the transition is better described in terms of a reaction path ensemble.[81]

In this work, we choose B-DNA and Z-DNA duplexes with sequence d(5′-CGCGCGCGCGCG-3′)$_2$ to test their behavior when exposed to fast melting by an infrared laser pulse. In addition to the intrinsic interest of the melting and healing process, we propose that this technique could be used as a relatively inexpensive tool to determine relative stability in fully solvated, atomically accurate nucleic acid structures. Naturally, a laser pulse implies a nonequilibrium process and the results cannot be used to construct an equilibrium melting curve, and therefore the ensuing free energy estimates cannot be obtained either. The perturbation, however, can map out the responses of the structures for the entire range of applied fields, covering extremely small perturbations all the way to complete melting. The response can be measured by a wide range of variables including normalized handedness, base stacking, Watson-Crick hydrogen bonds, etc. A systematic trend in all these variables (in this case that Z-DNA is more "melted" than B-DNA) for all strengths of the field is interpreted as being indicative of the greater stability of B-DNA under the solvation conditions in the simulations. A traditional equilibrium molecular dynamics simulation not only is orders of magnitude more expensive but also faces additional challenges. Since B-DNA is more prone to fray at the ends than Z-DNA, the melting of short B-DNA oligomers is more susceptible to length effects than Z-DNA, a difficulty that can be partially bypassed with the laser melting technique. In addition, Z-DNA is believed to be stabilized by higher temperatures,[73,83–85] which complicates the interpretation of the traditional melting experiments.

II. MATERIALS AND METHODS

A. Simulation details

Simulations of B-DNA and Z-DNA duplexes with sequence (5′-CGCGCGCGCGCG-3′)$_2$ in TIP3P waters[86] were carried out using the GROMACS 4.5.5 package[87] with the AMBER99SB-ILDN force field and DNA parameters ff99BSCO.[88] The initial DNA configurations were taken from our previous studies.[81,82] For each DNA (B-DNA or Z-DNA) structure, we performed simulations at five different NaCl salt concentrations: 0, 1.0, 2.5, 3.0, and 4.0M, with ion parameters by Joung and Cheatham.[89] Specifically, 22 Na^+ neutralizing ions were added, followed by an additional number of Na^+ and Cl^-.

Ions to make up the specified NaCl salt concentration. The DNA was placed in an octahedral box containing about 7050 water molecules with a 1 nm distance from the solute to the box boundary, resulting in a box size and volume of 6.6 nm and 200 nm^3, respectively. Periodic boundary conditions were imposed with a 1 nm cutoff for the van der Waals and electrostatic interactions. The long-range electrostatic interactions were computed with the Particle-Mesh Ewald summation method,[90] and the non-bonded interaction pair list updated every 10 fs. The covalent bonds were constrained using the LINCS algorithm[91] with a relative geometrical tolerance of 10^{-4}. The simulations were primarily NPT and made use of the Berendsen pressure coupling method[92] to model the barostat (pressure was kept at 1 atm). To ensure stability, a time step of 0.2 fs was used, and the equations of motion were integrated using the Leapfrog method.[93] To control the temperature, we used a V-rescaling temperature coupling scheme based on rescaling the velocity with a stochastic term.[94] Since a laser pulse was applied to DNA, only the solvent molecules were coupled to the heat bath which was kept at 310 K, with a temperature coupling constant of 0.1 ps. Experimentally, the DNA samples would be immersed in an effectively "infinite" reservoir of solvent (perhaps even with added flow). However, the simulations are limited in terms of the sizes the one can deal with, even with current supercomputers necessitating the use of a thermostat in order to damp out the thermal fluctuations of the solvent.

The laser pulse was modeled by the following equation:

$$E(t) = E_0 \exp\left[-\frac{(t-t_0)^2}{2\sigma^2}\right] \cos[c\omega(t-t_0)],$$

with E_0 denoting the amplitude of the electric field, σ the pulse width, t the time, t_0 the time when the pulse is maximal, ω the frequency, and c the velocity of light. An important issue that needs consideration is the direction of the laser pulse with respect to the orientation of the DNA molecule. While a given laser beam is polarized in a specified direction, the free electron laser experiments typically take place in an unpolarized environment. This is important, because the

applied laser pulse will be most efficient in exciting a given bond if the direction of the electric field is along the direction of the bond. In this study, the native hydrogen bonds (H-bonds) (as listed in supplementary material Table S2) connecting the C–G base pairs (corresponding to bond indices 14 to 33 in supplementary material Table S1) are targeted, which are oriented almost perpendicular to the DNA axis.[101] If the direction of the laser pulse (LPD) is parallel to the DNA axis, the laser pulse will have only a minimal effect on these bonds, even when the correct resonant frequency is used. Moreover, during the course of the simulation, the orientation of the DNA molecule may shift. To solve this issue, we used the following averaging procedure. For the simulations scanning the different frequencies, ten different LPDs were set and the results averaged. Specifically, the LPDs were set in parallel to the ten internal vectors $C1'_i - C1'_{25-i}$, with i representing the base index as shown in Fig. 1, and $C1'_i$ the $C1'$ atom of the ith base of the starting configuration. For the laser melting simulation, averaging was performed as follows. First, the average vector (AV) of the ten $C1'_i - C1'_{25-i}$ was calculated. Then, the angle between this AV vector and a given $C1'_i - C1'_{25-i}$ was divided into four equal parts, thereby defining five different directions. Given the ten different $C1' - C1'_{25-i}$ vectors, this defines a total of fifty different LPDs. Simulations were performed for each of these directions and the results were then averaged.

For this project we carried out three kinds of simulations: equilibration, frequency scanning, and laser melting simulations. To equilibrate the solvated DNA (both with and without salt), we carried out NPT simulations at 310 K for 30 ns. From these simulations, 100 independent conformations were selected at random from the last 10 ns. These configurations provided us with initial configurations for the subsequent simulations. Following equilibration, we carried out laser scanning simulations in order to select the best frequency for the laser melting simulations. These simulations were carried out for both B- and Z-DNA in the absence of excess salt with laser parameters E_0, t_0 and σ set, respectively, to 1 V/nm, 10 ps, and 3 ps. The frequency ω was varied between 1600 to 1935 cm^{-1}, with frequency steps set to 5 cm^{-1}. The latter was lowered to 1 cm^{-1} around the observed resonant peaks in order to provide for a better resolution. In total, 94 different frequencies were scanned. For each ω value, ten 30 ps simulations with different initial conditions selected from the 100 conformations in the equilibration step were then carried out, and the results averaged. From the mathematical expression of the laser pulse, we see that there are four parameters. Since the frequency ω is a property of the system, we actually have three parameters. For these simulations we chose two pulse widths, σ = 3 ps and σ = 6 ps and the two sets of simulations had the following parameters: (i) σ = 3 ps, t_0 = 10 ps, E_0 = (3.0, 4.0, 4.6, 5.0, 5.5, 6.0 and 6.5) V/nm, and total time for each run of 40 ps; (ii) σ = 6 ps, t_0 = 20 ps, E_0 = (3.0, 3.5, 4.0, 4.5 and 5.0) V/nm, and total time for each run of 100 ps. For each value of E_0, 50 simulations were carried out using different initial structures selected from the equilibration step. The values of the electric field used in our simulation range from about 6.0-20.0 J/cm^2 (depending

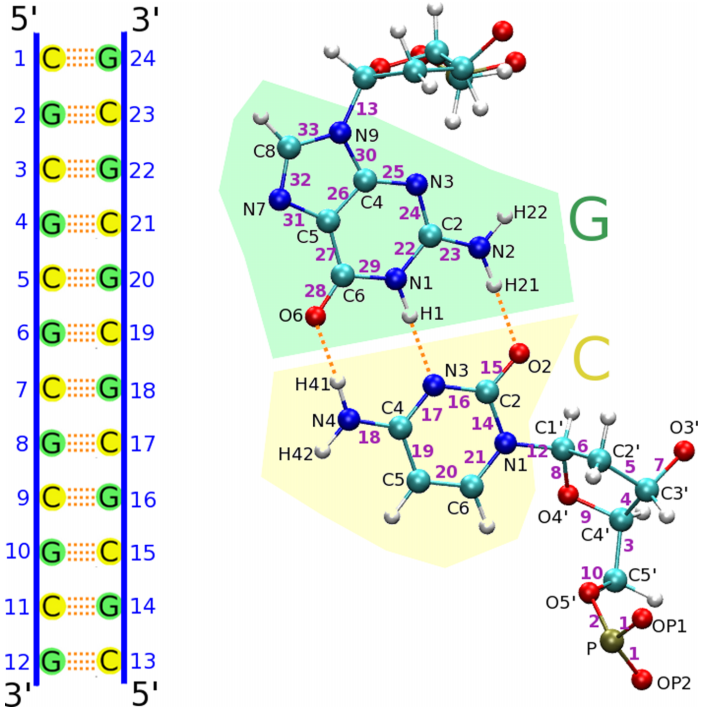


FIG. 1. DNA sequence (left) and structure of a CG pair (right). Bases in the sequence are numbered. In the CG base pair, the indices for the bond type are given as purple numbers, while the atomic labels are given in black. Details of the different bond types are given in Table S1 in supplementary material.[101]

on how one estimates the pulse shape), which is about 100 times higher than the current experimental values. Choosing such a high laser intensity is necessitated by the fact that experimentally a very large number of pulses are delivered over a relatively long period of time (microseconds). However, MD simulations are plagued by the so-called time scale problem, which precludes simulations over such long time scales. Here, we have chosen to add all the energy over a relatively short period of time.

The laser melting of the DNA structures was characterized in a number of standard ways. The DNA structure is held together by Watson-Crick H-bonds, so that their number represents a reasonable measure of the native structure. For both B- and Z-DNA, there are three H-bonds for each CG pair, giving a total of thirty-six H-bonds for the entire structure. Throughout the simulation, we tracked the percentage of the Watson-Crick H-bonds (WCHBs) as a function of time. As a further refinement, we also considered the base pair opening probability (BPOP), which represents the probability that a given base pair H-bond is broken.

Since the two major determinants of the stability of a helical duplex are the base-pair H-bonds and the base stacking energy, we have also defined a base-stacking index through the following procedure. A base is represented by a convex pentagon determined by five base atoms. In a C(G) base, these atoms are O2(O6), N4(N2), C5(N9), C6(C8), and N1(N7) (see Fig. 1). Given the stacking index (st) between two bases i and j ($st_{i,j}$), the base-stacking index (BSI) was calculated via the following equations:

$$st_{i,j} = \frac{S_{overlap}}{r_{i,j}^2},$$

$$V_i = st_{i,i+1} + st_{i,N-i} + st_{N+1-i,i+1} + st_{N+1-i,N-i},$$

$$BSI = \frac{1}{N-2} \times \sum_{i=1}^{(N-2)/2} V_i.$$

In these equations, $S_{overlap}$ is the overlapping area between successive bases, given by the projection of each pentagon plane on the next ones along the chain, $r_{i,j}$ is the distance between the centers of mass of the two pentagons, and N is the total number of bases. Note that the choice of subindexes allows for the projection of each base pentagon onto both next ones.

Another useful measure of the duplex structure is given by its handedness (H).[95] Handedness is a natural choice here, since B- and Z-DNA are right-handed and left-handed helical structures. Our working definition of H is based on our previous work on polyproline peptides[95] and extended to transitions between B- and Z-DNA.[81] It discriminates between left- ($H < 0$) and right-handed ($H > 0$) helical structures. For the DNA double helix, the position of the phosphorus (P) atoms of the backbone phosphate group was found to be a good choice for the definition of H.[81] In brief, the definition of H for a portion of DNA between the base pairs n and m makes use of a sequence of P atoms: P_n^1, P_n^2, P_{n+1}^1, P_{n+1}^2,...,P_m^1, P_m^2, where the upper index indicates the strand number (1 or 2) and the lower index indicates the base-pair number labeled in the $5' \rightarrow 3'$ direction of strand 1. Note that this definition of H is independent of the labeling of the strands. Since the 5′ terminal is missing the backbone phosphate group, the first and last elements of the sequence are ignored in the definition of H. Therefore, the sequence used in the definition starts from P_1^2 instead of P_1^1 and ends with P_N^1 instead of P_N^2 (N being the total number of bases). The position of these P atoms then defines H via

$$H(P_1P_2P_3 \ldots P_N) = \sum_{i=1}^{N-3} H(P_iP_{i+1}P_{i+2}P_{i+3}),$$

in which each P_i is a point in the sequence discussed above, and

$$H(ABCD) = \frac{\overrightarrow{AB}}{|\overrightarrow{AB}|} \times \frac{\overrightarrow{CD}}{|\overrightarrow{CD}|} \cdot \frac{\overrightarrow{EF}}{|\overrightarrow{EF}|}.$$

In this last equation, the points A, B, C, D define the vectors $\overrightarrow{AB}$ and $\overrightarrow{CD}$ and the midpoints of these vectors, called E and F, in turn form the vector $\overrightarrow{EF}$.

III. RESULTS

A. Laser frequency scanning

The results of the laser frequency scans for DNA are shown in Fig. 2, which displays the percent fluctuation of all the different bond types for both B- and Z-DNA, as well as the change in energy ΔE, all as a function of the frequency. The fluctuations of the 34 different bond types (details of which are given in Table S2 in supplementary material)[101] were computed as the average of the fluctuations of a given bond over a 6 ps time frame centered around the laser pulse maximum. This was then compared to the average fluctuation of the same bond in an equilibrium simulation in the absence of any laser pulse. The ΔE the system at a given laser frequency was defined to be the total energy difference (highest value) for a system with the laser minus the total energy of an equilibrium system. The ΔE spectrum displays a complex structure of approximately 7 peaks of varying height, which coincide with high levels of bond fluctuations. Both in terms of the bond fluctuations and ΔE, the spectrum for B- and Z-DNA track each other quite well, with only relatively small shifts. Table I lists the peak frequencies and the associated bond types for the DNA structures. Notably, the two highest peaks are at ω = 1659, 1870 cm^{-1} and the associated fluctuating bond types are 22, 24, and 27 (former), and 22, 25, 26, 27, 28, and 29 for the latter. Clearly, the latter is associated with a larger number of bonds that are fluctuating. Moreover, bond type 28 is directly associated with a Watson-Crick H-bond. Given that ω = 1870 cm^{-1} targets this important bond and that peak height for both B- and Z-DNA are almost identical, we chose to conduct our laser melting simulations at this frequency.

B. Laser melting of the DNA helices

We now turn to the laser melting of the DNA helices. Both sets of simulations with pulse width σ = 3 ps and σ = 6 ps gave systematically consistent results, and here we report on the results obtained with the wider pulse. Selected results for σ = 3 ps are given in supplementary material.[101]

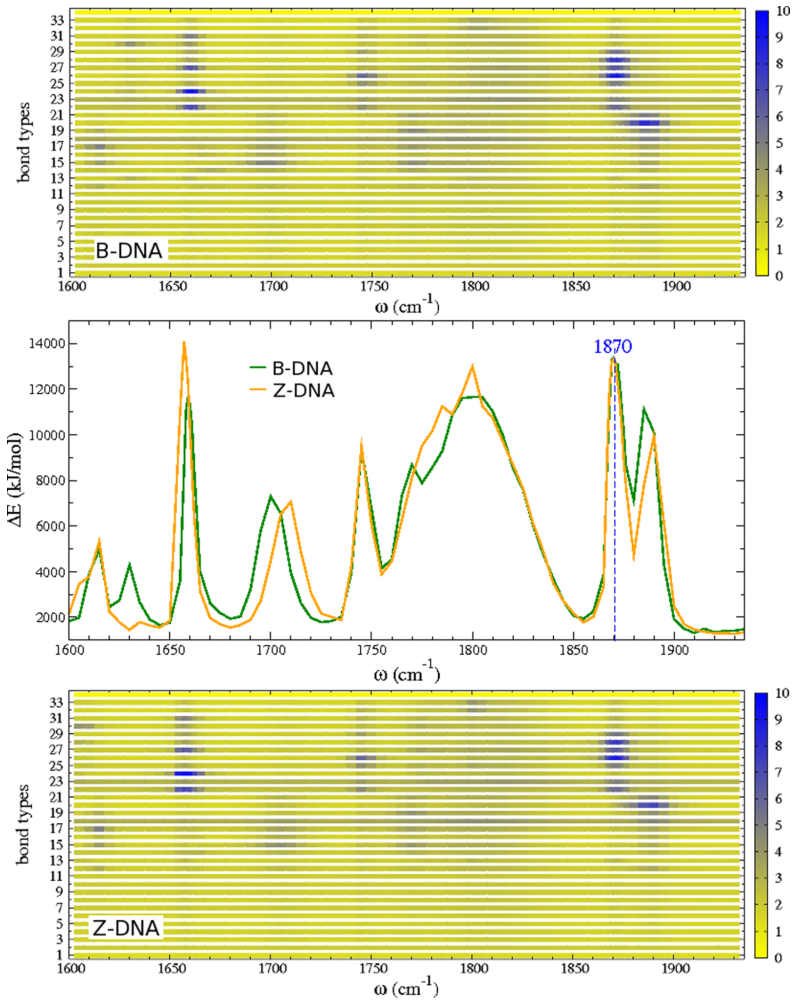


FIG. 2. Results of the laser frequency scan. The middle panel shows the maximum energy adsorption ΔE (kJ/mol) as a function of frequency ω (cm^{-1}) for B-DNA (green line) and Z-DNA (orange line). The dashed blue line indicates the location of the peak associated with $\omega = 1870$ cm^{-1}, which is the frequency chosen for the melting simulations. The top (bottom) panels indicate the fluctuations of the different bond types as a function of frequency for B-DNA (Z-DNA), respectively.

TABLE I. Absorption spectroscopy of B- and Z-DNA. RAS indicates relative absorption strength.

	B-DNA		Z-DNA	
ω (cm^{-1})	RAS	Excited bonds	RAS	Excited bonds
1615	Weak	C4–N3	Weak	The same as B-DNA
1630	Weak	N9–C4	None	
1659	Strong	N1–C2, C2–N3, C5–C6, N7–C5	Strong	The same as B-DNA
1700 (1710)	Medium	C2–O2, C4–N3	Medium	The same as B-DNA
1745	Medium	N1–C2, N3–C4, C5–C4, C6–N1	Medium	The same as B-DNA
1800	Strong	Almost base bonds	Strong	Almost base bonds
1870	Strong	N1–C2, N3–C4, C5–C4, C5–C6, C6–O6, C6–N1	Strong	The same as B-DNA
1885 (1890)	Medium	C5–C6	Medium	The same as B-DNA

Thus, Fig. 3 shows the time evolution of the laser pulse with $\sigma = 6$ ps and $t_0 = 20$ ps, for two intensities of the field, $E_0 = 3.0$ V/nm and 4.5 V/nm. For both field values, the energy absorption by B-DNA and Z-DNA is exactly the same (except for noise), and the total energy shows a small lag (about 2 ps) of the duplex response with respect to the perturbation. Fig. 4 shows snapshots of both DNA structures following the application of the laser pulse. The duplexes stay almost intact immediately following the application of the laser pulse (which reaches its maximum at 20 ps). This is quickly followed by melting and considerable disruption of the H-bonds of the structure. After the pulse dies out at approximately 40 ps, the structures continue struggling with the unfolding provoked by the laser.

Fig. 5 shows two quantitative measures of the laser melting process: the percentage of Watson-Crick H-bonds (WCHBs, top panels) and base-pair RMSDs (bottom panels) taken with respect to the initial structure, for various electric field amplitudes. These quantities start to show the effect of the pulse at about 11-12 ps, with a subsequent sharp drop in the number of H-bonds and a sharp increase in the base-pair RSMDs that signal the disruption of the structure. The disruption becomes maximum at about 24-25 ps. After that, the duplex partially reforms, depending on the magnitude of the field. After about 40 ps, most of the rapid changes are over and there are only small changes. Naturally, the amount of unfolding depends on the applied field strength, which varies from $E_0 = 3.0$ V/nm, where the duplexes almost completely recover, to $E_0 \geq 5.0$ V/nm, where the duplexes are denatured. Thus, for $E_0 = 4.0$ V/nm, about 43% of the H-bonds reform for B-DNA, and about 28% for Z-DNA; for $E_0 = 4.5$ V/nm, this drops to about 20% for B-DNA and 10% for Z-DNA. Note that in this regime, the percent WCHBs is always higher and the RMSD is smaller for B-DNA than for Z-DNA, reflecting the greater stability of the B-DNA structure (in zero or smaller fields B-DNA exhibits more end fraying than Z-DNA). These curves were averaged over 50 runs; we extended these simulations for up to 1 ns for only 10 of those 50 runs for each value of the electrical field (data not shown) and we found that essentially the values obtained at 100 ps remain constant or show a very slight increase. Thus, we can say that about 100 ps the system reaches a long-lived plateau. The WC hydrogen bond probabilities for each hydrogen bond as function of time for different magnitudes of the electric field at zero excess salt are shown in Figures S3 and S4 for B-DNA and Z-DNA, respectively, while the equilibrium distribution for various salts is shown in Figure S2.[101] The terminal fraying of B-DNA both in equilibrium and at small fields is apparent. In both duplexes, the disruption of the bonds at around 20 ps for $E_0 = 3.0$, 3.5 V/nm is followed by the re-formation of the bonds, but this re-bonding quickly degrades as the magnitude of the field increases.

Fig. 6 shows the base-stacking index for B-DNA and Z-DNA at zero and 4M NaCl (in addition to the 22 neutralizing Na^+). The initial base-stacking index is higher for B-DNA (0.25) than Z-DNA (0.15) reflecting the more stable base stacking of the right-handed duplex. This more favorable base stacking of B-DNA is maintained for all field strengths. Notice that the addition of 4M NaCl decreases base stacking for B-DNA while increasing it for Z-DNA. Finally, Fig. 7 shows the normalized handedness (nHDN) and BPOP of both DNA structures as a function of time for different field strengths at zero excess salt. The BPOP looks approximately

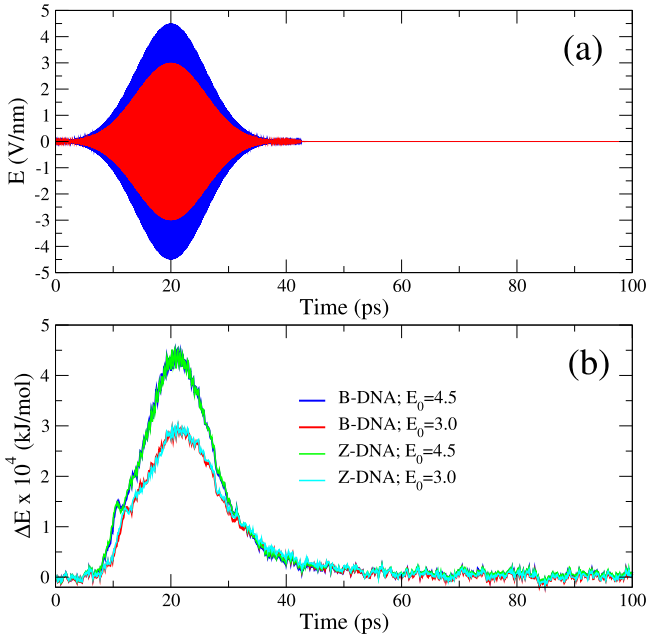


FIG. 3. Time dependence of the external electric field ($\omega = 1870$ cm^{-1}) (a) and energy absorption of B-DNA and Z-DNA duplexes (b) during a laser melting simulation. Panel (a) shows laser pulses with $E_0 = 3.0$ (red) and 4.5 (blue) V/nm.

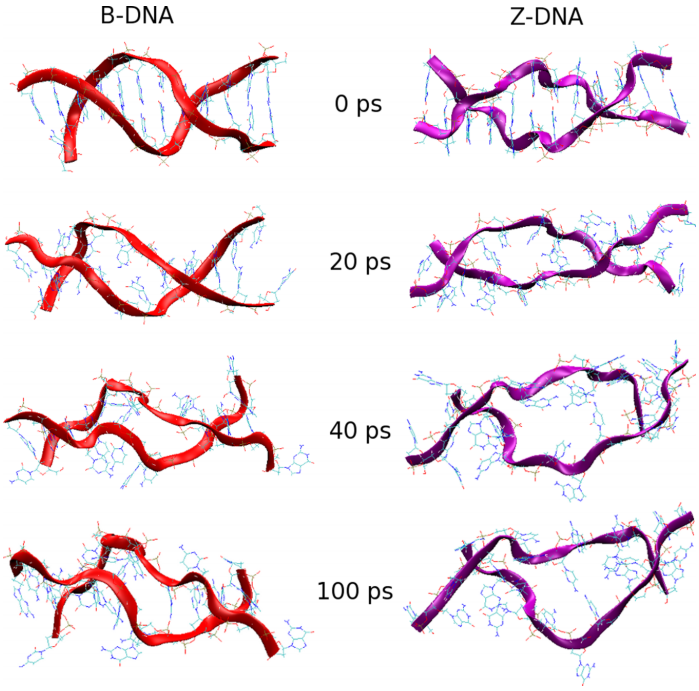


FIG. 4. Snapshots of typical conformational changes in B- and Z-DNA (zero excess salt) upon application of the laser pulse. Laser parameters are $\omega = 1870\ \mathrm{cm}^{-1}$ and $E_0 = 5.0$ V/nm, respectively.

as the mirror image of the WCHB images since, as the base pairs open, their WC H-bonds break. Lower values of BPOPs reflect higher stability of B-DNA with respect to Z-DNA. The nHDN is even more sensitive to the melting process, and is powerful to visually identify the more unwinded structure (the absolute value is shown here; handedness is positive for B-DNA and negative for Z-DNA). Fig. 8 gives a plot of the normalized handedness and base stacking index as

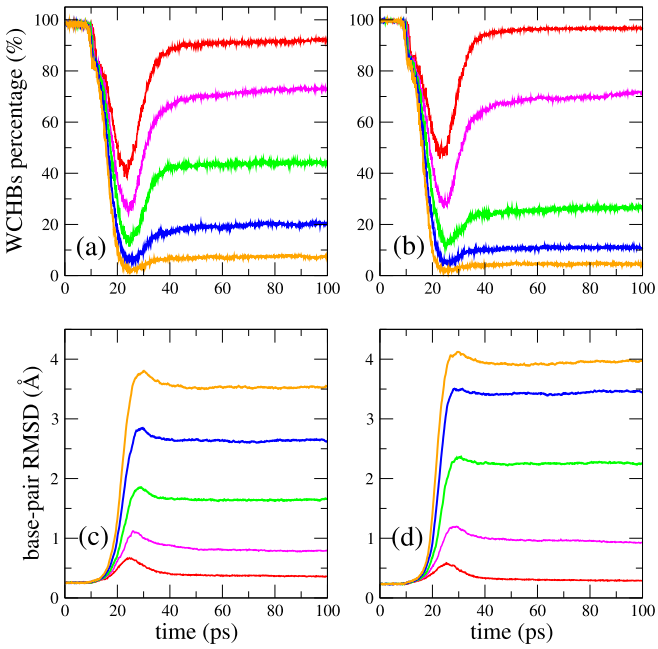


FIG. 5. Time evolution of the Watson-Crick H-bond (WCHB) percentage and base-pair RMSD of B-DNA and Z-DNA under the application of a laser pulse at zero excess salt. B-DNA is shown in (a) and (c), and Z-DNA in (b) and (d). Here, different colors represent different values of the electric field amplitude E_0 in V/nm: 3.0 (red); 3.5 (magenta); 4.0 (green); 4.5 (blue); and 5.0 (orange), respectively. The base-pair RMSD is taken with respect to the initial structure and represents an average over the ten middle base pairs of each DNA sequence. Data are averaged over 50 trajectories and E_0 is given in V/nm.

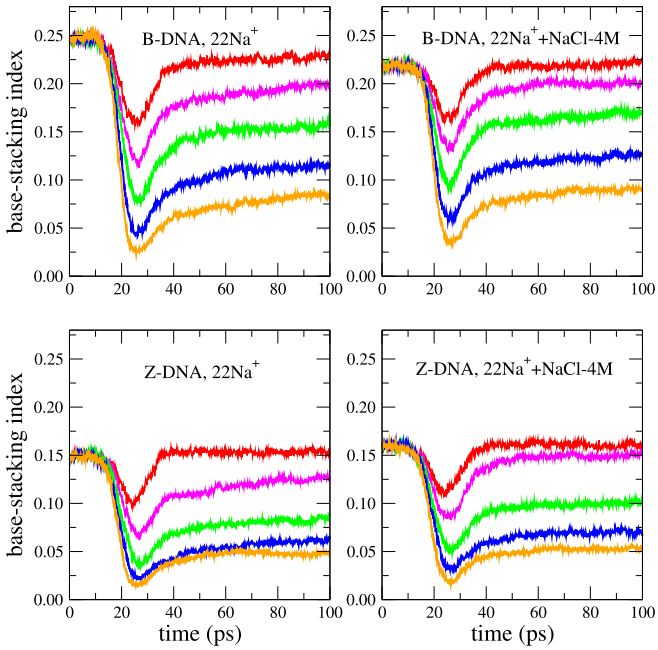


FIG. 6. Time evolution of the base-stacking index of B-DNA and Z-DNA under the application of a laser pulse at zero and 4M excess salt. Data are averaged over 50 trajectories and E_0 is given in V/nm. Color scheme for different values of E_0 is same as in Fig. 5.

a function of the magnitude of the electric field. B-DNA exhibits more favorable stacking throughout the entire range of electric field magnitudes, and B-DNA is considerably more resistant to unwinding by the laser pulse than Z-DNA.

We have also examined the behavior of all of these different structural measures as a function of excess salt concentration, as shown in Fig. 9. It is well known that increasing the salt concentration stabilizes the DNA duplexes, but stabilization is stronger in Z-DNA. Fig. 9 display this

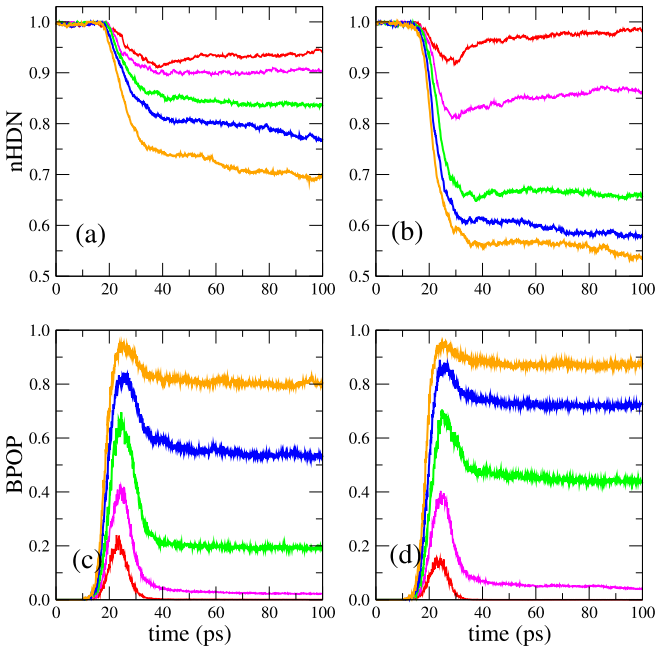


FIG. 7. Time evolution of the normalized handedness (nHDN) and base-pair opening probability (BPOP) of B-DNA and Z-DNA under the application of a laser pulse at zero excess salt. B-DNA is shown in (a) and (c), and Z-DNA in (b) and (d). Data is averaged over 50 trajectories and E_0 is given in V/nm. Color scheme for different values of E_0 is same as in Fig. 5.

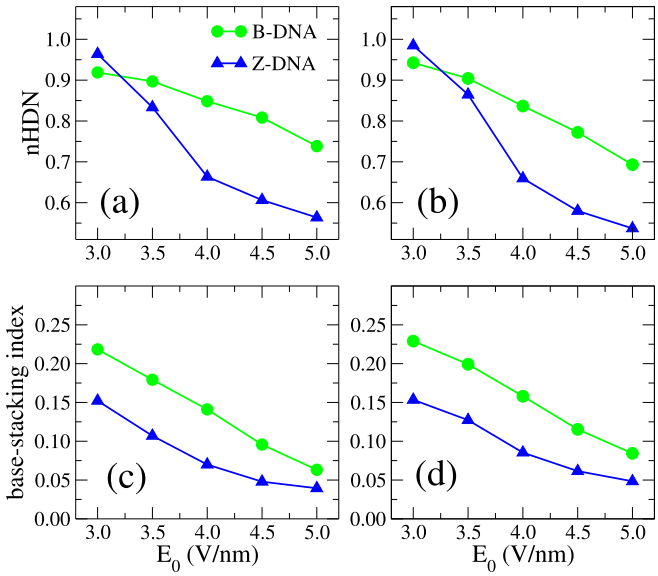


FIG. 8. Dependence of the normalized handedness ((a) and (b)) and base-stacking index ((c) and (d)) on the magnitude of the electric field E_0. The left ((a) and (c)) and right ((b) and (d)) panels are the data points averaged over the 42^{sd} ps and last 5 ps of the 100 ps simulations with zero excess salt.

general trend, with the curves growing or decaying (depending on the quantity) towards more stability and exhibiting a greater slope for Z-DNA.

Finally, how was the running time chosen? Clearly, after the perturbation is gone the structures would tend to fold back if enough time (*very* long time if they need to completely refold) has elapsed. We find that for each magnitude of the field, the quantities characterizing the structures reach a plateau somewhere between 60 and 100 ps, depending on the quantity, as shown in Figs. 3, 5, 6, and 7. These figures were computed as an average over 50 runs for each field magnitude. We extended simulations for only 10 of each set of 50 runs up to 1 ns, and indeed the values at 100 ps are the same or extremely close to the values at 1 ns. Therefore, the run times were chosen in the first segment of the plateau, which is indicative of much slower dynamics.

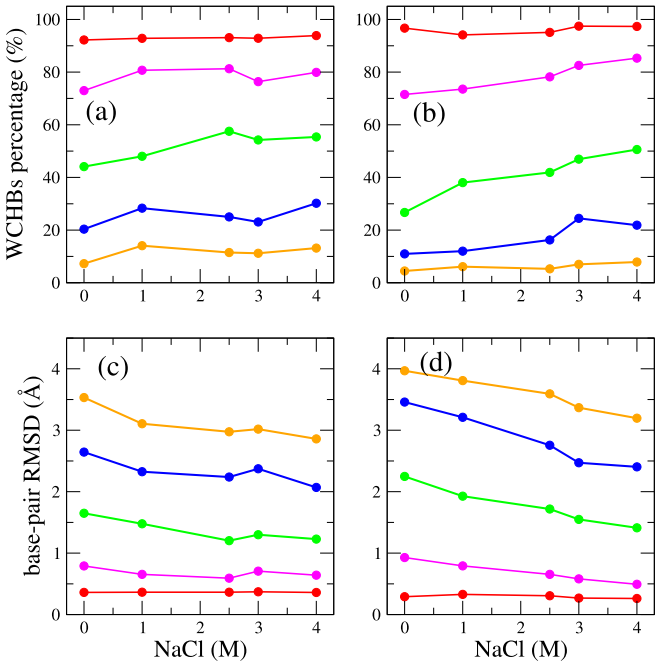


FIG. 9. Salt dependence of the WCHB percentages and the base-pair RMSD. Left panels: B-DNA; Right panels: Z-DNA. Data points are averaged over the last 5 ps of the 100 ps laser melting simulations. Color scheme for different values of E_0 is the same as in Fig. 5.

Thus, to build a comparative figure such as those in Fig. 8 one can choose the time where the pulse effectively goes to zero (at approximately 40 ps for σ = 6 ps) and add the "lag" in the molecular response (≃2 ps, such that quantities are compared at 42 ps). Alternatively, one could choose values somewhere in the plateau, say at 100 ps. In both cases, the same trends are observed, as shown in Fig. 8.

IV. DISCUSSION

A. Infrared laser pulse melting as a tool for discerning structural healing and correlated stabilities

The classical laser pulse simulated in this work is typical of a free-electron based laser,[1–5] with specific oscillation characteristics of a picosecond pulse structure, tunable wavelengths within the infrared regime, and a high photon density. Using this laser pulse, it has been possible to target specific bonds within a given structure and thereby study the unfolding of the structure. Specifically, this method was used to investigate the dissociation of amyloid fibrils into soluble monomers,[3–6] the break-up of a peptide-based nanotube,[8] and the dissociation of viruses.[7] Comparing healing and, possibly, relative stability of different structures via laser induced melting is clearly not a universal technique. Key to the application of the technique is the fact that the perturbation must affect both structures in the same way (i.e., the damage to the structures must be the same). As an example, we have used this technique previously to compare the responses of polyasparagine and polyglutamine amyloid aggregates. In this case, the results of the laser frequency scanning simulations resulted in an optimum frequency of ω = 1671 cm^{-1} that targeted the C═O main-chain bonds equally in both aggregates, thereby destabilizing the β sheets.[96]

To find out the relative stability of two DNA (or RNA) structures with the same sequence but with different secondary structure can be quite expensive computationally. In this work, we set out to determine what can be inferred about the relative stability of two duplexes with exactly the same sequence, d(5′-CGCGCGCGCGCG-3′)$_2$ but with two very different structures, right-handed B-DNA and left-handed Z-DNA, that are subject to a laser pulse. A meaningful comparison can be carried out because it is possible to find a laser pulse frequency, mainly ω = 1870 cm^{-1}, that generates the same resonant peak affecting the same bonds, especially bond C6–O6 involving a Watson-Crick H-bond, for both structures (Fig. 2). This results in *exactly the same energy absorption pattern for both B-DNA and Z-DNA* (Fig. 3). The perturbation is tuned to vary from minor disruption at small field strength to complete melting at high field strength, which allows for an extensive comparison of the responses of the two duplexes. Since a laser pulse is by definition a nonequilibrium process, its results cannot be translated into an equilibrium melting curve with the ensuing free energy estimates. However, in all cases, Z-DNA lags behind B-DNA in the recovery process. Although this can be associated with the kinetics of refolding, we believe that because this occurs in the whole range of field strengths, from small perturbations to complete melting, the process actually reflects the lower stability of Z-DNA with respect to B-DNA under the solvation conditions employed in these simulations. In that sense, Fig. 8 can be interpreted as the corresponding non-equilibrium melting curves of the duplexes.

Finally, it should be noted that we explore the partial melting and subsequent healing of DNA strands by means of MD techniques. Given the time scale problem and the necessity of simulating relatively small system size, this requires the use of high electric field intensities and a temperature rescaling of the solvent. However, even if it were currently feasible to simulate much larger systems over very much longer periods of time, we believe that the system would still very much follow similar generic behavior of the kind that we have uncovered with our simulations.

B. Relative stability of B-DNA and Z-DNA

Aside from uncovering the response of a biomolecular system to a fast, high intensity laser pulse, our simulations can viewed as representing a methodology in its own right, whose aim is to provide for a fast and cheap way of determining the relative stability of different DNA conformations. B-DNA is more stable than Z-DNA under physiological ionic strength and pH conditions. Experimentally, after a high nucleation barrier in a B- to Z-DNA transition, the propagation free energy of Z-DNA has been measured to be 0.66 kcal/mol per CG repeat at 0.1 NaCl,[97] and in general less than 2 kcal/mol per dinucleotide repeat.[98] Simulations of the base-extrusion zipper mechanism at 0.1M salt concentration in explicit water obtained 0.93 kcal/mol per dinucleotide,[80] while free energy maps with only neutralizing ions (no excess salt) resulted in 1.5 kcal/mol per dinucleotide.[81] The computational measures required enhanced sampling techniques, with definition of reaction coordinates.

A melting experiment is conceptually more straightforward, but it faces some challenges. First, B-DNA is more flexible than Z-DNA, and the configurational entropy difference is important in stabilizing B-DNA.[99] However, due to its inherent flexibility, B-DNA is more prone to terminal fraying (Fig. S5)[101] than more rigid Z-DNA, therefore B-DNA temperature melting is more susceptible to length effects than Z-DNA. While this does not affect long sequences, it could be a problem for relatively short oligomers. This is not an issue *in vivo*, where short DNA duplexes with four free ends are not encountered but could present a challenge for thermal stability measurements *in vitro* and *in silico*. Second, there is an evidence that Z-DNA is stabilized by higher temperature,[73,83–85] which complicates the interpretation of the traditional melting experiments. In addition, traditional melting simulations can be extremely taxing on the computational resources. We ran some tests to probe this. The melting temperature for d(5′-CGCGCGCGCGCG-3′)$_2$ in B-DNA form at 0.1M NaCl is estimated to be 331 K and 333 K by uMELT[49] and OligoCalc,[100] respectively. Preliminary tests show that well above 1 μs would be required to obtain the melting products for each temperature around the melting temperature, and because of noise, one needs tens of runs at each temperature interval to obtain a reliable value.

The nonequilibrium results presented here also lend themselves to the interpretation that B-DNA is more stable than Z-DNA. Some quantities are better than others to quantify this difference. Thus, for small fields the WC H-bonds, or the related variable, base-pair opening, might seem to slightly favor Z-DNA. This is because near equilibrium B-DNA frays more than Z-DNA, as shown in the equilibrium (zero field) fraying probability in Fig. S3.[101] However, slightly larger values of the field definitely disrupt the H-bonds more in Z-DNA than in B-DNA. Since stacking interactions play such a significant role in polynucleotide stability, the base-stacking index defined in this paper is a better measurement of stability, as it captures the favorable stacking of B-DNA compared to Z-DNA, from zero applied field all the way to the maximum value of the field. Also, the normalized handedness is related to the normalized helicity that some experiments and codes measure. Fig. 7 shows that the application of the laser pulse unwinds Z-DNA (reflected in the loss of absolute handedness) considerably more than B-DNA.

In summary, we have explored a novel, infrared laser pulse melting technique that is suitable for melting polynucleotides is a very fast time. In addition to the intrinsic interest of the melting and healing process, this technique could be used to qualitatively compare relative stabilities of different polynucleotide structures, especially considering that the reduced number of building blocks of nucleotides can readily give common paths for the application of the pulse. The nonequilibrium process does not give such quantities as relative free energies. Rather, it provides one with a "more-or-less" (stable) comparison. This, however, could be extremely useful for stability rankings in competing candidate structures that cannot be easily predicted with other tools, including non-standard secondary structures with noncanonical base pairs and mismatches. In addition, the method applies to very realistic, fully solvated, fully atomistic systems that, in principle, can faithfully reproduce experimental setups.

ACKNOWLEDGMENTS

We thank the NC State HPC Center for extensive computer support. This work has been supported by the National Science Foundation via Grant Nos. SI2-SSE-1148144 and 1534941, and also XSEDE Grant No. TG-MCB150114.

[1]D. Ozawa, H. Yagi, T. Ban, A. Kameda, T. Kawakami, H. Naiki, and Y. Goto, J. Biol. Chem. **284**, 1009 (2009).
[2]H. Yagi, D. Ozawa, K. Sakuri, T. Kawakami, H. Kuyama, O. Nishimura, T. Shimanouchi, R. Kuboi, H. Naiki, and Y. Goto, J. Biol. Chem. **285**, 19660 (2010).
[3]T. Kawasaki, J. Fujioka, T. Imai, and K. Tsukiyama, Protein J. **31**, 710 (2012).
[4]T. Kawasaki, J. Fujioka, T. Imai, K. Torigoe, and K. Tsukiyama, Lasers Med. Sci. **29**, 1701 (2014).
[5]T. Kawasaki, T. Yaji, T. Imai, T. Ohta, and K. Tsukiyama, Am. J. Anal. Chem. **5**, 384 (2014).
[6]V. H. Man, P. Derreumaux, M. S. Li, C. Roland, C. Sagui, and P. Nguyen, J. Chem. Phys. **143**, 155101 (2015).
[7]V. H. Man, N.-T. Van-Oanh, P. Derreumaux, M. S. Li, C. Roland, C. Sagui, and P. H. Nguyen, "Picosecond infrared laser-induced all-atom nonequilibrium molecular dynamics simulation of dissociation of viruses," Phys. Chem. Chem. Phys. (published online).
[8]V. Man, P. Truong, P. Derreumaux, M. Li, C. Roland, C. Sagui, and P. Nguyen, Phys. Chem. Chem. Phys. **17**, 27275 (2015).
[9]J.-F. Mergny and L. Lacroix, Oligonucleotides **13**, 515 (2003).
[10]R. Thomas, Gene **135**, 77 (1993).
[11]J. Jin, K. J. Breslauer, R. A. Jones, and B. L. Gaffney, Science **250**, 543 (1990).
[12]C. Hardin, E. Henderson, T. Watson, and J. K. Prosser, Biochemistry **30**, 4460 (1991).
[13]J.-L. Mergny and J.-C. Maurizot, ChemBioChem **2**, 124 (2001).
[14]J. G. Duguid, V. A. Bloomfield, J. M. Benevides, and G. J. Thomas, Biophys. J. **71**, 3350 (1996).
[15]L. Movileanu, J. M. Benevides, and G. J. Thomas, Nucleic Acids Res. **30**, 3767 (2002).
[16]J.-L. Mergny, A.-T. Phan, and L. Lacroix, FEBS Lett. **435**, 74 (1998).
[17]P. Cahen, M. Luhmer, C. Fontaine, C. Morat, J. Reisse, and K. Bartik, Biophys. J. **78**, 1059 (2000).
[18]D. Poland and H. A. Scheraga, J. Chem. Phys. **45**, 1456 (1966).
[19]D. Poland, Biopolymers **13**, 1859 (1974).
[20]M. Fixman and J. J. Freire, Biopolymers **16**, 2693 (1977).
[21]M. Y. Azbel, Phys. Rev. A **20**, 1671 (1979).
[22]K. J. Breslauer, R. Frank, H. Bloecker, and L. A. Marky, Proc. Natl. Acad. Sci. U. S. A. **83**, 3746 (1986).
[23]M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[24]T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, R44 (1993).
[25]T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, 684 (1993).
[26]T. Dauxois and M. Peyrard, Phys. Rev. E **51**, 4027 (1995).
[27]N. Sugimoto, S.-I. Nakano, M. Yoneyama, and K.-I. Honda, Nucleic Acids Res. **24**, 4501 (1996).
[28]J. Santa Lucia, Proc. Natl. Acad. Sci. U. S. A. **95**, 1460 (1998).
[29]R. Owczarzy, P. M. Vallone, F. Gallo, T. M. Paner, M. J. Lane, and A. S. Benight, Biopolymers **44**, 217 (1998).
[30]A. Campa and A. Giansanti, Phys. Rev. E **58**, 3585 (1998).
[31]T. V. Chalikian, J. Volker, G. E. Plum, and K. J. Breslauer, Proc. Natl. Acad. Sci. U. S. A. **96**, 7853 (1999).
[32]V. Ivanov, Y. Zeng, and G. Zocchi, Phys. Rev. E **70**, 051907 (2004).
[33]D. Poland, Biopolymers **73**, 216 (2004).
[34]C. Richard and A. J. Guttmann, J. Stat. Phys. **115**, 925 (2004).
[35]S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, Proc. Natl. Acad. Sci. U. S. A. **83**, 9373 (1986).
[36]J. A. Jaeger, D. H. Turner, and M. Zuker, Proc. Natl. Acad. Sci. U. S. A. **86**, 7706 (1989).
[37]M. Zuker, Science **244**, 48 (1989).
[38]K. B. Hall and L. W. W. McLaughlin, Biochemistry **30**, 10606 (1991).
[39]L. Ratmeyer, R. Vinayak, Y. Zhong, G. Zon, and W. D. Wilson, Biochemistry **33**, 5298 (1994).
[40]N. Sugimoto, K.-I. Honda, and M. Sasaki, Nucleosides Nucleotides **13**, 1311 (1994).
[41]N. Sugimoto, M. Katoh, S.-I. Nakano, T. Ohmichi, and M. Sasaki, FEBS Lett. **354**, 74 (1994).
[42]M. J. Doktycz, M. Morris, S. J. Dormady, K. L. Beattie, and B. Jacobson, J. Biol. Chem. **270**, 8439 (1995).
[43]J. Santa Lucia, H. T. Allawi, and P. A. Seneviratne, Biochemistry **35**, 3555 (1996).
[44]J. Santa Lucia and D. H. Turner, Biopolymers **44**, 309 (1997).
[45]T. Xia, J. Santa Lucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner, Biochem. **37**, 14719 (1998).
[46]D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt, and D. H. Turner, RNA **5**, 1458 (1999).
[47]D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, J. Mol. Biol. **288**, 911 (1999).
[48]M. Zuker, Nucleic Acids Res. **31**, 3406 (2003).
[49]Z. Dwight, R. Palais, and C. T. Wittwer, Bioinformatics **27**, 1019 (2010).
[50]A. H. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel, and A. Rich, Nature **282**, 680 (1979).
[51]A. Nordheim and A. Rich, Proc. Natl. Acad. Sci. U. S. A. **80**, 1821 (1983).
[52]A. Rich, A. Nordheim, and A. H. Wang, Annu. Rev. Biochem. **53**, 791 (1984).
[53]G. Schroth, P. Chou, and P. J. Ho, J. Biol. Chem. **267**, 11846 (1992).
[54]L. F. Liu and J. C. Wang, Proc. Natl. Acad. Sci. U. S. A. **84**, 7024 (1987).
[55]D. Oh, Y. Kim, and A. Rich, Proc. Natl. Acad. Sci. U. S. A. **99**, 16666 (2002).
[56]H. J. Lipps, Cell **32**, 435 (1983).
[57]F. Lancillotti, M. Lopez, C. Alonso, and B. Stollar, J. Cell Biol. **100**, 1759 (1985).
[58]A. Herbert and A. Rich, Genetica **106**, 37 (1999).
[59]E. Kmiec, K. Angelides, and W. Holloman, Cell **40**, 139 (1985).

[60]J. Blaho and R. Wells, J. Biol. Chem. **262**, 6082 (1987).
[61]A. Jaworski, W. Hsieh, J. Blaho, J. Larson, and R. Wells, Science **238**, 773 (1987).
[62]T. Schwartz, M. Rould, K. Lowenhaupt, A. Herbert, and A. Rich, Science **284**, 1841 (1999).
[63]A. Vinogradov, Nucleic Acids Res. **31**, 1838 (2003).
[64]P. Champ, S. Maurice, J. Vargason, T. Champ, and P. Ho, Nucleic Acids Res. **32**, 6501 (2004).
[65]G. Wang, L. Christensen, and K. Vasquez, Proc. Natl. Acad. Sci. U. S. A. **103**, 2677 (2006).
[66]A. Herbert, M. Schade, K. Lowenhaupt, J. Alfken, T. Schwartz, L. S. Shlyakhtenko, Y. L. Lyubchenko, and A. Rich, Nucleic Acids Res. **26**, 3486 (1998).
[67]J. D. Kahmann, D. A. Wecking, V. Putter, K. Lowenhaupt, Y.-G. Kim, P. Schmieder, H. Oschkinat, A. Rich, and M. Schade, Proc. Natl. Acad. Sci. U. S. A. **101**, 2712 (2004).
[68]S. C. Ha, J. Choi, H.-Y. Hwang, A. Rich, Y.-G. Kim, and K. K. Kim, Nucleic Acids Res. **37**, 629 (2009).
[69]C.-X. Wu, S.-J. Wang, G. Lin, and C.-Y. Hu, Fish Shellfish Immunol. **28**, 783 (2010).
[70]L. Yang, S. Wang, T. Tian, and X. Zhou, Curr. Med. Chem. **19**, 557 (2012).
[71]M. A. Fuertes, V. Cepeda, C. Alonso, and J. M. Pérez, Chem. Rev. **106**, 2045 (2006).
[72]S. Harvey, Nucleic Acids Res. **11**, 4867 (1983).
[73]S. Goto, Biopolymers **23**, 2211 (1984).
[74]A. Ansevin and A. Wang, Nucleic Acids Res. **18**, 6119 (1990).
[75]W. Lim and Y. Feng, Biophys. J. **88**, 1593 (2005).
[76]W. Saenger and U. Hienemann, FEBS Lett. **257**, 223 (1989).
[77]S. C. Ha, K. Lowenhaupt, A. Rich, Y. G. Kim, and K. K. Kim, Nature **437**, 1183 (2005).
[78]R. Elber, A. Ghosh, and A. Cardenas, Acc. Chem. Res. **35**, 396 (2002).
[79]M. A. Kastenholz, T. U. Schwartz, and P. H. Hünenberger, Biophys. J. **91**, 2976 (2006).
[80]J. Lee, Y. Kim, K. Kim, and C. Seok, J. Phys. Chem. B **114**, 9872 (2010).
[81]M. Moradi, V. Babin, C. Roland, and C. Sagui, Nucleic Acids Res. **41**, 33 (2013).
[82]F. Pan, C. Roland, and C. Sagui, Nucleic Acids Res. **42**, 13981 (2014).
[83]J. H. van de Sande and T. M. Jovin, EMBO J. **1**, 115 (1982).
[84]D. J. Patel, S. A. Kozlowski, A. Nordheim, and A. Rich, Proc. Natl. Acad. Sci. U. S. A. **79**, 1413 (1982).
[85]K. B. Roy and H. T. Miles, Biochem. Biophys. Res. Commun. **115**, 100 (1983).
[86]W. L. Jorgensen, J. Chandrasekhar, J. Madura, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[87]H. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[88]A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, Biophys. J. **92**, 3817 (2007).
[89]I. S. Joung and T. E. Cheatham, J. Phys. Chem. B **112**, 9020 (2008).
[90]T. A. Darden, D. M. York, and L. G. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[91]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[92]H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[93]R. W. Hockney, S. P. Goel, and J. Eastwood, J. Comput. Phys. **14**, 148 (1974).
[94]G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[95]M. Moradi, V. Babin, C. Roland, T. Darden, and C. Sagui, Proc. Natl. Acad. Sci. U. S. A. **106**, 20746 (2009).
[96]Y. Zhang, V. H. Man, C. Roland, and C. Sagui, "Amyloid properties of asparagine and glutamine in prion-like proteins," ACS Chem. Neurosci. (published online).
[97]L. Peck and J. Wang, Proc. Natl. Acad. Sci. U. S. A. **80**, 6206 (1983).
[98]P. Ho, Proc. Natl. Acad. Sci. U. S. A. **91**, 9549 (1994).
[99]K. K. Irikura, B. Tidor, B. R. Brooks, and M. Karplus, Science **229**, 571 (1985).
[100]W. A. Kibbe, Nucleic Acids Res. **35**, W43 (2007).
[101]See supplementary material at http://dx.doi.org/10.1063/1.4945340 for tables of bond types and WCHB index, and additional figures.

Chapter 4

Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats

Feng Pan, Viet Hoang Man, Christopher Roland and Celeste Sagui. (2017) Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats. *Biophysical Journal 113*, 19-36.

Biophysical Journal
Article

Biophysical Society

Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats

Feng Pan,[1] Viet Hoang Man,[1] Christopher Roland,[1] and Celeste Sagui[1,*]
[1]Department of Physics, North Carolina State University, Raleigh, North Carolina

ABSTRACT CAG trinucleotide repeats are known to cause 10 late-onset progressive neurodegenerative disorders as the repeats expand beyond a threshold, whereas GAC repeats are associated with skeletal dysplasias and expand from the normal five to a maximum of seven repeats. The TR secondary structure is believed to play a role in CAG expansions. We have carried out free energy and molecular dynamics studies to determine the preferred conformations of the A-A noncanonical pairs in $(CAG)_n$ and $(GAC)_n$ trinucleotide repeats ($n = 1, 4$) and the consequent changes in the overall structure of the RNA and DNA duplexes. We find that the global free energy minimum corresponds to A-A pairs stacked inside the core of the helix with anti-anti conformations in RNA and (high-anti)-(high-anti) conformations in DNA. The next minimum corresponds to anti-syn conformations, whereas syn-syn conformations are higher in energy. Transition rates of the A-A conformations are higher for RNA than DNA. Mechanisms for these various transitions are identified. Additional structural and dynamical aspects of the helical conformations are explored, with a focus on contrasting CAG and GAC duplexes. The neutralizing ion distribution around the noncanonical pairs is described.

INTRODUCTION

Trinucleotide repeats (TRs) belong to the family of simple sequence repeats (SSRs), that comprises all sequences with core motifs of one to six (and even 12) nucleotides that are repeated up to 30 times (and more for pathological cases) (1). SSRs exhibit dynamic mutations that do not follow Mendelian inheritance (which asserts that mutations in a single gene are stably transmitted between generations). In the 1990s, scientists discovered that inherited neurological disorders known as "anticipation diseases", where the age of the onset of the disease decreased and its severity increased, were caused by the intergenerational expansion of SSRs (2–5). After a certain threshold in the length of the repeated sequence, the probability of further expansion and the severity of the disease increases with the length of the repeat. To date, ~30 DNA expandable SSR diseases have been identified and the list is expected to grow (6,7). In particular, the dynamic mutations in human genes associated with TRs cause severe neurodegenerative and neuromuscular disorders, known as trinucleotide (or triplet) repeat expansion diseases (TREDs) (3,8–10). The expansion is believed to be primarily caused by some sort of slippage during DNA replication, repair, recombination, or transcription (5–7,11–15). Cell toxicity and death have been linked to the atypical conformation and functional changes of the transcripts and, when TRs are present in exons, of the translated proteins (6,16–25).

Of all the TRs, CAG repeats give rise to the largest group of neurodegenerative diseases. CAG repeats in the 5′-UTR of the gene PPP2R2B cause spinocerebellar ataxia type 12, whereas CAG repeats in the exon part of various genes cause another nine late-onset, progressive neurodegenerative disorders, including Huntington's disease, dentatorubral-pallidoluysian atrophy, spinal and bulbar muscular atrophy, and several spinocerebellar ataxias. These disorders are also known as polyglutamine (polyQ) diseases (26), because although CAG repeats could likely encode three different amino acid repeats depending on the reading frame (codons CAG, AGC, and GCA would code for polyQ, polyS, and polyA, respectively), the CAG expansions in these genes only lead to polyQ expansions. These polyQ diseases, like other TREDs, are caused by expansions greater than a given threshold (26). For instance, in Huntington's disease, the normal polyQ (or CAG repeat) length is 10–34 repeats, and pathological lengths are 36–250 repeats. Although each disease has a different pathology, they all share a common feature: the formation of polyQ aggregates (27), where the mature fibrils display

Submitted April 4, 2017, and accepted for publication May 26, 2017.
*Correspondence: sagui@ncsu.edu
Editor: Wilma Olson.
http://dx.doi.org/10.1016/j.bpj.2017.05.041

cross-β conformations (28–34); and the eventual neuronal death.

Interestingly, after the discovery of the CAG repeats and their relation to neurological disease, it was found that the GAC trinucleotide is also involved in a completely different class of diseases from the known TREDs. These diseases are caused by a very small change in the repeat number, and therefore do not qualify as TREDs. In particular, the human gene for cartilage oligomeric matrix protein exhibits a $(GAC)_5$ repeat. Expansion by one repeat causes multiple epiphyseal dysplasia, whereas expansion by two repeats or, alternatively, deletion by one repeat, causes pseudoachondroplasia (35). The structure of the various duplexes seems to strongly depend on the pH of the solution and the ionic strength (36). Whereas the CAG trinucleotide leads to expansion, the GAC trinucleotide does not (except for, at most, two extra repeats).

Although the mechanisms underlying TREDs are believed to be extremely complex, simple and robust trends beyond the repeat threshold have been identified, such as the correlation between the repeat length and the probability of further expansion and increased severity of the disease. Another important breakthrough has been the recognition that stable atypical DNA secondary structure in the expanded repeats is "a common and causative factor for expansion in human disease" (37). In addition, mutant transcripts also contribute to the pathogenesis of TREDs through toxic RNA gain-of-function (6,16–21). In this mechanism, the RNA TRs sequester proteins that are generally involved in pre-mRNA splicing and regulation. Thus, a first step toward the understanding of these diseases involves the structural characterization of the atypical DNA and RNA structures. Because there is experimental consensus that the most typical DNA and RNA TR secondary structures, at least in the initial stages of expansion, are hairpins whose stem lengths can wildly vary (21,38–40), a characterization of the mismatched helical duplexes forming the stems provides a foundation toward a structural understanding of the TR atypical secondary structures.

At present, little is known about the atomic structure and associated dynamics of trinucleotide CAG and GAC repeats. To date, experimental investigations have only considered CAG repeats in RNA; there are no experimental studies with atomic resolution of GAC repeats for DNA or RNA; and, perhaps most importantly, there are no experimental atomic resolution experiments of CAG repeats in DNA. Given that the expansions that characterize TRs originate at the DNA level, a structural understanding of these repeats at the atomic level in DNA is particularly important. Also, as described above, GAC repeats and CAG repeats behave in radically different ways in a biological context, and teasing out the structural differences between these two repeats both in the RNA and DNA context may help in the elucidation of their different behaviors with respect to expansion diseases.

Here we briefly review the experimental results for CAG repeats in RNA. The x-ray RNA-CAG duplex crystal structures include the following sequences: the sequence r(5′-GG-$(CAG)_2$-CC)$_2$ (41), and the sequence r(5′-UUG GGC-$(CAG)_3$-GUCC)$_2$ (42,43). This last sequence was also analyzed via NMR (43). The first study found that the duplexes favor the A-RNA form and that the A-A noncanonical pairs are in the anti-anti conformation. In the second sequence, both anti-anti and syn-anti A-A conformations were observed: the A-A pairs in the internal CAG always displayed the anti-anti conformation, whereas one (43) or two (42) of the terminal A-A pairs displayed the anti-syn conformation. These results are in general agreement with the complementary molecular dynamics (MD) simulations (42). Thus, whereas the anti-anti conformation (with fluctuations) for an internal A-A pair in a TR is common to the three studies, the nature of the anti-syn conformations is not clearly established. This is because these conformations occur in the terminal A-A pairs of r(5′-UUGGGC-$(CAG)_3$-GUCC)$_2$, where the A-A pairs are flanked by CC/GG steps. Using high-level ab initio calculations, it has been shown that CC/GG steps are the least stable of the 10 dinucleotide steps, with well-separated energies (44) from the other dinucleotide steps. Because these steps are never present in a genuine $(CAG)_n$ TR (which only exhibits GpC steps), it is clear that their presence could bias the conformation of the adjacent A-A pairs.

In addition to these RNA-CAG studies, there is one molecular dynamics study for CAG repeats in DNA (45). This study uses a sequence that is more relevant to the expanded disease, mainly $d(CAG)_6$. According to the conclusions of this study, the A-A mismatch in DNA behaves in exactly the opposite way than its RNA counterpart: it disfavors the anti-anti and the anti-(+syn) conformations and adopts the (−syn)-(−syn) conformations, resulting in a local Z-form around the mismatch (45). These results are intriguing and raise questions as to the true nature of the A-A mismatches in DNA-CAG.

In this work, we present a unified and comparative description of the nucleic acid duplexes for both DNA and RNA for both CAG and GAC trinucleotide repeats based on MD simulations and free energy calculations. A review of the field of MD simulations of nucleic acids is beyond the scope of this work, and the reader is referred to the authoritative reviews presented in the literature (46–48). Out of the four possible DNA/RNA CAG/GAC cases, there is experimental data only for RNA-CAG. We therefore begin by making the connection with this experimental data through an explicit investigation of a specific sequence employed in these studies and then move on to a four-trinucleotide repeat duplex. After that, we consider the other three cases—specifically RNA-GAC, DNA-CAG, and DNA-GAC. In particular, we present results corresponding to both free energy calculations and regular

1 μs MD simulations of single mismatch duplexes (5′-CCG-CAG-CGG-3′)$_2$ and (5′-GGC-GAC-GCC-3′)$_2$ both for RNA and DNA, and for regular 1-μs MD simulations of four-trinucleotide repeat duplexes (5′-(CAG)$_4$-3′)$_2$ and (5′-(GAC)$_4$-3′)$_2$. For each of the four duplexes, the free energy calculations involve two maps, each computed with a different pair of collective variables. The eight resulting free energy maps allow us to identify and rank the minima corresponding to the different A-A mismatch conformations. We also identify mechanisms of transition of the A-A mismatches toward the global free energy minimum, and link these mechanisms to paths over the free energy maps. We complete the work with a characterization of the neutralizing Na^+ ion distributions around the mismatches. Strictly speaking, the noncanonical A-A pairs in RNA are not "mismatches", because RNA is not necessarily self-complementary. However, because we are considering both DNA and RNA in their duplex form, we will call these noncanonical basepairs "mismatches" for simplicity.

MATERIALS AND METHODS

The sequences we investigated are shown in Fig. 1. For both DNA and RNA, we ran regular MD simulations for the four sequences (with various combinations of χ-angles for the mismatches) up to 1 μm. We used the sequences with a single mismatch, (5′-CCG-CAG-CGG)$_2$ ("CAG" for short), and the complementary sequence (5′-GGC-GAC-GCC)$_2$ ("GAC" for short) to determine the most favorable A-A mismatch conformation via the computation of free energy maps is described below. The initial conformations for the regular 1 μs MD simulations for the trinucleotide repeats (5′-(CAG)$_4$-3′)$_2$ (short-hand notation (CAG)$_4$) and (5′-(GAC)$_4$-3′)$_2$ (short-hand notation (GAC)$_4$) made use of the four possible combinations of A-A conformations, as described below.

The simulations were carried out using the PMEMD module of the AMBER v.14 (49) software package with the ff12SB force field with parameters ff99BSC0 (50) for DNA and ff99BSC0+Yildirim's χ-modification (51) for RNA. The TIP3P model (52) was used for the water molecules, along with the standard parameters for ions as in the AMBER force fields (53). The long-range Coulomb interaction was evaluated by means of the particle-mesh Ewald method (54) with a 9 Å cutoff and an Ewald coefficient of 0.30768. Similarly, the van der Waals interactions were calculated by means of a 9 Å atom-based nonbonded list, with a continuous correction applied to the long-range part of the interaction. The production runs were generated using the leap-frog algorithm with a 1 fs timestep with Langevin dynamics, and a collision frequency of 1 ps^{-1}. Conformations were saved every picosecond of the simulation. The SHAKE algorithm was applied to all bonds involving hydrogen atoms. We also emphasize here that our study is based on classical MD, and quantum effects are not accounted for. For instance, recent quantum chemistry investigations point to a possible transition of the A-base into a rare imino tautomeric form based on a double proton transfer giving rise to local wobbling between hydrogen bonds, which in turn can influence the dissociation rate of the mismatch (55,56).

To calculate the free energy maps, we made use of the adaptively biased molecular dynamics (ABMD) method (57,58), which has been implemented for PMEMD in AMBER v.16 (59). ABMD is a proven, elegant, nonequilibrium MD method that belongs to the general category of umbrella sampling methods with a history-dependent biasing potential, a method that, in the long-time limit, reproduces the negative of free energy. The free energy—or potential of mean force—is calculated as a function of

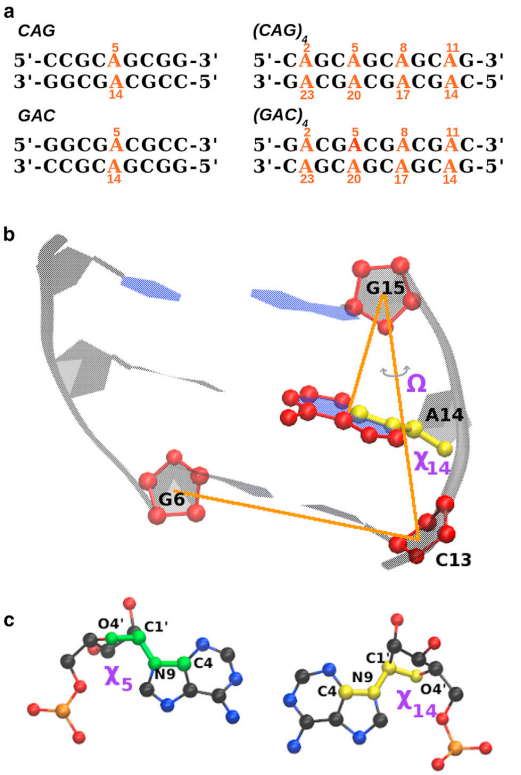


FIGURE 1 (*a*) Shown here are the sequences considered in this study (for both DNA and RNA). (*b*) Shown here is the schematic view of the center-of-mass pseudodihedral angle Ω (for A14 in CAG) and χ_{14}. (*c*) Given here is the view of χ_5 and χ_{14}. To see this figure in color, go online.

one or more collective variables, which must be carefully chosen to reflect the underlying physics of the problem. ABMD has been implemented with multiple walkers (both noninteracting (60) and interacting walkers, with the latter interacting by means of selection algorithm (61)), replica exchange molecular dynamics (62), and well-tempered extensions (63). It is now a mature method that has been applied to a variety of biomolecular systems including small peptides (57,58), sugar puckering (64), polyproline systems (65–69), polyglutamine systems (70), DNA systems (71), and others.

We computed free energy maps for a single mismatch in the CAG and GAC sequences for both DNA and RNA. The free energy of these mismatches was calculated as function of three main collective variables, chosen to reflect the structure of the mismatches and to make direct contact with a previous study of RNA-CAG (42). We define 1): Ω as the center-of-mass pseudodihedral angle, which is defined using the centers-of-mass of four atom groups: G6(C1′, C2′, C3′, C4′, O4′), C13(C1′, C2′, C3′, C4′, O4′), G15(C1′, C2′, C3′, C4′, O4′), and A14(N1, C2, N3, C4, C5, C6, N6, N7, C8, N9). This variable describes the base unstacking of A with respect to the helical axis 2); χ_5 as the glycosyl torsion angle χ of A5, namely the dihedral angle O4′-C1′-N9-C4; and 3) χ_{14}, which represents the χ-angle of A14. A schematic view of these collective variables is shown in Fig. 1. With these variables, we constructed two phase diagrams, (Ω, χ_{14}) and (χ_5, χ_{14}). For the first diagram, we found that if we choose χ_5 in the anti range, A5 stays in its anti conformation for all calculations, so the first diagram explores anti-anti and anti-syn conformations, and also whether they are stacked inside the helical core. By construction, therefore, the first

diagram cannot explore syn-syn conformations. The (χ_5, χ_{14}) diagram, on the other hand, can explore all options of χ (anti-anti, anti-syn, syn-anti, and syn-syn) but is degenerate with respect to Ω, i.e., it cannot tell whether the bases are inside the helix or have flipped out. A given free energy landscape was deemed to have converged when both the position and differences in the free energy values of the minima remain approximately constant as further ABMD cycles are performed. For the RNA (DNA), ~150 ns (180 ns) are required for each of the (χ_5, χ_{14}) maps; the DNA (Ω, χ_{14}) are much harder to converge, and results are shown simply after ~220 ns.

Initial conformations for both MD and free energy calculations were obtained as follows. We first solvated the initial structures and then followed this up with a sequence of ABMD runs of ever finer resolution. The details are as follows. First, we created the duplexes with the four possible combinations of χ-angle for the A-A mismatch: anti-anti, anti-syn, syn-anti, and syn-syn. These were then solvated in an octahedral box with 16 neutralizing Na^+ ions as in previous work (72), with a distance of at least 10 Å between the duplexes and walls of the box. The box was then filled with a suitable number of waters. The system was then minimized: first keeping the nucleic acid and ions fixed; then, allowing them to move. Subsequently, the temperature was gradually raised using constant volume simulations from 0 to 300 K over 50 ps, followed by a further 50 ps run. Then a 100 ps run at constant volume was used to gradually reduce the restraining harmonic constants for nucleic acids and ions. This was followed by a 1.0 ns constant pressure run, with the χ-angles of A5 and A14 slightly restrained so that these retain their initial anti- or syn- conformation. We took random conformations from the last 200 ps of these runs as the initial conformations for both the ABMD and MD runs. In particular, for the (Ω, χ_{14}) phase diagrams (where the collective variables are angles associated with A14), we picked four structures from A5(anti)-A14(anti) and four from A5(anti)–A14(syn), because the point of this calculation was to assess the anti-syn flipping of A14 (which is completely equivalent to A5). Because DNA is less stable than RNA, a small restraint was applied to χ_5 for the (Ω, χ_{14}) phase diagram. For the (χ_5, χ_{14}) phase diagrams, we picked two structures from each of the four runs (anti-anti, anti-syn, syn-anti, and syn-syn).

Multiple walker ABMD runs at constant volume and 300 K were carried out with eight replicas. The first ABMD simulation was for 20.0 ns with parameters $\tau_F = 1$ ps and $4\Delta\xi = 0.5$ radians. This simulation provided for a rough estimate of the free energy landscape over the relevant parameter space. We then followed this up with a finer 100-ns well-tempered ABMD simulation (parameters $\tau_F = 1$ ps, $4\Delta\xi = 0.2$ radians, pseudo-temperature 10,000 K). For these runs, the total number of hydrogen bonds in neighboring CG Watson-Crick basepairs were slightly restrained to be six using a 1.0 kcal/mol harmonic constraint. This was used to avoid the large-scale twisting of the whole structure during the long simulations. This constraint, however, was chosen to be flexible enough so as to readily allow for the relevant anti-syn transitions. Finally, a slower and smoother flooding to refine the landscapes was carried out with parameters $\tau_F =$ 5 ps, $4\Delta\xi = 0.2$ radians, and pseudo-temperature 10,000 K.

Although the above protocol was sufficient for RNA, the DNA duplexes proved to be much more flexible and the A-A mismatches readily became entangled with nucleotide backbone or formed short-lived stacking structures. To avoid these conformations, all the heavy atoms in the DNA duplex, except for the A-A mismatch and neighboring Watson-Crick pairs, were restrained using a very small harmonic constraint of 0.1 kcal/mol for the initial equilibration. For DNA, this constraint was large enough as to preserve the general shape of the structure, although readily allowing for transitions within the A-A mismatch. This small constraint was eliminated for the production runs. We also added a small plane-plane distance restraint between the A-A mismatch and the neighboring CG pairs, which prevents A-A stacking.

RESULTS

The sequences we investigated are shown in Fig. 1. We begin our discussion with a consideration of the single-mismatch sequence CAG and GAC. For each model CAG/GAC and RNA/DNA, we computed two free energy landscapes: one asymmetric map using the variables Ω and χ_{14} for A14 (although A5 stays in the anti conformation; see Materials and Methods), and one symmetric map using the variables χ_5 for A5 and χ_{14} for A14. Positive values of Ω represent well-stacked bases inside the helix core whereas negative values of Ω represent bases that had flipped out of the helix core. Values of χ between 90° and 270° (or, equivalently, between 90° and 180° and between −180° and −90°) are considered anti conformations; the other half ranges –90° –90° (or, equivalently 270°–360° and 0°–90°), which corresponds to syn conformations. These free energy landscapes display several stable minima. We have set the deepest minimum in each free energy map as the zero level of the free energy. On the diagrams, we have labeled the more prominent minima with letters that correspond to conformations shown in Fig. 2. For these structures, we have marked the most important hydrogen bonds. These are in good agreement with those obtained from quantum chemistry calculations of unsolvated DNA bases (73,74). The location and values of these minima are given in Table 1.

Fig. 3, *a* and *b*, shows the (Ω, χ_{14}) free energy maps for RNA-CAG and RNA-GAC. These figures share some general features: 1) the deeper minima with $\Omega > 0$ correspond to well-stacked bases inside the helix core; 2) the shallower minima with $\Omega < 0$ correspond to bases that have approximately flipped out of the helix core; and 3) the deepest minimum A1 corresponds to anti-anti conformation (because A5 is in anti conformation). The differences in free energies between the absolute minimum in A1 (stacked bases, anti-anti) and the next minimum in B1 or B2 (stacked bases, anti-syn) is ~1.2 kcal/mol for RNA-CAG, but it is 5.6 kcal/mol in RNA-GAC. We have computed least free energy paths on the (Ω, χ_{14}) free energy landscapes in Fig. 3, and examined the corresponding profiles. Sample free energies along these paths are presented in Fig. S3. The paths are relatively similar for the different maps with barriers in the 5–11 kcal/mol range for the B → A transition. The lowest values correspond to RNA-GAC, which therefore exhibits a larger transition rate for the B → A reaction.

Fig. 3, *c* and *d*, shows the (χ_5, χ_{14}) free energy maps for RNA-CAG and RNA-GAC. Because the two A-bases of the mismatch are completely equivalent, one can expect the free energy maps to show mirror symmetry across the diagonal, a feature that can generally be observed in these phase diagrams. The deepest minimum A1 is at (−168, −168) in RNA-CAG and ~(−163, −163) in RNA-GAC, corresponding in both cases to anti-anti conformations. In these maps, primed letters indicate minima related by mirror symmetry (e.g., B indicating anti-syn and B′ indicating syn-anti). These minima are degenerate with respect to the base-stacking parameter Ω. In RNA-CAG, the three anti-anti minima A1, A2, and A3 are degenerate in the phase diagram, all ending in the same position (the same happens with the

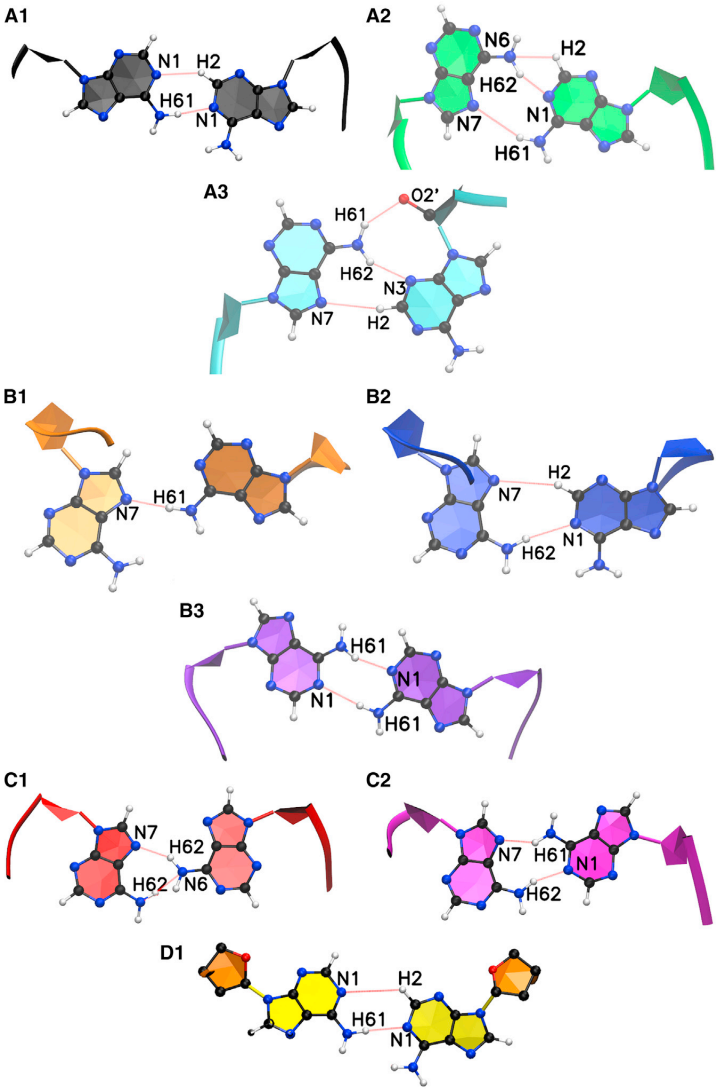


FIGURE 2 Given here are the A-A mismatch conformations for the main minima associated with the free energy landscapes. The letters denote different conformations: A, anti-anti; B, syn-anti; and C, syn-syn. Here D1 is a special case, because the χ-angle corresponds to syn-syn, but the base conformation looks like anti-anti due to the twisting of the sugar rings that become parallel to the bases. Note also that the hydrogen bonds associated with each of the conformations are also marked. To see this figure in color, go online.

B and B′ minima, and C minima in RNA-CAG and RNA-GAC; see Fig. 1).

Fig. 4 shows the free energy maps for DNA-CAG, and DNA-GAC. Although there are clear similarities between these free energy maps and their RNA counterparts, differences arise because of the greater flexibility of the DNA sugar ring, which slows down the convergence of the DNA free energy maps. For instance, consider the conformation in Fig. 2 D1. Here, both sugar rings are twisted to lie in the same plane as the A-A mismatch, leading to a (−syn)-(−syn) combination that shows a marked similarity with that in Fig. 2 A1). The (Ω, χ_{14}) free energies are the landscapes most affected by this convergence issue. On these maps, the anti-anti and the anti-syn DNA minima are 0.5 kcal/mol apart, which is within the error of the calculation. Thus the (Ω, χ_{14}) maps cannot truly distinguish the free energy difference between these two minima. This issue, however, is resolved by the (χ_5, χ_{14}) free energy maps, which clearly identify the anti-anti conformation as the global minimum structure. Because the (χ_5, χ_{14}) maps are degenerate with respect to the stacking variable Ω, we inspected all the conformations corresponding to this minimum and found that in all cases the bases are stacked inside the helical core.

TABLE 1 Main Minima for All the Free Energy Maps

Local Minimum		A1	A2	A3	B1	B2	B3	C1	C2	D1
A-A Form		Anti-Anti			Anti-Syn			Syn-Syn		
Main H-Bond		N1-H2, H61-N1	N7-H61, H62-N1	N3-H62	N7-H61	N7-H2, H62-N1	H61-N1, N1-H61	N7-H62, N6-H61	N7-H61, H62-N1	N1-H2, H61-N1
RNA-CAG	approximate location (Ω, χ_{14})	(75,195)	(−35, 190)	—	(50,45)	—	(−35, 60)	—	—	—
	relative free energy (kcal/mol)	0	3.5 ± 0.1	—	1.2 ± 0.2	—	5.4 ± 0.4	—	—	—
	approximate location (χ_5, χ_{14})		(−168, −168)		(−165, 55) for B, (58, −165) for B′			(55,55)		—
	relative free energy (kcal/mol)		0		2.7 ± 0.2			10.3 ± 0.3		—
DNA-CAG	approximate location (Ω, χ_{14})	(40, 250)	—	—	(61,40)	(23,43)	—	—	—	—
	relative free energy (kcal/mol)	0	—	—	0.2 ± 0.3	2.9 ± 0.3	—	—	—	—
	approximate location (χ_5, χ_{14})	(−110, −110)	—	—	(−100, 40) for B, (45,−100) for B′	—	—	(45,45)	—	(−50, −50)
	relative free energy (kcal/mol)	0	—	—	2.0 ± 1.0	—	—	2.2 ± 0.5	—	5.4 ± 0.6
RNA-GAC	approximate location (Ω, χ_{14})	(70, 195)	(−35, 190)	—	(78,50)	(50,40)	(−3, 52)	—	—	—
	relative free energy (kcal/mol)	0	5.9 ± 0.1	—	5.6 ± 0.1	5.5 ± 0.1	5.9 ± 0.1	—	—	—
	approximate location (χ_5, χ_{14})	(−163, −163)	—	—	(−165, 46) for B, (44, −162) for B′			(44,78)		—
	relative free energy (kcal/mol)	0	—	—	4.7 ± 0.2			9.8 ± 0.1		—
DNA-GAC	approximate location (Ω, χ_{14})	(40, 255)	—	—	(78,43)	(61,52)	—	—	—	—
	relative free energy (kcal/mol)	0	—	—	0.6 ± 0.4	0.2 ± 0.2	—	—	—	—
	approximate location (χ_5, χ_{14})	(−101, −101)	—	—	(−140,50) for B, (50, −140) for B′		—	(50,50)	—	(−46, −46)
	relative free energy (kcal/mol)	0	—	—	0.9 ± 0.5		—	0.9 ± 0.2	—	2.4 ± 0.1

The mirror images of B(B1,B2,B3), B′(B1′,B2′,B3′), are not shown. The free energy values in the B columns are the average of B and B′. All the values and errors are calculated based on the last 20 ns of the ABMD simulations.

Finally we note that the values of χ for RNA in anti conformation correspond to ∼180–200°, which is properly anti, whereas the equivalent values for DNA correspond to ∼230–260°, which corresponds to high anti. This difference can be explained by the presence of the hydroxyl group at the 2′ position in the sugar ring of RNA, as shown in Fig. 5. This hydroxyl interacts with the RNA backbone, especially the phosphate oxygens (or the other bases) pulling the sugar ring at one end and causing a twist at the other end, which results in an overall decrease of the χ-angle.

To gain further insight into the single A-A mismatches, we have followed up these calculations with regular, 1-μs MD simulations. Initial conformations for these single mismatch runs were chosen to be anti-anti, anti-syn, and syn-syn, respectively. The RMSD of the A-A mismatches with respect to the initial A-A conformations is shown in Fig. S2. A summary of these results is as follows. RNA-CAG was found to be stable in the initial anti-anti conformation (global minimum), making occasional excursions from anti-anti A1 to anti-anti A2/A3 (Fig. 2). However, when started in the anti-syn conformation B1, it did not find the global minimum in the 1-μs simulation. On the other hand, when RNA-CAG was started in the initial syn-syn conformation, it quickly transitioned to the anti-syn B1 conformation using the mechanism depicted in Fig. 9 *a*. RNA-GAC, which starts its trajectory in either anti-anti or anti-syn conformations, transitioned readily to its global minimum conformation A1. However, when it was started in the syn-syn conformation, it did not find its way back to the global minimum in the 1-μs timescale. With respect to DNA there is no major change in the symmetry of the χ-angle for either CAG and GAC sequences and all the runs explored only neighboring minima (e.g., runs that start in the anti-anti A1 conformation transitioned to the A2/A3

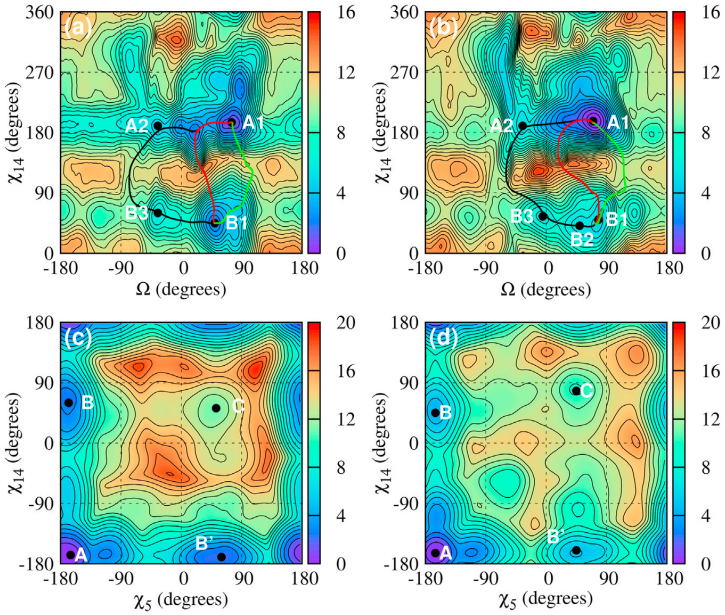


FIGURE 3 Free energy maps for single mismatches in RNA-CAG (r(5′-CCG-CAG-CGG)$_2$) and RNA-GAC (r(5′-GGC-GAC-GCC)$_2$). (*a*) (Ω, χ_{14}) map for RNA-CAG; (*b*) (Ω, χ_{14}) map for RNA-GAC; (*c*) (χ_5, χ_{14}) map for RNA-CAG; and (*d*) (χ_5, χ_{14}) map for RNA-GAC. The letters represent the local minima, with associated structures as shown in Fig. 2 and free energy values given in Table 1. The primed letters represent minima that are mirror images of minima labeled with the corresponding unprimed letters. The solid lines (*black, red and green*) describe three possible transition paths from B1 to A1. To see this figure in color, go online.

conformations, runs that start in the anti-syn B1 conformation transitioned to the B2 conformation, etc.)

We also ran 1-μs simulations of the TRs $(CAG)_4$ and $(GAC)_4$. Figs. 6 and 7 show the RMSD of the inner mismatches A_5–A_{20} and A_8–A_{17} as a function of time. RNA duplexes starting in the anti-anti A1 global minimum conformation (Fig. 6) were observed to occasionally transition to the anti-anti A2/A3 conformations in RNA-$(CAG)_4$, but not in RNA-$(GAC)_4$, where they remain locked in the global minimum position. RNA duplexes that started in the anti-syn B1

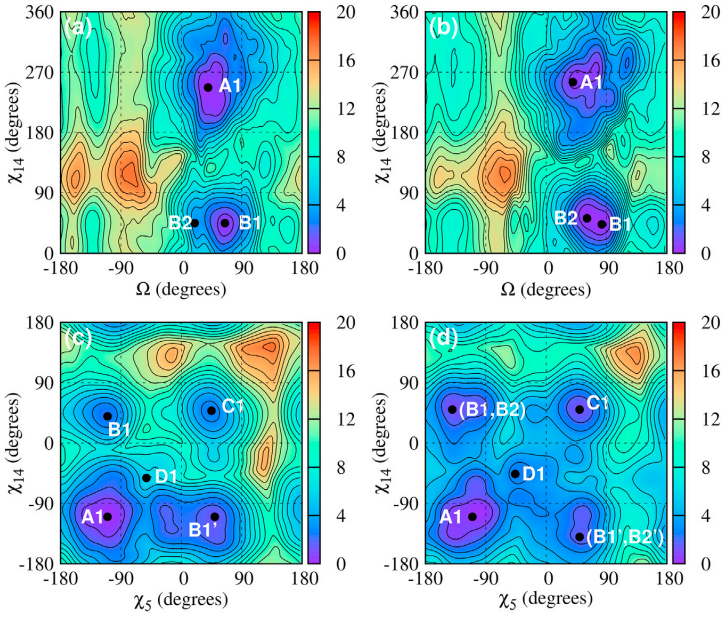


FIGURE 4 Given here are free energy maps for single mismatches in DNA-CAG (d(5′-CCG-CAG-CGG)$_2$) and DNA-GAC (d(5′-GGC-GAC-GCC)$_2$). (*a*) Shown here are: (Ω, χ_{14}) map for DNA-CAG; (*b*) (Ω, χ_{14}) map for DNA-GAC; (*c*) (χ_5, χ_{14}) map for DNA-CAG; and (*d*) (χ_5, χ_{14}) map for DNA-GAC. The letters represent the local minima, with associated structures as shown in Fig. 2 and free energy values given in Table 1. The primed letters represent minima that are mirror images of minima labeled with the corresponding unprimed letters. To see this figure in color, go online.

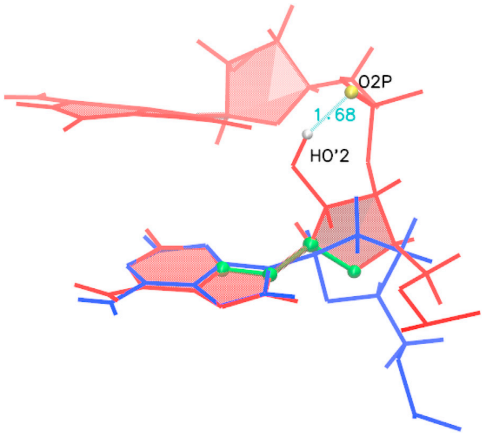


FIGURE 5 For the anti-anti conformations, the value of χ for DNA corresponds to high anti (230–260°) whereas for RNA, its value corresponds to just anti (180–200°). This difference is caused by the hydroxyl group at the 2′ position in the RNA sugar, which interacts with the backbone or other bases. (*Blue lines*) DNA; (*red lines*) RNA. The χ-torsion angle is indicated by green atoms. There is a strong direct interaction between the HO′2 atom and the O2P atom. To see this figure in color, go online.

conformation first sampled a few intermediate conformations like B2 and B3, before transitioning to the global minimum conformation A1 (with some A2/A3 in RNA-$(CAG)_4$ and almost none in RNA-$(GAC)_4$). Notice that the transition to the global minimum was not observed for the single-mismatch sequence RNA-(CAG). Presumably, this is because the adjacent Watson-Crick pairs in the single mismatch sequence constrain the motion of the mismatched bases. Correspondingly, the extra mismatches in RNA-$(CAG)_4$ seem to loosen the double helix, allowing for the rotations that lead to the minimum free energy. Also, our calculated free energy barrier in the B → A transition is smaller by 1.7 kcal/mol in RNA-GAC than in RNA-CAG, which helps account for the faster transition of the mismatch in RNA-$(GAC)_4$ over RNA-$(CAG)_4$. These simulations also allowed us to identify two different transition mechanisms (to be described below) from the major groove in RNA-$(CAG)_4$ (see Fig. 8), whereas only a single mechanism was observed from the major groove (see Fig. 8 *a*) for RNA-$(GAC)_4$. The corresponding increase in the entropy of the transition for RNA-$(CAG)_4$ may be the cause of the slightly higher free energy barrier in RNA-$(CAG)_4$. Finally, simulations in the initial syn-syn conformation were initially observed to oscillate between C1 and C2. In RNA-$(CAG)_4$ one of the internal mismatches managed to transition to anti-syn B1 using the mechanism depicted in Fig. 9 *b*, whereas the other remained trapped in the syn-syn conformation. In RNA-$(GAC)_4$, one of the mismatches transitions to the global minimum A1 (with some mixed A3), whereas the other transitions to the anti-syn conformations B1/B2, both according to the mechanism depicted in Fig. 9 *c* with a typical stacking conformation. DNA duplexes (Fig. 7), on the other hand, stayed in their initial geometry without major transitions over the 1 μs timescale. Thus, anti-anti conformations were observed to remain in A1, with a few transitions to A2/A3. Anti-syn conformations B1 stay anti-syn, with a few transitions to anti-syn B2, and syn-syn conformations stay syn-syn, with a few transitions from C1 to C2.

The slower transition rate of DNA with respect to RNA as exemplified by these MD simulations may be qualitatively understood as follows. A χ-rotation of an A-base can be clockwise or counterclockwise. Clockwise rotations (that would take the A-base along the path 50° → 0° → −110°) are hindered both for RNA and DNA due to steric clashes with neighboring bases. Fig. S4 shows how this clash would occur when the transition is attempted from the major groove in a clockwise rotation. Counterclockwise rotations (50° → 180° → −165°) are free from these

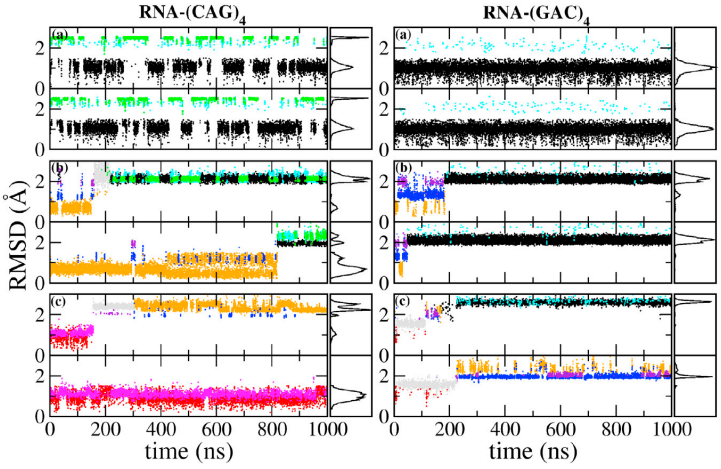


FIGURE 6 Given here is the RMSD for the internal mismatches in RNA-$(CAG)_4$ and RNA-$(GAC)_4$ as obtained from 1-μs MD simulations. In each panel, the upper row shows the RMSD for A_5–A_{20} and the lower row for A_8–A_{17}. Conformations are color-coded to agree with the mismatch conformations in Fig. 2. Initial conformations for each of the three panels in a column are as follows: (*top*) anti-anti (conformation *A1* in Fig. 2); (*middle*) anti-syn (B1); and (*bottom*) syn-syn (C1). (*Right panels*) Shown here is the distribution of the observed conformations. (*Gray*) Here we show irregular structures observed during the transition. To see this figure in color, go online.

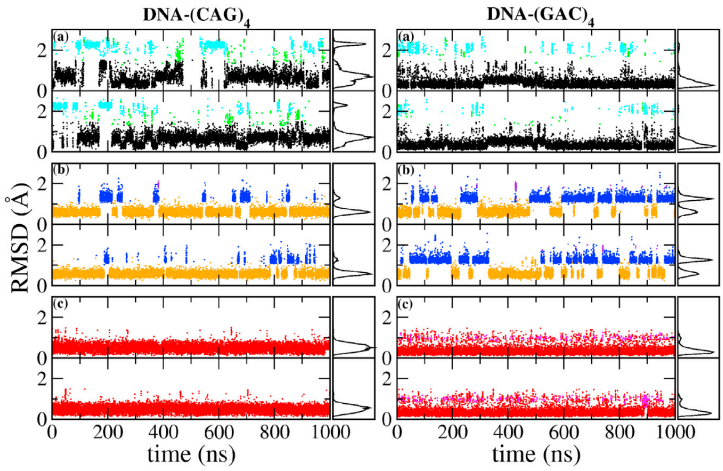


FIGURE 7 Given here is the RMSD for the internal mismatches in DNA-$(CAG)_4$ and DNA-$(GAC)_4$ as obtained from 1-μs MD simulations. In each panel, the upper row shows the RMSD for A_5–A_{20} and the lower row for A_8–A_{17}. Conformations are color-coded to agree with the mismatch conformations in Fig. 2. Initial conformations for each of the three panels in a column are as follows: (*top*) anti-anti (conformation A1 in Fig. 2); (*middle*) anti-syn (B1); and (*bottom*) syn-syn (C1). (*Right panels*) Shown here is the distribution of the observed conformations. To see this figure in color, go online.

clashes. For an RNA A-base, which is in an anti conformation, the counterclockwise rotation is also the shorter path to achieve a B → A (anti-syn to anti-anti) transition. However, for the DNA A-base that is in a high anti conformation, the shorter path is the clockwise rotation, which is strongly hindered. Thus the DNA-A mismatch is forced to rotate through a considerably longer path than the RNA mismatch, presumably resulting in a higher free energy barrier. In addition, both the bending and the hollow core that accompany A-RNA give more breathing space for the base to rotate as compared to B-DNA.

Now we consider the atomic mechanisms involved in various transitions. Fig. 8 shows two different mechanisms involved in an anti-syn to anti-anti conformational transition. In the top row the transition occurs through syn-to-anti base flipping in the major groove. The initial mismatch is in the B1 form. Then one of the A-bases rotates toward the major groove, breaking the N7-H61 hydrogen bond in the process. Its glycosyl angle χ twists and eventually rotates to anti, bringing the conformation to the A3 form, which can easily transition to A1. This corresponds to the green path in Fig. 3. The bottom row illustrates another mechanism where the transition occurs through the minor groove. In this case, one of the mismatched bases in syn conformation rotates toward the minor groove, allowing a transition from B1 to B2 and then B3. At this point the base may rotate

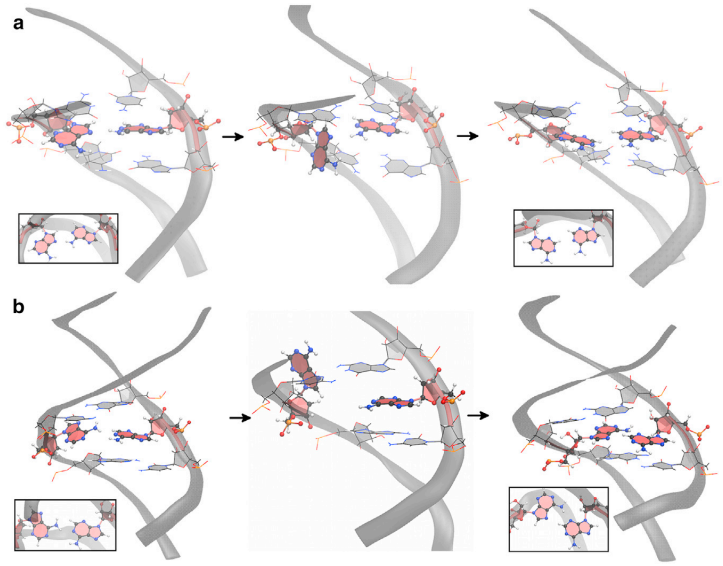


FIGURE 8 Two mechanisms associated with the transition from the anti-syn conformation to the anti-anti conformation. (*a*) The transition occurs through base flipping in the major groove. The structure goes from B1 to A3 to A1. (*b*) The transition occurs through base flipping in the minor groove. The insets show the A-A conformation in the vertical direction. See detailed descriptions in the text. To see this figure in color, go online.

back to the B1 form or interact with the backbone or neighboring bases, forming an irregular structure. This is shown as gray shading in Fig. 6 for A5–A10 in RNA-$(CAG)_4$. The irregular structure may last ~10 ns. In all cases, the transition is completed when the rotated base flips to the anti form, going to conformation A3 and then A1. This mechanism corresponds to the black path in Fig. 3.

Fig. 9 shows three mechanisms involved in syn-syn to anti-syn conformation. Fig. 9 *a* shows the transition path through the minor groove. First one mismatched base rotates toward the minor groove side, leading to a C2 form. Then it flips out, generally interacting with the backbone, and then it quickly flips back and changes to the anti B1 form. Fig. 9 *b* shows the transition path through the major groove. One A-base starts the transition by rotating from C1 to C2. Then the two A-bases separate and one of the bases flips to anti. This results in an unstacked anti-syn mismatch (not often observed in the simulations), as shown in the second graph of Fig. 9 *b*. The presence of hydrogen bonds in this uncommon anti-syn form makes it relatively stable, and the conformation lasts for ~100 ns, after which it finally transitions to the B1 form where it remains. Finally Fig. 9 *c* shows a relatively rare mechanism where the two syn A-bases first become stacked and then one base changes to anti. After this, the bases become unstacked and transition to conformation B1 and then B2.

To elucidate to what extent the mismatches distort the initial A-RNA and B-DNA forms, we have carried out a principal component analysis (75) (PCA) on the backbone of the duplexes. Figs. 10 and 11 show the time evolution of the first and second eigenvalues as well as the distribution of conformations projected onto the first principal component for the backbone of $(CAG)_4$ and $(GAC)_4$ for RNA and DNA. Only considered for this analysis are the internal residues that encompass internal mismatches, mainly residues 4–9 on one strand and 16–21 on the other. For RNA the eigenvalues stay relatively constant, and the projection of the conformations onto the first principal component results in a stable Gaussian distribution. The only exception is the RNA-$(GAC)_4$ duplex that starts in a syn-syn conformation, where a transition in the backbone conformation takes place at ~200 ns, after which the backbone remains stable. On the other hand, the DNA duplexes are stable when they start in anti-syn and syn-syn conformations, but they undergo

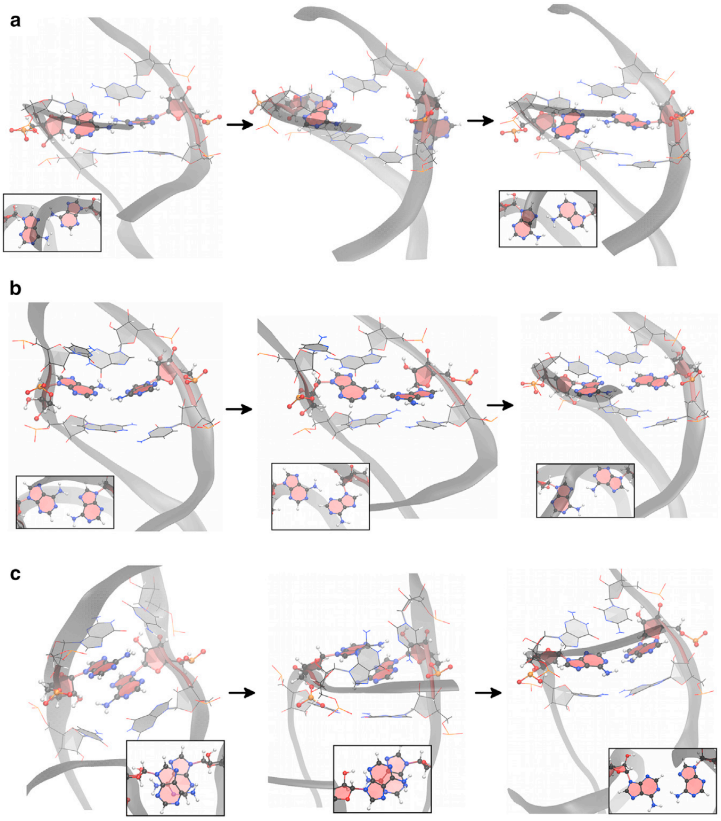


FIGURE 9 Three mechanisms associated with the transition from the syn-syn conformation to the anti-syn conformation. (*a*) The transition occurs through base flipping in the minor groove, following a path C1 → C2 → B1. (*b*) Shown here is the transition that occurs through base flipping in the major groove. (*c*) The two syn bases first stack on each other, one of them rotates while stacked, and then they become unstacked adopting anti-syn conformations. The insets show the A-A conformation in the vertical direction. See detailed descriptions in the text. To see this figure in color, go online.

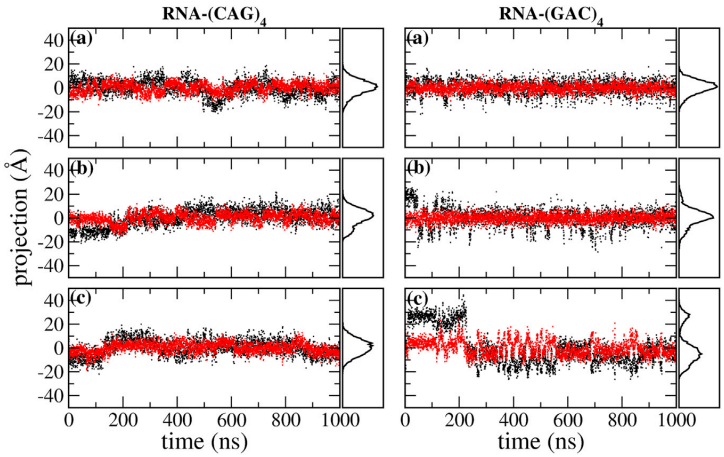


FIGURE 10 Given here are time plots of the PCA first- and second eigenvalues and distribution of the projections of the conformations onto the first principal component axis for the backbone corresponding to internal mismatches in RNA. Considered here are the residues 4–9 on one strand and the complementary residues 16–21 on the other. (*Left column*) DNA-$(CAG)_4$; (*right column*) DNA-$(GAC)_4$. (*Black*) First eigenvalue; (*red*) second eigenvalue. Initial conformations for the MD runs are (*a*) anti-anti, (*b*) anti-syn, and (*c*) syn-syn. To see this figure in color, go online.

considerable reaccommodation when they start in anti-anti conformations, both in the $(CAG)_4$ and $(GAC)_4$ forms. Conformational fluctuations along the direction of the first PCA eigenvector in the DNA-$(CAG)_4$, and DNA-$(GAC)_4$ duplexes with initial anti-anti conformations are shown in Fig. 12. This figure shows that the first eigenvector corresponds to the simultaneous coupling of unbending and unwinding modes.

To further quantify this unwinding, we show the simple twist based on C1′ atoms (see definition in the Supporting Material) in Fig. 13 for the middle eight steps for DNA and RNA with initial mismatches in anti-anti conformation. The green bars show the initial, constant twist corresponding to ideal B-DNA with a value of 36°, and ideal A-RNA with a value of 31.5°. Immediately after equilibration, the twist has already acquired sequence-dependent values (data not shown). The blue bars in the figure show the average value of twist for the last 200 ns of the 1 μs simulations. Notice that the final conformations display a mirror symmetry around the central step (step 6) that reflects the inversion symmetry of the sequences. Both DNA and RNA experience some degree of unwinding, but this is considerably more marked for DNA. Although both CAG and GAC sequences show a general decrease of twist, they do not share the same pattern of twist decrease. We take the general definition of Watson-Crick steps as GpC = GC/GC and CpG = CG/CG. In addition, we define steps containing mismatches as m_1 = AG/CA = CA/AG and m_2 = AC/GA = GA/AC. Thus, the pattern of steps for $(CAG)_4$ is m_1m_1-GpC-m_1m_1-GpC-m_1m_1-GpC-m_1m_1, and for $(GAC)_4$ it is m_2m_2-CpG-m_2m_2-CpG-m_2m_2-CpG-m_2m_2. Fig. 13 shows that DNA-$(CAG)_4$ experiences most unwinding in the m_1 steps surrounding the central GpC step (with a considerable decrease of twist) whereas DNA-$(GAC)_4$

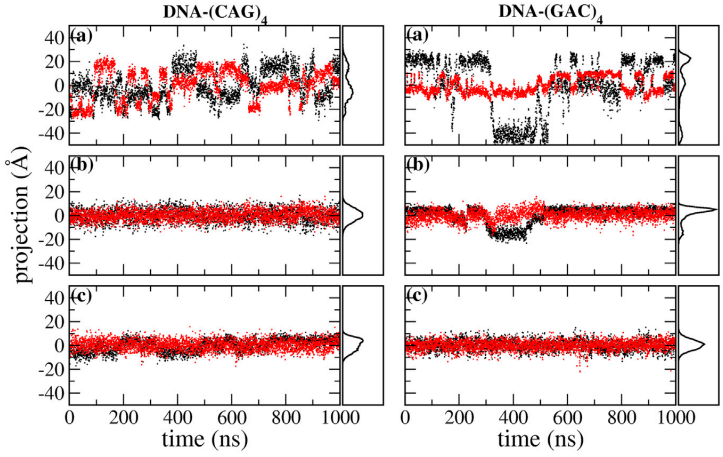


FIGURE 11 Given here are time plots of the PCA first- and second eigenvalues and distribution of the projections of the conformations onto the first principal component axis for the backbone corresponding to internal mismatches in DNA. Considered here are the residues 4–9 on one strand and the complementary residues 16–21 on the other. (*Left column*) DNA-$(CAG)_4$; (*right column*) DNA-$(GAC)_4$. (*Black*) First eigenvalue; (*red*) second eigenvalue. Initial conformations for the MD runs are as follows: (*a*) anti-anti, (*b*) anti-syn, and (*c*) syn-syn. To see this figure in color, go online.

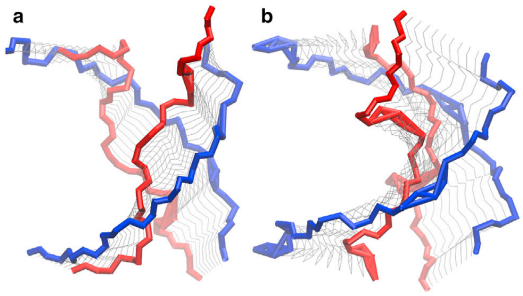


FIGURE 12 Shown here are fluctuations of duplex conformations around the first eigenvector direction, based on the PCA analysis of the backbone. (*a*) DNA-$(CAG)_4$; (*b*) DNA-$(GAC)_4$. Both duplexes have an initial anti-anti conformation. (*Blue line*) Most bending conformation; (*red line*) most unwinding conformation. To see this figure in color, go online.

experiences most unwinding at the CpG steps. The twist of RNA-$(CAG)_4$ barely decreases and it is not much affected by the sequence, the structure staying quite close to ideal A-RNA. The twist of RNA-$(GAC)_4$ decays at the mismatches, and stays almost the same or even increases at the CpG steps. We also considered the twist for the duplexes starting in initial mismatch conformations other than anti-anti. The RNA duplexes are evolving toward the global minimum with transitions taking place at different times (Fig. 6), and therefore it is enough to consider the anti-anti mismatch conformations, as done in Fig. 13. The DNA duplexes, on the other hand, get stuck in their initial mismatch conformations (Fig. 7). While in these nonequilibrium conformations, DNA does not experience unwinding (see Fig. S5). The unwinding of the anti-anti DNA duplexes can also be illustrated using the concept of handedness, defined in the Supporting Material and shown in Figs. S6–S8. In these figures, positive values of handedness mean a right-handed helix, zero stands for a duplex with no helicity, and negative values of handedness represent a left-handed helix. On Figs. S6 and S8, we see a clear decay of the positive, right-handed values for DNA sequences starting in the anti-anti mismatch conformation. Temporarily, different local turns can show zero or negative handedness for both sequences. The total handedness for the middle basepairs exhibits an oscillatory nature in DNA-$(CAG)_4$ (with a cycle of ~200 ns), whereas in addition DNA-$(CAG)_4$ experiences sudden zero handedness (parallel strands) for ~200 ns, and then it also recovers suddenly, reinitiating the smoothly oscillatory behavior. Naturally, when the helix unwinds, its radius of gyration increases, as shown in Fig. S9. This is in agreement with the PCA analysis, where the first mode is seen as a coupling of (un)winding and (un)bending. As shown, on average the helix stays slightly unwound but still right-handed, even at the local level (Fig. S8). By contrast, the RNA handedness stays constant throughout the simulation (Fig. S7). This analysis indicates that the global minimum A1 (anti-anti) corresponds to

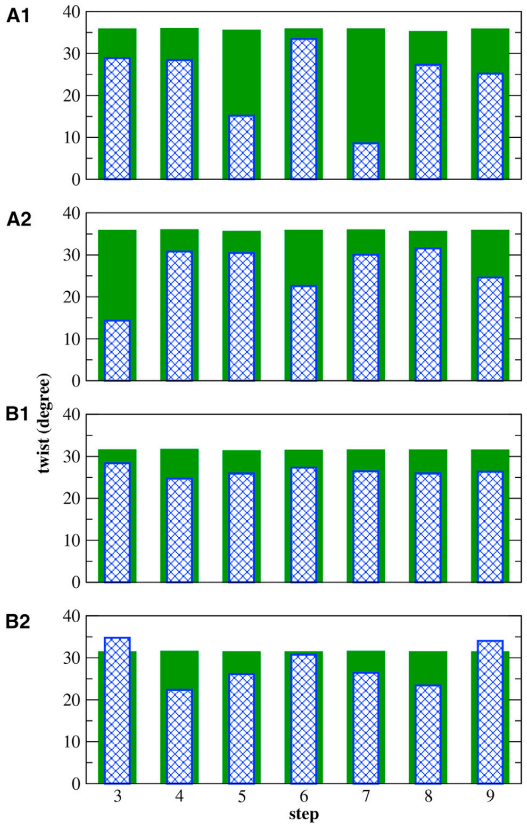


FIGURE 13 Shown here is a simple twist based on the C1′ atoms for the middle eight basepairs of the duplexes starting in anti-anti mismatch conformations. (A1) DNA-$(CAG)_4$; (A2) DNA-$(GAC)_4$; (B1) RNA-$(CAG)_4$; and (B2) RNA-$(GAC)_4$. Green bars show the initial value of ideal B-DNA (36°) and ideal A-RNA (31.5°). Blue bars show the final average values taken from the final 200 ns of the 1 μs simulations. To see this figure in color, go online.

a fairly stable helix in RNA (with relatively small fluctuations), and a very dynamical helix for DNA (with rather large fluctuations). For RNA, in contrast to DNA, it is therefore possible to select from the simulation data a helix that is close to the free energy minimum as being representative of these structures. The duplexes corresponding to RNA-$(CAG)_4$ and RNA-$(CAG)_4$ are shown in Fig. S10. The widths of the major and minor grooves, and the inclination angles, are given in Table S1 for the duplexes close to the global minimum A1 and for the duplexes close to the next minimum (anti-syn). Results for anti-anti RNA-$(CAG)_4$ are in general agreement with those observed previously (43): there is a wider major groove and a substantial decrease of the inclination angle with respect to the canonical A-RNA form. Notably, the next minimum (anti-syn) is quite close to the canonical A-RNA form, with a narrower

major groove and larger inclination angles compared to A1. Our results indicate that the RNA-$(GAC)_4$ structures follow similar trends.

Now we consider the distribution of the neutralizing Na^+ ions. Fig. S11 shows the distance between Na^+ ions to the center of mass of the A-A single mismatch in RNA and DNA. Different colors represent different ions to show the single-ion binding time for separate ions. Ions within a distance of 5 Å always have direct interactions with the bases in the mismatch. From the figures we see that the binding time for any single ion in RNA-(GAC) is very short. Both DNA duplexes have slightly longer (and comparable) binding times. RNA-(CAG), on the other hand, has the longest binding times, especially for initial anti-syn and syn-syn conformations. Fig. 14 shows the (any) ion occupancy of the A-A single mismatch in RNA and DNA. If the A-A mismatches stayed in the initial anti-anti or syn-syn conformations, ion distributions around A5 (*red*) should be the same as ion distribution around A14 (*blue*) due to the inversion symmetry of the single-mismatch duplexes (which is not present in the initial anti-syn conformations). Both DNA duplexes display this symmetry for initial anti-anti and syn-syn conformations. RNA-(CAG) with initial syn-syn conformation does not show this symmetry because it transitions to anti-syn. For the anti-syn conformations there is a large peak of ion occupation at atom N7 in base A5, which is in the anti conformation.

Some typical Na^+ ion binding conformations are shown in Fig. 15. In an anti-anti conformation, a typical binding site involves the A-N7 atoms (Fig. 15 *a*). In DNA the ion may also interact directly with the OP2 atom in the backbone, but not in RNA, where distances between bases and backbone are increased by the duplex bending (in this case, the ion interacts through intermediate waters with A-N6 or OP2). However, the ion occupancy of N7 is higher in RNA than in DNA. Fig. 15 *b* shows binding of Na^+ by A-N3 and A-O4′ in the minor groove. This binding site has only been observed in DNA-CAG. For anti-syn conformations, a strong ion bridge is observed where the Na^+ ion forms a bridge between the A-N7 (A in anti conformation) and the G-N7 and G-O6 atoms in the neighboring G-base in the major groove (Fig. 15 *c*). This ion bridge has highest occupancy and binding time for RNA-CAG (Fig. 14). A similar ion bridge was observed in Gacy et al. (39). For other structures, this bridge is also observed but not as strong as in RNA-CAG. In particular, in GAC sequences the A-mismatch in anti conformation has a weaker stacking with the neighboring G-base, increasing the distances of the atoms that would contribute to trapping the Na^+ ion. Ion binding in the minor groove can also connect the A-N1 atom (A in syn conformation) and atoms in a C-G Watson-Crick basepair: C-N4 and G-O6 in CAG sequences (Fig. 15 *d*), and C-N4 and G-N7 in GAC sequences

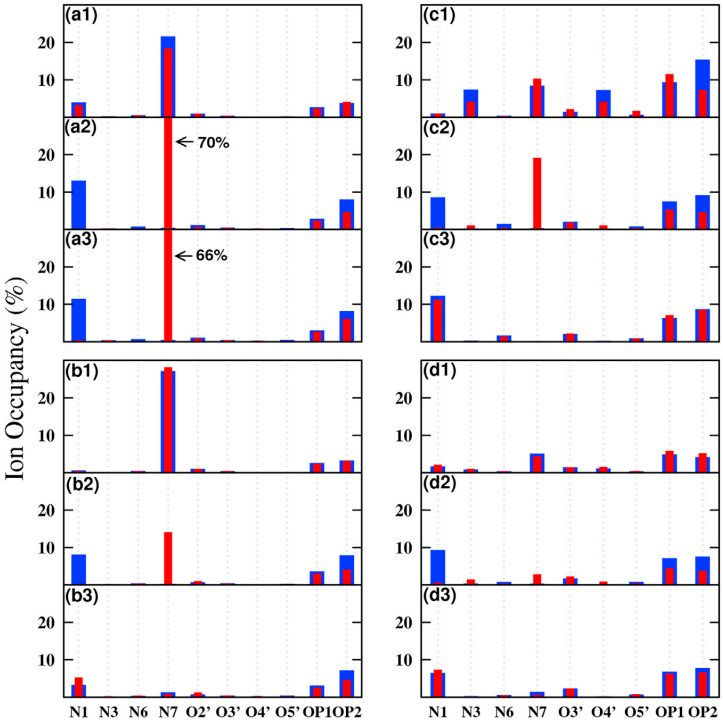


FIGURE 14 Given here is the ion occupancy around a single A-A mismatch in RNA and DNA. (*Red*) Base A5; (*blue*) base A14. RNA-CAG: (a1) anti-anti; (a2) anti-syn; and (a3) syn-syn. RNA-GAC: (b1) anti-anti; (b2) anti-syn; and (b3) syn-syn. DNA-CAG: (c1) anti-anti; (c2) anti-syn; and (c3) syn-syn. DNA-GAC: (d1) anti-anti; (d2) anti-syn; and (d3) syn-syn. To see this figure in color, go online.

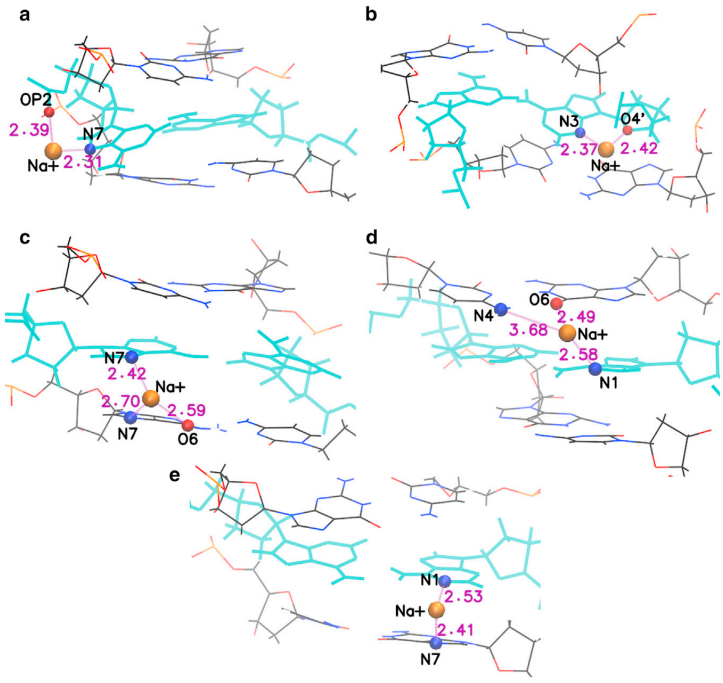


FIGURE 15 Given here are some typical Na^+ ion binding sites. A-A mismatches are highlighted in cyan and Na^+ ions are represented by orange spheres. (*a*) In an anti-anti conformation, a typical binding site involves the A-N7 atoms. In DNA, the ion may also interact with the OP2 atom in the backbone, but not in RNA. (*b*) Shown here is the ion binding by A-N3 and A-O4′ in the minor groove. This binding site has only been observed in DNA-CAG. (*c*) For anti-syn conformations, a strong ion bridge is observed where the Na^+ ion forms a bridge between the A-N7 (*A* in anti conformation) and the G-N7 and G-O6 atoms in the neighboring G-base in the major groove. (*d*) Shown here is the ion binding in the minor groove involving the A-N1 atom (A in syn conformation) and G-O6 and C-N4 in neighboring bases, as it occurs in CAG sequences. (*e*) Shown here is the ion binding in the minor groove involving A-N1 (A in syn conformation) and C-N4 and G-N7 in neighboring bases, as it occurs in GAC sequences. To see this figure in color, go online.

(Fig. 15 *e*). Finally ion binding to syn-syn mismatches also exhibits a large peak in A-N7 in RNA-CAG, but not in the three other sequences. Other than this, the next important binding site in the bases is A-N1, which engages in binding similar to that described in Fig. 15, *d* and *e*.

DISCUSSION

Although the mechanisms underlying TREDs are believed to be extremely complex, an important breakthrough has been the recognition that stable atypical DNA secondary structure in the expanded repeats is "a common and causative factor for expansion in human disease" (37). In addition, mutant transcripts also contribute to the pathogenesis of TREDs through toxic RNA gain-of-function (6,16–21). Thus, a first step toward the understanding of these diseases involves the structural characterization of the atypical DNA and RNA structures.

As stated in the Introduction, experimental investigations with atomic resolution have only considered CAG repeats in RNA; there are no experimental studies with atomic resolution of GAC repeats—not for DNA or RNA; and, perhaps most importantly, there are no experimental atomic resolution experiments of CAG repeats in DNA. The x-ray RNA-CAG duplex crystal structures include the following sequences: the sequence r(5′-GG-$(CAG)_2$-$CC)_2$ (41), and the sequence r(5′-UUGGGC-$(CAG)_3$-$GUCC)_2$ (42,43), which was also analyzed via NMR (43). The first study found that the duplexes favor the A-RNA form and that the A-A noncanonical pairs are in the anti-anti conformation. In the second sequence, both anti-anti and syn-anti A-A conformations were observed: the A-A pairs in the internal CAG always displayed the anti-anti conformation, whereas one (43) or two (42) of the terminal A-A pairs displayed the anti-syn conformation. These results are in general agreement with the complementary MD simulations (42). In this work, the authors computed an (Ω, χ_{14}) phase diagram for r(5′-CCG-CAG-$CGG)_2$; i.e., for RNA-CAG (the existence of this phase diagram is why we chose the sequences for the single mismatches). Our results for the (Ω, χ_{14}) RNA-CAG phase diagram are in very good agreement with these previous results, identifying the same set of minima with very similar free energy differences (~1 kcal/mol between the second minimum anti-syn at B1 and the deepest minimum anti-anti at A1). This in turn gives us confidence about the results obtained for the other seven free energy maps not previously investigated.

The other study is for DNA (45) and uses a sequence with a single CAG mismatch inserted in the middle of an otherwise complementary CAG• CTG B-DNA, and a sequence that is more relevant to the expanded disease, mainly $d(CAG)_6$. According to the conclusions of this study, the A-A mismatch in DNA behaves in exactly the opposite way than its RNA counterpart: it disfavors the anti-anti and

the anti-(+syn) conformations and adopts the (−syn)-(−syn) conformations, resulting in a local Z-form around the mismatch (45). Our results, which are based on careful free energy calculations, contradict these conclusions. First, the (Ω, χ_{14}) RNA-CAG free energy map precludes the exploration of syn-syn conformations because one of the mismatched bases always remains in an anti conformation. However, the (χ_5, χ_{14}) landscapes do probe these conformations. Our results are unequivocal: the global minimum for all possible combinations of CAG/GAC RNA/DNA is always anti-anti, followed by anti-syn. We speculate that the observed difference in the previous study may well be due to convergence issues (the lack of convergence of the simulations can be observed, for instance, in the figures that show twist; see Figs. S8 and S15 in the Supporting Information of (40)), which certainly do not reflect the inversion symmetry of the sequences) because the study only reports on 300-ns DNA simulations (45). Having said this, we also notice that difference between "high anti" in our case (250−255°) and the "−syn" reported previously in Khan et al. (45) (270−300°) may not be so large.

In this work, we have carried out free energy calculations and MD studies to determine the preferred conformations of the A-A mismatches in $(CAG)_n$ and $(GAC)_n$ trinucleotide repeats ($n = 1$ or 4) and the way in which these mismatch conformations affect the overall structure of RNA and DNA duplexes. Our main findings are the following.

1) The global minimum (A1) of the various free energy maps corresponds to A-A mismatches stacked inside the core of the helix with anti-anti conformations in the RNA sequences and (high-anti)-(high-anti) conformations in the DNA sequences. In terms of the free energy, the next higher minimum corresponds to anti-syn conformations, whereas syn-syn conformations are even higher.
2) DNA helices near the global minimum are very dynamic, exhibiting large fluctuations. RNA helices still fluctuate, but with considerably lesser amplitude than DNA. Fluctuations of the DNA helix around the first eigenvector direction in the PCA of the backbone shows a coupling of bending and unwinding modes. On the other hand, the anti-anti RNA helices close to the global free energy minimum are very stable. They exhibit a wider major groove and a substantial decrease of the inclination angle with respect to the canonical A-RNA form. RNA helices close to the next anti-syn minimum, on the other hand, are quite close to the canonical A-RNA form.
3) Free energy barriers between minima corresponding to different states of the glycosyl torsion angle χ are rather high, which results in low transition rates during regular MD. The systems can readily transition between conformations within the same χ-range (say, A1, A2, and A3 for anti-anti; B1, B2, and B3 for anti-syn; and C1 and C2 for syn-syn (see Fig. 2)) because these categories represent minima that are quite close in phase space. However, transitions between different χ categories are much slower.
4) Rates of MD transitions of the A-mismatches between different χ-categories are higher for RNA than DNA. The i2′ hydroxyl group in the sugar ring of RNA interacts with the backbone, keeping the corresponding value at a lower χ-value (just anti) with respect to DNA, whose sugar ring adopts a high anti conformation. This results in a shorter path for the RNA ring to rotate from syn to anti, as compared to that for DNA. This, in addition to the hollow core and bending of the A form in RNA, results in a higher transition rate for the syn → anti χ-rotation in RNA. In the 1-μs RNA-$(CAG)_4$ and RNA-$(GAC)_4$ simulations, initial conformations starting in anti-anti and in anti-syn all end up in the global minimum, whereas all mismatches except one in RNA-$(CAG)_4$ manage to transition from syn-syn to anti-syn. Instead, the DNA sequences remain in the initial conformations during the 1-μs simulations (except for transitions to neighboring local minima).
5) Several mechanisms for the transitions anti-syn → anti-anti and syn-syn → anti-syn have been identified both through the major and minor grooves. These are identified in Figs. 8 and 9, and all involve intermediate conformations (a transition from syn-syn → anti-anti is achieved through intermediate transitions steps: syn-syn → anti-syn and anti-syn → anti-anti). These transitions involve local distortions of the helical duplexes in the regions surrounding the mismatch. We note that quantum chemistry calculations for the anti-syn → anti-anti transition for isolated A•A mismatched bases (without the sugar ring) have recently been published (76).
6) DNA-$(CAG)_4$ and DNA-$(GAC)_4$ duplexes in anti-anti conformations experience some degree of unwinding. DNA-$(CAG)_4$ unwinds at the mismatches surrounding the GpC steps and DNA-$(GAC)_4$ unwinds at the CpG steps. Except for some instantaneous local oscillations, none of the sequences becomes left-handed, and there is no local Z-DNA structure (in addition, the mismatches remain in anti-anti conformations). We notice that the duplex structure seems to strongly depend on the pH of the solution and the ionic strength (36). In particular, CD and UV absorption spectroscopy experiments reveal the presence of GAC (but not CAG) Z-DNA under conditions of low alkaline pH, high NaCl salt, and various divalent ions.
7) Under conditions of neutral pH and only neutralizing ions, the main distinctions between CAG and GAC RNA sequences is given by the difference in free energy between the second minimum anti-syn B1 and the first minimum anti-anti A1. This difference is ∼1 kcal/mol for RNA-CAG but ∼5 kcal/mol for RNA-GAC in the (Ω, χ_{14}) map. In addition, the transition rates

B1 → A1 are higher for RNA-GAC than RNA-CAG. Taken together, this means that given the proper environment, the mismatches in CAG-rich RNA can easily adopt, and remain relatively stable in, the anti-syn conformation with long lifetimes, whereas those in GAC-rich RNA are less stable in the anti-syn conformation and would evolve more readily toward the global anti-anti minimum.

8) Under conditions of neutral pH and only neutralizing ions, the main distinctions between CAG and GAC DNA sequences are given by 1) the difference in the pattern of unwinding described in point 5 above; and 2) the presence of a unique minimum D1 in DNA-GAC that corresponds to a (−syn)-(−syn) conformation quite similar to the anti-anti A1 conformation due to twisting of the sugar rings. D1 is also present in CAG-rich DNA but at higher free energies, and is absent in RNA. On the (χ_5, χ_{14}) phase diagram the situation is inverted as compared to RNA: the differences between the second and first minima are 2.2 kcal/mol for DNA-CAG and 1.1 kcal/mol for DNA-GAC.
9) We have characterized the neutralizing Na^+ ion distribution around the A-A mismatches. The mismatches in RNA-CAG and RNA-GAC have the longest and shortest single-ion binding times, respectively. A-N7 represents a major binding site for all three RNA-CAG geometries, for RNA-GAC anti-anti, and to a lesser extent in RNA-GAC anti-syn and DNA-CAG anti-syn. The other important binding site is A-N1, which contributes to important ion bridges between the A-mismatches and adjacent G-C pairs.

We finish with two more comments. First, a comparison between the two homopurine mismatches A-A in trinucleotide repeats $(CAG)_n$ and $(GAC)_n$ and the G-G mismatches in trinucleotide repeats $(GGC)_n$ and hexanucleotide repeats $(GGGGCC)_n$ shows that they prefer different conformations: A-A favor anti-anti whereas G-G favors anti-syn (77–79). Second, as stated in the Introduction, CAG expansions cause late-onset, progressive neurodegenerative disorders after the expansions become greater than a given threshold (26). In diseases like Huntington's disease they can reach up to 250 repeats. GAC repeats, on the other hand, lead to rare skeletal dysplasias but do not expand by more than two repeats (from five normal repeats to a maximum of seven pathological repeats), therefore GAC diseases do not belong to the family of TREDs. Although the duplexes formed by GAC repeats seem to strongly depend on pH and ionic strength, it is interesting to check whether these results (under neutral pH and only neutralizing ions) reflect some differences between the two sequences. The main differences are summarized in points 6 and 7 above. We hope that our future studies under different pH and ionic conditions will help elucidate further differences between the CAG and GAC secondary structures.

SUPPORTING MATERIAL

Supporting Materials and Methods, eleven figures, and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30610-0.

AUTHOR CONTRIBUTIONS

Designed research: all authors. Performed research: F.P. and V.H.M. Analyzed data: F.P. and V.H.M. Wrote paper: all authors.

ACKNOWLEDGMENTS

We thank the NC State HPC Center for computational support.

The work was supported by the National Institutes of Health (NIH) (R01GM118508), the National Science Foundation (NSF) (SI2-SEE-1534941), and the Extreme Science and Engineering Discovery Environment (XSEDE) (TG-MCB160064).

SUPPORTING CITATIONS

References (80,81) appear in the Supporting Material.

REFERENCES

1. Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5:435–445.
2. Oberle, I., F. Rouseau, …, J. Mandel. 1991. Molecular-basis of the fragile-X syndrome and diagnostic applications. *Am. J. Hum. Genet.* 49:76.
3. Giunti, P., M. G. Sweeney, …, A. E. Harding. 1994. The trinucleotide repeat expansion on chromosome 6p (SCA1) in autosomal dominant cerebellar ataxias. *Brain.* 117:645–649.
4. Campuzano, V., L. Montermini, …, M. Pandolfo. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271:1423–1427.
5. Mirkin, S. M. 2006. DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol.* 16:351–358.
6. Mirkin, S. M. 2007. Expandable DNA repeats and human disease. *Nature* 447:932–940.
7. Pearson, C. E., K. Nichol Edamura, and J. D. Cleary. 2005. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6:729–742.
8. Wells, R. D., and S. Warren. 1998. Genetic Instabilities and Neurological Diseases, 2nd. Elsevier, San Diego, CA.
9. Orr, H. T., and H. Y. Zoghbi. 2007. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* 30:575–621.
10. Pearson, C., and R. Sinden. 1998. Slipped strand DNA (S-DNA and SI-DNA), trinucleotide repeat instability and mismatch repair: a short review. In Structure, Motion, Interaction and Expression of Biological Macromolecules, Vol 2. R. H. Sarma and M. H. Sarma, editors. 10th Conversation in Biomolecular Stereodynamics Conference, 191–207. US NIH, SUNY Albany, June 17–21, 1997.
11. Wells, R. D., R. Dere, …, L. S. Son. 2005. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res.* 33:3785–3798.
12. Kim, J. C., and S. M. Mirkin. 2013. The balancing act of DNA repeat expansions. *Curr. Opin. Genet. Dev.* 23:280–288.
13. Cleary, J. P., D. M. Walsh, …, K. H. Ashe. 2005. Natural oligomers of the amyloid-beta protein specifically disrupt cognitive function. *Nat. Neurosci.* 8:79–84.

14. Dion, V., and J. H. Wilson. 2009. Instability and chromatin structure of expanded trinucleotide repeats. *Trends Genet.* 25:288–297.
15. McMurray, C. T. 2008. Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair (Amst.).* 7:1121–1134.
16. Ranum, L. P. W., and T. A. Cooper. 2006. RNA-mediated neuromuscular disorders. *Annu. Rev. Neurosci.* 29:259–277.
17. Li, L.-B., and N. M. Bonini. 2010. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends Neurosci.* 33:292–298.
18. Jin, P., D. C. Zarnescu, …, S. T. Warren. 2003. RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in Drosophila. *Neuron* 39:739–747.
19. Jiang, H., A. Mankodi, …, C. A. Thornton. 2004. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum. Mol. Genet.* 13:3079–3088.
20. Daughters, R. S., D. L. Tuttle, …, L. P. Ranum. 2009. RNA gain-of-function in spinocerebellar ataxia type 8. *PLoS Genet.* 5:e1000600.
21. Krzyzosiak, W. J., K. Sobczak, …, P. Kozlowski. 2012. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.* 40:11–26.
22. Campuzano, V., L. Montermini, …, M. Koenig. 1997. Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum. Mol. Genet.* 6:1771–1780.
23. Kim, E., M. Napierala, and S. Y. Dent. 2011. Hyperexpansion of GAA repeats affects post-initiation steps of FXN transcription in Friedreich's ataxia. *Nucleic Acids Res.* 39:8366–8377.
24. Kumari, D., R. Biacsi, and K. Usdin. 2011. Repeat expansion in intron 1 of the Frataxin gene reduces transcription initiation in Friedreich ataxia. In Proceedings of the 2011 Experimental Biology Meeting. *FASEB J.* 25:895.
25. Punga, T., and M. Bühler. 2010. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *EMBO Mol. Med.* 2:120–129.
26. Zoghbi, H. Y., and H. T. Orr. 2000. Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.* 23:217–247.
27. Davies, S. W., M. Turmaine, …, G. P. Bates. 1997. Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell* 90:537–548.
28. Sikorski, P., and E. Atkins. 2005. New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils. *Biomacromolecules* 6:425–432.
29. Sharma, D., L. M. Shinchuk, …, D. A. Kirschner. 2005. Polyglutamine homopolymers having 8-45 residues form slablike β-crystallite assemblies. *Proteins* 61:398–411.
30. Schneider, R., M. C. Schumacher, …, M. Baldus. 2011. Structural characterization of polyglutamine fibrils by solid-state NMR spectroscopy. *J. Mol. Biol.* 412:121–136.
31. Buchanan, L. E., J. K. Carr, …, M. T. Zanni. 2014. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. *Proc. Natl. Acad. Sci. USA* 111:5796–5801.
32. Kar, K., C. L. Hoop, …, R. Wetzel. 2013. β-hairpin-mediated nucleation of polyglutamine amyloid formation. *J. Mol. Biol.* 425:1183–1197.
33. Man, V. H., C. Roland, and C. Sagui. 2015. Structural determinants of polyglutamine protofibrils and crystallites. *ACS Chem. Neurosci.* 6:632–645.
34. Zhang, Y., V. H. Man, …, C. Sagui. 2016. Amyloid properties of asparagine and glutamine in prion-like proteins. *ACS Chem. Neurosci.* 7:576–587.
35. Délot, E., L. M. King, …, D. H. Cohn. 1999. Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. *Hum. Mol. Genet.* 8:123–128.
36. Vorlícková, M., I. Kejnovská, …, J. Kypr. 2001. Conformational properties of DNA fragments containing GAC trinucleotide repeats associated with skeletal dysplasias. *Eur. Biophys. J.* 30:179–185.
37. McMurray, C. T. 1999. DNA secondary structure: a common and causative factor for expansion in human disease. *Proc. Natl. Acad. Sci. USA* 96:1823–1825.
38. Mitas, M., A. Yu, …, I. S. Haworth. 1995. The trinucleotide repeat sequence d(CGG)15 forms a heat-stable hairpin containing Gsyn•Ganti base pairs. *Biochemistry* 34:12803–12811.
39. Gacy, A. M., G. Goellner, …, C. T. McMurray. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* 81:533–540.
40. Petruska, J., N. Arnheim, and M. F. Goodman. 1996. Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Res.* 24:1992–1998.
41. Kiliszek, A., R. Kierzek, …, W. Rypniewski. 2010. Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res.* 38:8370–8376.
42. Yildirim, I., H. Park, …, G. C. Schatz. 2013. A dynamic structural model of expanded RNA CAG repeats: a refined x-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *J. Am. Chem. Soc.* 135:3528–3538.
43. Tawani, A., and A. Kumar. 2015. Structural insights reveal the dynamics of the repeating r(CAG) transcript found in Huntington's disease (HD) and spinocerebellar ataxias (SCAs). *PLoS One* 10:e0131788.
44. Svozil, D., P. Hobza, and J. Sponer. 2010. Comparison of intrinsic stacking energies of ten unique dinucleotide steps in A-RNA and B-DNA duplexes. Can we determine correct order of stability by quantum-chemical calculations? *J. Phys. Chem. B.* 114:1191–1203.
45. Khan, N., N. Kolimi, and T. Rathinavelan. 2015. Twisting right to left: A…A mismatch in a CAG trinucleotide repeat overexpansion provokes left-handed Z-DNA conformation. *PLOS Comput. Biol.* 11:e1004162.
46. Pérez, A., F. J. Luque, and M. Orozco. 2012. Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.* 45:196–205.
47. Cheatham, T. E., 3rd, and D. A. Case. 2013. Twenty-five years of nucleic acid simulations. *Biopolymers.* 99:969–977.
48. Sponer, J., M. Krepl, …, M. Otyepka. 2017. How to understand atomistic molecular dynamics simulations of RNA and protein-RNA complexes? *WIREs RNA* 8. http://dx.doi.org/10.1002/wrna.1405.
49. Case, D. A., T. A. Darden, …, P. A. Kollman. 2014. AMBER 14. University of California, San Francisco, San Francisco, CA.
50. Pérez, A., I. Marchán, …, M. Orozco. 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.* 92:3817–3829.
51. Yildirim, I., H. A. Stern, …, D. H. Turner. 2010. Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine. *J. Chem. Theory Comput.* 6:1520–1531.
52. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.
53. Joung, I. S., and T. E. Cheatham, 3rd. 2008. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B.* 112:9020–9041.
54. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.
55. Brovarets', O. O., R. O. Zhurakivsky, and D. M. Hovorun. 2014. Does the tautomeric status of the adenine bases change upon the dissociation of the $A^*{\cdot}A$(syn) Topal-Fresco DNA mismatch? A combined QM and QTAIM atomistic insight. *Phys. Chem. Chem. Phys.* 16:3715–3725.
56. Brovarets', O. O., and D. M. Hovorun. 2015. Wobble↔Watson-Crick tautomeric transitions in the homo-purine DNA mismatches: a key to

the intimate mechanisms of the spontaneous transversions. *J. Biomol. Struct. Dyn.* 33:2710–2715.

57. Babin, V., C. Roland, and C. Sagui. 2008. Adaptively biased molecular dynamics for free energy calculations. *J. Chem. Phys.* 128:134101.
58. Babin, V., V. Karpusenka, …, C. Sagui. 2009. Adaptively biased molecular dynamics: an umbrella sampling method with a time-dependent potential. *Int. J. Quantum Chem.* 109:3666–3678.
59. Case, D. A., R. Betz, …, P. Kollman. 2016. AMBER 16. University of California, San Francisco, San Francisco, CA.
60. Raiteri, P., A. Laio, …, M. Parrinello. 2006. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B.* 110:3533–3539.
61. Minoukadeh, K., Ch. Chipot, and T. Lelievre. 2010. Potential of mean force calculations: a multiple-walker adaptive biasing force technique. *J. Chem. Theory Comput.* 6:1008–1017.
62. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.
63. Barducci, A., G. Bussi, and M. Parrinello. 2008. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100:020603.
64. Babin, V., and C. Sagui. 2010. Conformational free energies of methyl-αL-iduronic and methyl-β-D-glucuronic acids in water. *J. Chem. Phys.* 132:104108.
65. Moradi, M., V. Babin, …, C. Sagui. 2009. Conformations and free energy landscapes of polyproline peptides. *Proc. Natl. Acad. Sci. USA.* 106:20746–20751.
66. Moradi, M., V. Babin, …, C. Sagui. 2010. A classical molecular dynamics investigation of the free energy and structure of short polyproline conformers. *J. Chem. Phys.* 133:125104.
67. Moradi, M., J.-G. Lee, …, C. Sagui. 2010. Free energy and structure of polyproline peptides: an ab initio and classical molecular dynamics investigation. *Int. J. Quantum Chem.* 110:2865–2879.
68. Moradi, M., V. Babin, …, C. Roland. 2011. A statistical analysis of the PPII propensity of amino acid guests in proline-rich peptides. *Biophys. J.* 100:1083–1093.
69. Moradi, M., V. Babin, …, C. Roland. 2011. PPII propensity of multiple-guest amino acids in a proline-rich environment. *J. Phys. Chem. B.* 115:8645–8656.
70. Moradi, M., V. Babin, …, C. Sagui. 2012. Are long-range structural correlations behind the aggregration phenomena of polyglutamine diseases? *PLOS Comput. Biol.* 8:e1002501.
71. Moradi, M., V. Babin, …, C. Sagui. 2013. Reaction path ensemble of the B-Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Res.* 41:33–43.
72. Pan, F., C. Roland, and C. Sagui. 2014. Ion distributions around left- and right-handed DNA and RNA duplexes: a comparative study. *Nucleic Acids Res.* 42:13981–13996.
73. Brovarets, O., Y. Yurenko, and D. Hovorun. 2014. The significant role of the intermolecular CH O/N hydrogen bonds in governing the biologically important pairs of the DNA and RNA modified bases: a comprehensive theoretical investigation. *J. Biomol. Struct. Dyn.* 33:1624–1652.
74. Brovarets, O., Y. Yurenko, and D. Hovorun. 2015. The significant role of the intermolecular CHMIDLINE HORIZONTAL ELLIPSISO/N hydrogen bonds in governing the biologically important pairs of the DNA and RNA modified bases: a comprehensive theoretical investigation. *J. Biomol. Struct. Dyn.* 33:1624.
75. Amadei, A., A. B. Linssen, and H. J. Berendsen. 1993. Essential dynamics of proteins. *Proteins* 17:412–425.
76. Brovarets, O., and D. Hovorun. 2015. How do long improper purine-purine pairs of DNA bases adapt the enzymatically competent conformation? Structural mechanisms and its quantum mechanical grounds. *Ukr. J. Phys.* 60:748–756.
77. Kiliszek, A., R. Kierzek, …, W. Rypniewski. 2011. Crystal structures of CGG RNA repeats with implications for fragile X-associated tremor ataxia syndrome. *Nucleic Acids Res.* 39:7308–7315.
78. Kumar, A., P. Fang, …, M. D. Disney. 2011. A crystal structure of a model of the repeating r(CGG) transcript found in fragile X syndrome. *Chembiochem.* 12:2140–2142.
79. Zhang, Y., C. Roland, and C. Sagui. 2017. Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases. *ACS Chem. Neurosci.* 8:578–591.
80. Lu, X. J., and W. K. Olson. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 31:5108–5121.
81. Lu, X. J., and W. K. Olson. 2008. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* 3:1213–1227.

Chapter 5

Structure and dynamics of DNA and RNA double helices obtained from the CCG and GGC trinucleotide repeats

Feng Pan, Viet Hoang Man, Christopher Roland, and Celeste Sagui. Structure and dynamics of DNA and RNA double helices obtained from the CCG and GGC trinucleotide repeats. Ready for submission to Biophysical Journal, 2017.

ABSTRACT

Expansions of both GGC and GCC sequences lead to a number of expandable, trinucleotide repeat (TR) neurodegenerative diseases. Understanding of these diseases involves, among other things, the structural characterization of the atypical DNA and RNA secondary structures. We have performed molecular dynamics simulations of $(GGC)_n$ and $(GCC)_n$ homoduplexes in order to characterize their conformations, stability and dynamics. Each TR has two reading frames, which results in eight non-equivalent RNA/DNA homoduplexes, characterized by CpG or GpC steps between the Watson-Crick basepairs. Free energy maps for the eight homoduplexes indicate the C-mismatches prefer anti-anti conformations while G-mismatches prefer anti-syn conformations. Comparison between three modifications for DNA of the AMBER force field shows good agreement for the mismatch free energy maps. The mismatches in DNA GCC (but not CCG) are extra-helical forming an extended e-motif. The mismatched duplexes exhibit characteristic sequence-dependent patterns where the disparities among the step twists are more pronounced in the G-rich sequences, which undergo some local unwinding. We have characterized the distribution of the neutralizing ions around the homoduplexes.

5.1 Introduction

Trinucleotide repeats (TRs) belong to the microsatellite family of simple sequence repeats (SSRs), that comprises all sequences with core motifs of 1 to 6 (and even 12) nucleotides that are repeated up to 30 times in the human genome, both in genetic and intergenic regions [1]. The length of these repeats varies greatly among people and the fact that they are over-represented in genes indicates that they may have played an important role in evolution and gene regulation [1]. SSRs exhibit "dynamic mutations" that do not follow Mendelian inheritance (which asserts that mutations in a single gene are stably transmitted between generations). Intergenerational expansion of SSRs is behind inherited neurological disorders known as "anticipation diseases", where the age of the onset of the disease decreases and its severity increases with each successive generation [2–5]. After a certain threshold in the length of the repeated sequence, the probability of further expansion and the severity of the disease increase with the length of the repeat. To date, approximately 30 DNA expandable SSR diseases have been identified and the

list is expected to grow [6, 7]. In particular, the dynamic mutations in human genes associated with TRs cause severe neurodegenerative and neuromuscular disorders, known as Trinucleotide (or Triplet) Repeat Expansion Diseases (TREDs) [3,8–10]. The expansion is believed to be primarily caused by some sort of slippage during DNA replication, repair, recombination or transcription [5–15]. Cell toxicity and death have been linked to the atypical conformation and functional changes of the RNA transcripts, of DNA itself [6, 16] and, when TRs are present in exons, of the translated proteins [6, 17–26]. The expanded RNA transcripts exhibit secondary structures that sequester regulatory proteins and cause abnormal nuclear foci [27–30]. Contributing to the complexity of the pathological mechanisms, there is also evidence that antisense transcripts of the expansion, i.e., expanded repeats resulting from the bidirectional transcription of the DNA TR expansions, can also form nuclear RNA foci that contribute to toxicity, and that both sense and antisense expansions can trigger protein translation in the absence of the start ATG codon, giving rise to the unconventional repeat-associated non-ATG (RAN) translation [31].

In this work we are interested in CGG and CCG TRs, which are overexpressed in the exons of the human genome: CGG TRs are found in the 5'-untranslated region (5'-UTR) of the fragile X mental retardation gene (FMR1) [32], while CCGs are found both in the 5'-UTR and translated regions of more than one gene. The normal range of the CGG TRs in the population is 5-54, with the last ten repeats increasing the probability of disease in descendants [33, 34]. TRs of 55-200 CGGs constitute premutations associated with fragile X-associated tremor ataxia syndrome (FXTAS) in males [35] and premature ovarian failure in females [36]. TRs longer than 200 CGG cause the inherited fragile X mental retardation syndrome [37]. CCG TRs are related to three TREDs: the longest expansion occurs in the FRM2 gene giving rise to chromosome X-linked mental retardation (FRAXE) [38], and they also seem to play a role in Huntington's disease [39], and myotonic dystrophy type 1 [40].

An important breakthrough in the understanding of TREDs has been the recognition that stable atypical DNA secondary structure in the expanded repeats is "a common and causative factor for expansion in human disease" [41]. This atypical secondary structure forms when the parental DNA strands are separated freeing single-stranded DNA, which can occur during the processes of replication, translation, recombination and repair. In addition, mutant transcripts also contribute to the pathogenesis of TREDs through toxic

RNA gain-of-function [6, 17–22]. Thus, a first step towards the understanding of these diseases involves the structural characterization of the atypical DNA and RNA structures. Various experimental methods *in vitro*, such as CD, UV absorbance, NMR, electrophoretic mobility assay, and chemical or enzymatic digestion [42], show a general trend to formation of duplexes and hairpins, depending on the sequence length and environment conditions. Among these secondary structures, those formed by CGG expansions seem to be the most stable.

Crystallographic studies for short RNA duplexes provide valuable atomic detail. For the CGG expansion, two crystallographic studies using unmodified sequences 5'-G-$(CGG)_2$-C-3' (PDB ID 3R1C, Ref. [43]) and 5'-UU-GGGC-$(CGG)_3$-GUCC-3' (PDB ID 3JS2, Ref. [44]) found that the RNA helices have the A-form, with some variations, with the G-G pairs in a typical anti-syn conformation, with two hydrogen bonds between the Watson-Crick edge of G_{anti} and the Hoogsteen edge of G_{syn}. For the CCG sequence, there is one crystallographic RNA duplex with an unmodified sequence 5'-G-$(CCG)_2$-C-3' (PDB ID 4E59, Ref. [45]), and one solution NMR DNA duplex 5'-$(CCG)_2$-3' (PDB ID 1NOQ, Ref. [46, 47]). The C-rich structures are less conclusive because they involve only two repeats, which results in the slipping of one strand with respect to the other. In the RNA crystal structure, this dislocation and the stacking of the oligomers along the c-axis in the crystal results in a single C·C pair effectively surrounded by four C-G Watson-Crick pairs (with two overhanging C's). Thus, it is not clear whether this structural environment for the single "mismatch" can reproduce the one that would occur in the cell for longer $(CCG)_n$ sequences, where each C·C pair may (or may not) be surrounded by only two Watson-Crick pairs. The C·C pair surrounded by four Watson-Crick C-G base pairs as shown in the RNA crystal might be overconstrained with respect to that in a real CCG expansion. In the DNA duplex, the slipping of the strands leaves the two 5'-C terminal unpaired, and a single central C·C mismatch surrounded by two Watson-Crick pairs. This gives rise to the "e-motif", where the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the $i-2$ residue (i.e., towards the 5' direction in each strand).

Another important issue when considering possible TR conformations is the nature of the Watson-Crick pairs that surround the mismatches [48, 49]: sequences of the form 5'-$(CGG)_n$-3' and 5'-$(CCG)_n$-3' (without slipping) exhibit GpC steps between the Watson-Crick base pairs, while sequences of the form 5'-$(GGC)_n$-3' and 5'-$(GCC)_n$-3' (without

slipping) exhibit CpG steps between the Watson-Crick base pairs. The two RNA G-rich crystal structures [43, 44] involve GpC steps; terminal mismatches in 5'-UU-GGGC-$(CGG)_3$-GUCC-3' in Ref. [44] are surrounded by CC/GG steps, not present in a $(CGG)_n$ expansion. Indeed, with the use of high level *ab initio* calculations, it has been shown that CC/GG steps are the least stable of the ten dinucleotide steps, with well-separated energies [50] from the other dinucleotide steps. The slipping of strands with respect to each other in the (CCG) sequences results in GpC steps for the RNA crystal [45] and in CpG steps for the DNA NMR structure [46] (as opposed to the GpC steps that would result if the DNA strands were paired at the ends).

The work presented here is part of our effort to achieve a unified and comparative description of the nucleic acid duplexes obtained from SSRs for both DNA and RNA, considering all the possible reading frames that result in CpG or GpC steps between the Watson-Crick base pairs. Our previous work includes a characterization of the four helical duplexes obtained from the CAG (GpC steps) and GAC (CpG steps) TRs for both RNA and DNA [51]; and of the twelve helical duplexes derived from the (GGGGCC) hexanucleotide repeat (HR) expansion in the C9ORF72 gene, and its associated antisense (GGCCCC) expansion [52]. CAG TRs are known to cause ten late-onset progressive neurodegenerative diseases, including spinocerebellar ataxia type 12 (SCA12), Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA) and several other spinocerebellar ataxia (SCA) diseases [53]. On the other hand, GAC repeats behave quite differently: expansion by one repeat in the human gene for cartilage oligomeric matrix protein, which exhibits a $(GAC)_5$ repeat, causes multiple epiphyseal dysplasia, while expansion by two repeats or, alternatively, deletion by one repeat causes pseudoachondroplasia [54]. A (GGGGCC) HR expansion in the first intron of the C9ORF72 gene has been shown to be the major cause behind frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) [55, 56]. While the unaffected population carries fewer than 20 repeats (generally no more than a couple), large expansions greater than 70 repeats and usually encompassing 250-1600 repeats have been found in C9FTD and ALS patients. The twelve duplexes that we studied result from the three different reading frames in sense and antisense HRs for both DNA and RNA. These duplexes display atypical structures relevant not only for a molecular level understanding of these diseases but also for enlarging the repertoire of nucleic-acid structural motifs.

In this work, we present results for molecular dynamics (MD) simulations and free energy calculations for both CCG and GGC trinucleotide repeats, with either CpG or GpC steps, for both RNA and DNA. This results in eight different non-equivalent helical duplexes. We compare results with the one case, G-rich RNA with GpC steps, which is well characterized experimentally. The good agreement with the experimental structures helps validate our results for the other seven cases. In addition to the free energy maps, we identify mechanisms of transition of the mismatches towards the global free energy minimum, and link these mechanisms to paths over the free energy maps. We complete the work with a characterization of the neutralizing Na^+ ion distributions around the mismatches. Strictly speaking, the non-canonical C·C and G·G pairs in RNA are not "mismatches" since RNA is not necessarily self-complementary. However, since we are considering both DNA and RNA in their duplex form, we will call these non-canonical base pairs mismatches for simplicity.

5.2 Materials and Methods

Molecular Dynamics. The sequences employed in this work are shown in Fig. 6.1. The simulations were carried out using the PMEMD module of the AMBER v.16 [57] software package with force fields ff99 BSC1 [58] for DNA and ff99 BSC0 [59]+χOL3 modification [60] for RNA. In addition we have used the BSC0 and OL15 [61] to compute and compare various free energy maps for single mismatch DNA, and have run regular MD for the DNA C-rich four-repeat sequences both with BSC0 and BSC1. The TIP3P model [62] was used for the water molecules, along with the standard parameters for ions in the AMBER force fields [63]. The long-range Coulomb interaction was evaluated by means of the Particle-Mesh Ewald (PME) method [64] with a 9 Å cutoff and an Ewald coefficient of 0.30768. Similarly, the van der Waals interaction were calculated by means of a 9 Å atom-based nonbonded list, with a continuous correction applied to the long-range part. The production runs for MD were generated using the leap-frog algorithm with a 2 fs timestep with Langevin dynamics with a collision frequency of 1 ps^{-1}. Conformations were saved every picosecond. The SHAKE algorithm was applied to all bonds involving hydrogen atoms. Regular MD was run for all sequences, starting with different conformations of the χ angle up to 1 μs.

Free energy maps. The sequences with a single mismatch, CCG, GCC, CGG and GGC,

were used to find the conformation of the mismatch that minimizes the free energy. To calculate the free energy maps, we made use of the Adaptively Biased Molecular Dynamics (ABMD) method [65, 66] which has been implemented for PMEMD in AMBER v.16 [57]. The free energy – or potential of mean force (PMF) – is calculated as a function of one or more collective variables, which must carefully be chosen as to reflect the underlying physics of the problem. ABMD has been implemented with multiple walkers (both noninteracting [67] and interacting walkers, with the latter interacting by means of selection algorithm [68]), Replica Exchange Molecular Dynamics (REMD) [69] and 'Well Tempered' (WT) extensions [70]. The free energy of these mismatches was calculated as function of two main collective variables. We define: (1) χ_5 as the glycosyl torsion angle χ of C5 or G5, namely the dihedral angle O4'-C1'-N1-C2 for C and O4'-C1'-N9-C4 for G; and (2) χ_{14} which represents the χ angle of C14 or G14. With these variables, we constructed the 2-dimensional phase diagrams, (χ_5,χ_{14}), which can explore all options of χ (anti-anti, anti-syn, syn-anti, syn-syn). A given free energy landscape is deemed to have converged when both the position and differences in the free energy values of the minima remain approximately constant as further ABMD cycles are performed. For both DNA and RNA, at least 270 ns simulations are performed for each of the (χ_5, χ_{14}) maps; for some sequences, runs needed to be extended up to 600 ns to reach a better convergence.

After the initial conformations were set up as explained below, multiple walker ABMD runs at constant volume and 300 K were carried out with 8 replicas. The first ABMD simulation was for 30.0 ns with parameters $\tau_F = 1$ ps and $4\Delta\xi = 0.5$ radians. This simulation provided for a rough estimate of the free energy landscape over the relevant parameter space. We then followed this up with a finer 120 ns WT-ABMD simulation (parameters $\tau_F = 1$ ps, $4\Delta\xi = 0.2$ radians, pseudo-temperature 10,000 K). For these runs, the total number of hydrogen bonds in neighboring CG Watson-Crick base pairs were slightly restrained to be six using a 1.0 kcal/mol harmonic constraint. This was used in order to avoid the large-scale twisting of the whole structure during the long simulations. This constraint, however, was chosen to be flexible enough so as to readily allow for the relevant anti-syn transitions. Finally, a slower and smoother flooding in order to refine the landscapes was carried out with parameters $\tau_F = 2$ ps, $4\Delta\xi = 0.2$ radians, and pseudo-temperature 10,000 K. The final biasing potential was processed by the *nfe-umbrella-slice* tool [57] to get the two-dimensional free energy.

Initial conformations. Initial conformations for one and three repeat sequences were

created as follows. First, we created the duplexes with the four possible combinations of χ angle for the C·C or G·G mismatches: anti-anti, anti-syn, syn-anti and syn-syn. These were then solvated in an octahedral box with neutralizing Na^+ ions as in previous work [71], with a distance of at least 10 Å between the duplexes and walls of the box. The box was then filled with a suitable number of waters. The system was then minimized: first keeping the nucleic acid and ions fixed; then, allowing them to move. Subsequently, the temperature was gradually raised using constant volume simulations from 0 to 300 K over 50 ps, followed by a further 50 ps run. Then a 100 ps run at constant volume was used to gradually reduce the restraining harmonic constants for nucleic acids and ions. This was followed by a 1.0 ns constant pressure run, with the χ angles of the mismatches slightly restrained so that these retain their initial anti- or syn- conformation. We took random conformations from the last 200 ps of these runs as the initial conformations for both the ABMD and MD runs. That means, we picked two structures from each of the four runs (anti-anti, anti-syn, syn-anti and syn-syn). For the four repeats, $(CCG)_4$, $(GCC)_4$, $CGG)_4$ and $(GGC)_4$, the initial mismatch conformation was chosen as the one that minimizes the free energy, and two 1 μs simulations were run at 310K: one starting from an ideal A form and one starting from an ideal B form.

5.3 Results

In this section we discuss our results. The sequences considered in this study are shown in Fig. 6.1. Unless otherwise stated, results for DNA are shown for the BSC1 modification of the force field.

A. Free energy maps. We begin our discussion with a consideration of the free energy maps for the single-mismatch sequences CCG1, GCC1, CGG1 and GGC1. As described in the Methods section, we use as collective variables the dihedral angles χ_5 for C5 (G5) and χ_{14} for C14 (G14). Values of χ between 90° and 270° (or, equivalently, between 90° and 180° and between -180° and -90°) are considered anti conformations; the other half range, -90° to 90° (or, equivalently 270° to 360° and 0° to 90°), corresponds to syn conformations. These free energy landscapes display several stable minima. We have set the deepest minimum in each free energy map as the zero level of the free energy. Because the two bases of the mismatch are completely equivalent, one can expect the free energy maps to show mirror symmetry across the diagonal once the maps have converged, a feature that

can generally be observed in these phase diagrams.

Table 5.1 gives the position of the principal minima in the phase diagram and their relative free energy value. Fig. 5.2 shows the (χ_5,χ_{14}) free energy maps for the C-rich duplexes. For RNA, the deepest minimum is located at $\chi = -163°$ for both mismatches and both sequences, and for DNA, the deepest minimum is located between $\chi = -122°$ and $\chi = -125°$ for both sequences. These χ values correspond to anti-anti conformations. For all duplexes, the next minima correspond to anti-syn conformations, while syn-syn conformations are considerably higher in energy. For RNA, the anti-syn minima are closer in value to the absolute anti-anti minimum than for DNA.

Fig. 5.3 shows the (χ_5,χ_{14}) free energy maps for the G-rich duplexes. For all duplexes but RNA-CGG1, the absolute minimum corresponds to anti-syn conformations. In RNA-CGG1, the anti-anti and anti-syn minima have the same value within the error of the calculation. We believe the inability of the free energy calculation to pin down the anti-syn conformation as absolute minimum is due to a strong triple G-base stacking not present in the GGC repeat. Fig. 5.9 shows this stacking. Notice also that the hydrogen bonds G14-N2 and G15-O6 also contribute to the stacking stability of RNA-CGG. For DNA, the anti-syn minima are located at (-96°,73°) and mirror image (73°,-96°) for CGG1; and (-113°,70°) and mirror image (70°,-113°) for GGC1. For RNA, the anti-syn minima are located at (-160°,40°) and mirror image (40°,-158°) for CGG1; and (-160°,40°) and mirror image (40°,-160°) for GGC1.

Recently, several improved force fields have been introduced for DNA. A comparison of free energy maps computed with different force fields BSC0, BSC1 and OL15 is shown in Fig. S0 for two of the DNA sequences. The free energy maps are relatively similar: All force fields predict the absolute minimum to be anti-anti for the C·C mismatches and anti-syn for the G·G mismatches. The positions of the minima are similar, especially for the G-rich duplexes. The main difference is that the minima are deeper in BSC1 and OL15 providing for a more rigid mismatched DNA duplex with respect to that in BSC0. In addition, in the C-rich duplexes, the anti-syn minima are closer in depth to the anti-anti absolute minimum in BSC0 than in the other two fields; OL15 seems to be an intermediate case in this respect.

B. MD simulations for one- and three-mismatch sequences. To gain further insight into the dynamics of the mismatches, we have followed these calculations up with regular, 1 μs MD simulations, both for the one- and three-mismatch sequences, starting

from the four possible combinations for the mismatches: anti-anti, anti-syn, syn-anti, and syn-syn. Figs. S1 to S16 show the χ_5 and χ_{14} torsion angles, the hydrogen bond number (hbond) between the mismatches, and the distance between the centers of mass of the bases in the mismatch, as a function of time. For the C-rich sequences, general observations are: (i) sequences starting in anti-anti conformations are stable; (ii) sequences starting in anti-syn quickly transition to anti-anti in DNA; (iii) sequences starting in anti-syn take several hundred nanoseconds (close to the 1μs time scale) to transition to anti-anti in RNA with one mismatch, but transition quickly in the looser three-mismatch sequence; (iv) sequences starting in syn-syn transition to either the absolute anti-anti minimum or the intermediate anti-syn minimum. For the G-rich sequences, general observations (ignoring the ambivalence in the anti/syn definition when $\chi = \pm 90°$) are: (i) sequences starting in anti-anti relative minimum remain in anti-anti in the 1μs time scale; (ii) sequences starting in the anti-syn absolute minimum remain in this minimum; (iii) RNA sequences starting in syn-syn quickly transition to anti-syn (one repeat) or stay in syn-syn in the 1μs time scale (three repeats); while DNA sequences remain around $\chi = +90°$. For sequences starting in the (rather artificial) syn-syn conformations, long-lived stacking among the bases can be observed in a few runs. Although RNA does not display an e-motif, extrusion of a C base is observed in RNA-GCC1 (anti-syn) and RNA-CCG3 (syn-syn), probably caused by the duplex seeking a transition path towards the anti-anti global minimum.

The hydrogen bond populations during the 1 μs regular MD for the single mismatch sequences are shown in Table 5.2. First, we consider the C·C mismatch conformations shown in Fig. 5.4(a)(b). For the DNA BSC1 force field correction, there is no e-motif formation in the 1μs time scale, thus the hydrogen bonds described here correspond to an intra-helical C·C mismatch. For the anti-anti conformations the main hydrogen bonds are N3-H41 and H41-N4 with an additional important contribution by H41-O3 in RNA-CCG. For the anti-syn conformations, the hydrogen bond H42(syn)-N3(anti) is present in all RNA and DNA duplexes; the hydrogen bond H41(anti)-N4(syn) is present in all duplexes but RNA-CCG; and the hydrogen bond H42(syn)-O2(anti) is present in DNA duplexes only. The presence of three, relatively stable hydrogen bonds in DNA results in shorter C·C mismatch distances. Next, we consider the G·G mismatch conformations shown in Fig. 5.4(c)(d). For the anti-anti conformations the main hydrogen bonds are N2H21/H22-O6 and N1H1-N2, while for the anti-syn conformations the main hydrogen bonds are N1H1-O6 and N2H21-N7 for both RNA and DNA; the RNA duplexes also have

an important contribution from N2H22-OP2; and there is a smaller contribution from N7-H1. Notice the very good agreement in the populations for the anti-syn and syn-anti conformations, which are expected to be equivalent due to the symmetry of the sequences. For either of the mismatch types, syn-syn conformations cannot form hydrogen bonds.

C. MD simulations for four-mismatch sequences. For the sequences with four TRs, we carry out MD simulations with initial mismatch conformations corresponding to the absolute minimum of the free energy, i.e., anti-anti for the C·C mismatches and anti-syn for the G·G mismatches. For each sequence, two runs were performed: one with the initial duplex in ideal A form and one with the initial duplex in ideal B form. The two simulations quickly converge. Convergence and stability of these runs is displayed in Figs. S17 to S20, that present results for the dihedral angles of the internal mismatches, the number of hydrogen bonds and the C1'–C1' distances. Structural features of the resulting duplexes are presented in Fig. 5.5 and Fig. 5.6. These figures show the distribution of the four TR sequences grouped by double helix handedness, C_1'–C_1' distance, and χ_6 and χ_{23} dihedral angles (see Fig. 1). Handedness (defined in the SI) has values of 5.1 and 6.1 for ideal A and B helices. First, we consider the C-rich sequences: (i) For RNA, there is almost no difference between the CCG4 and GCC4 sequences, with both sequences resulting in duplexes distributed around the ideal A-form, and dihedral angle χ distributions corresponding to anti -*ap* conformations; (ii) The DNA duplexes slightly unwind from the initial B-form, and end in forms intermediate between the A- and B-forms: GCC is more B- like, while CCG is more A- like; in other words, DNA duplexes with GpC steps unwind more. This conformational difference is seen both in the handedness and C1'–C1' distances, that are considerably shorter than those for regular double helices; both sequences have the same χ distribution centered at $\chi \simeq -120°$ and corresponding to anti -*ac* conformations. Now we consider the results for the G-rich sequences presented in Fig. 5.6: (i) Except for RNA GGC4, the other three duplexes experience some degree of unwinding, with DNA CGG4 closer to the A-form than DNA GGC4 (in other words, duplexes with GpC steps between the Watson-Crick base pairs tend to unwind slightly more than duplexes with CpG steps); (ii) syn χ values are centered around 40° for RNA and 72° for DNA, while anti χ values are centered around -161° for RNA and -104° for DNA GGC4 and -90° for DNA CGG4.

To further quantify these structures, we show the "simple twist" based on C1' atoms (see definition in the Supporting Material of Chapter 4) in Fig.5.7 for the middle steps

for DNA and RNA. For reference, ideal B-DNA has a twist with a value of 36°, and ideal A-RNA with a value of 31.5°. Immediately after equilibration, the twist quickly acquires sequence-dependent values. Convergence of the simulations is confirmed by the mirror symmetry of the twist around the central step (step 7) that reflects the inversion symmetry of the sequences. To describe the twist, we name the step types, starting with the general definition of Watson-Crick steps as L=GpC=GC/GC and M=CpG=CG/CG. In addition, we define "steps" containing mismatches as M_C=CG/CC=CC/CG (like a CpG step M but containing C mismatches) and L_C=GC/CC=CC/GC (like a GpC step L containing C mismatches). Thus, the pattern of steps (4-5-6-7-8-9-10) in Fig. 5.7 for $(CCG)_4$ is L-M_C-M_C-L-M_C-M_C-L, and for $(GCC)_4$ it is M-L_C-L_C-M-L_C-L_C-M. Proceeding in a similar manner for the G-rich sequences, we define M_G=CG/GG=GG/CG (like CpG step M with GG mismatches) and L_G=GC/GG=GG/GC (like a GpC step L containing GG mimatches). Thus, the pattern of steps (4-5-6-7-8-9-10) in Fig. 5.7 for $(CGG)_4$ is L-M_G-M_G-L-M_G-M_G-L, and for $(GGC)_4$ it is M-L_G-L_G-M-L_G-L_G-M. In the C-rich sequences, the twist is more uniform along the sequence, especially for the $(GCC)_4$ sequences with step pattern M-L_C-L_C-M-L_C-L_C-M. Sequences CCG4 experience increased twist at steps 5 and 9 (M_C step types) with twist decrease in the other steps both for RNA and DNA (although the differences are more marked for RNA). G-rich sequences, on the other hand, experience dramatic variation on the sequence-dependent twist, accompanied by some local unwinding. The GGC4 sequences experience a considerable decrease of twist at mismatch steps 6 and 8 (L_G step type) surrounding the central CpG step, with the twist in the other steps either staying close to the initial value or increasing. In CGG4 sequences, the decrease of twist at steps 6 and 8 (M_G step type) surrounding the central GpC step is even more pronounced, particularly for DNA. This is agreement with the twist behavior observed in DNA-$(CAG)_4$, where the most unwinding occurs in the mismatch steps surrounding the central GpC step [72].

To elucidate to what extent the mismatches distort the initial A-RNA and B-DNA forms we have carried out a principal component analysis [73] (PCA) on the backbone of the duplexes. Fig. 5.8 shows the distribution of conformations projected onto the first principal component for the backbone of the eight four-mismatch duplexes. This figure shows that the first eigenvector corresponds to the simultaneous coupling of bending and unwinding modes.

D. E-motif in DNA GCC. In a parallel paper [72] we present results about the

conformations, stability and dynamics of formation of the e-motif in DNA homoduplexes of TRs and hexanucleotide repeats (HRs). In an e-motif, the mismatched cytosines symmetrically flip out in the minor groove, pointing their base moieties towards the 5'-direction in each strand. E-motifs are not observed in RNA homoduplexes (at least not in sequences with chemically unmodified repeat sequences solvated in simple water solutions). Trinucleotide repeats have two reading frames, $(GCC)_n$ and $(CCG)_n$; while HRs have three: $(CCCGGC)_n$, $(CGGCCC)_n$, $(CCCCGG)_n$. We have defined three types of pseudo basepair steps related to the mismatches and show that the e-motif is only stable in $(GCC)_n$ and $(CCCGGC)_n$ homoduplexes due to the favorable stacking of pseudo GpC steps (whose nature depends on whether TRs or HRs are involved) and the formation of hydrogen bonds between the mismatched cytosine at position i and the cytosine (TRs) or guanine (HRs) at position $i-2$ along the same strand. In the complementary paper [72], we showed that the e-motif is stable under the three modifications of the DNA force field, mainly BSC0, BSC1 and OL15. In Fig. S0 in the present work, we show that free energy maps for these three force fields all share the same free energy minima (in terms of the dihedral angles χ) for the mismatch conformations. The main difference is that barriers between these minima are lowest for BSC0 and largest for BSC1, with OL15 providing some intermediate barriers. This suggests that BSC1 is perhaps "more rigid" in the description of mismatch conformations. Thus, transitions between mismatch conformations corresponding to different global and relative minima statistically will happen faster in BSC0 than in BSC1. Indeed, in this work we see spontaneous formation of e-motifs during regular molecular dynamics in a few hundreds of nanoseconds in trinucleotide repeats (GCC4) under the BSC0 force field (these also showed up for the HRs under the BSC0 force field).

Thus, the analysis in this paper and in the complementary paper indicate that the results under the BSC1 force field presented so far for the eight non-equivalent homoduplexes correspond to the equilibrium conformations, except for DNA GCC. For this particular sequence the intrahelical C·C mismatches under the BSC1 force field represent a metastable or transient conformation, while extrahelical e-motifs characterize the stable conformation. In Fig. 6.13 and Fig. 6.14 we show an extended e-motif studied in our previous work. The extrahelical C·C mismatches in an extended e-motif are stabilized by (i) pseudo GpC steps formed by the Watson-Crick basepairs adjacent to the mismatches; (ii) hydrogen bonds between the extruded C bases at a given position and

C bases belonging to Watson-Crick base pairs a few positions away from that; and (iii) by the stacking of the extruded C bases themselves. The pattern of stabilizing hydrogen bonds for GCC4 with an extended e-motif depends on the force field: OL15 displays consistently intra-strand C_i(N4)-$C_{(i-2)}$($O2$) bonding, BSC0 shows a mix of intra- and inter-strand bonding, and BSC1 shows inter-strand bondings between the N4 atom of the C_i mismatched base in one strand and the O4' atom of the second Watson-Crick paired C in the opposite strand (*i.e.,* C6-C27, C9-C24, etc.) [72].

E. Distribution of the neutralizing Na^+ ions. Now we consider the distribution of the neutralizing Na^+ ions. Figs. S21 to S28 show the distance between Na^+ ions to the center of mass of single mismatches. Different colors represent different ions in order to show the single-ion binding time for separate ions. Ions within a distance of 5 Å always have direct interactions with the bases in the mismatch. For the C·C mismatches, there is a larger presence of ions around RNA than around DNA. These figures indicate that the binding time for any single ion is short (except for the non-equilibrium DNA-GCC1 starting in syn-syn). For both RNA and DNA, there is more population of Na^+ ions around the G·G mismatches. Interestingly, for the equilibrium anti-syn sequences, ion binding in the GGC sequences is much longer than ion binding in the CGG sequences. The most important difference between the single-mismatch sequences in S21-S24 and the inner mismatches in the three-mismatch sequences, is that the latter do not display long-time ion binding for any of the conformations, a fact that we attribute to the enhanced flexibility of the multiple mismatches.

Fig.5.10 and Fig. 5.11 show the (any) ion occupancy of the single mismatches in RNA and DNA. If the mismatches stayed in the initial anti-anti or syn-syn conformations, ion distributions around C5/G5 (blue) should be the same as ion distribution around C14/G14 (red) due to the inversion symmetry of the single-mismatch duplexes (which is not present in the initial anti-syn conformations). For all the C-rich sequences, this clearly seen in the initial anti-anti conformations (that correspond to the minimum of the free energy) but not in the conformations that start in syn-syn, as these transition as explained above. In fact, deviations from this symmetry such as in Fig. 5.11 (c1) for DNA CGG1 starting in anti-anti correspond to "unhappy" conformations that are transitioning to the global equilibrium; in this case, anti-syn. For the C-rich conformations corresponding to equilibrium (anti-anti) the major binding sites are O2 and N3 for both RNA and DNA, with DNA displaying much stronger binding at these sites. For G-rich sequences

corresponding to the global minimum (anti-syn), considerable binding is seen at N7 and O6, the latter becoming a stronger attractor of Na^+ ions in the GGC sequences.

Some typical Na^+ ion binding conformations are shown in Fig. 5.12. Fig. 5.12(a) shows the binding to O2 and N3 atoms in the minor groove for a C·C mismatch in anti-anti conformation. This binding is found both in DNA and RNA, and is particularly high in DNA (with occupancy near 100%). Fig. 5.12(b) shows the binding of Na^+ to atoms O2, N3, O5' and OP2 in the C-base(syn) in the major groove of RNA-CCG in anti-syn conformation. This a highly populated conformation (see (a2) in Fig. 5.10); it may involve the four atoms as shown here or just two or three of those. For RNA-GCC there is a similar binding site, but closer to the backbone with less binding to N3. By contrast, the Na^+ ion binding to anti-syn DNA-CCG shown in Fig. 5.12(c) occurs in the minor groove and involves the O2 atom of the C base(anti) in the mismatch and a neighboring O2 atom of the C base belonging to an adjacent Watson-Crick basepair. Fig. 5.12(d) shows that in RNA-CGG and RNA-GGC, Na^+ binds to the N7 and O6 atoms in the major groove. This binding is very close to the backbone and always includes the neighboring OP2 atoms. Fig. 5.12(e) shows a particular high binding site comprised of N3, O6 and O4' atoms in the minor groove of DNA-CGG in anti-anti conformation. This is a very stable binding that also involves the O2 and N2 atoms of the neighboring Watson-Crick basepair and precludes the transition to the global minimum (anti-syn). This binding is only found in DNA-CCG because of its B-form shape and the way neighboring bases stack. Fig. 5.12(f) shows binding to the O6 atoms in the major groove for both RNA-CGG and DNA-CGG in anti-syn, while (g) shows a similar binding to (f), but as it occurs in GGC. The binding occupancy in GGC is much higher because Na^+ also binds a third G-O6 atom.

Figs. 5.13 and Fig. 5.14 show the ion cloud densities around the C·C and G·G mismatched duplexes, respectively. First, we consider the C-rich sequences. For RNA-CCG4 and RNA-GCC4, the main ion binding occurs in the major groove, although the cyan surfaces in the major groove are not directly connected to the mismatches. In RNA-CCG4, there are also some cyan surfaces in the minor groove near the mismatches, which correspond to the binding site in Fig. 5.12(a). For DNA-CCG4 and DNA-GCC4, these minor groove binding sites, as shown in Fig. 5.12(a), are more localized and obvious. DNA-CCG4 in (c) and DNA-GCC4 in (d) show an ion density highly localized around the mismatches. Pink surfaces with low ion densities are observed in Watson-Crick GpC steps (c), a behavior that is also observed in regular B-DNA. Next, we consider the G-rich

sequences. For all four structures, binding mainly occurs in the major groove, which corresponds to the binding site in Fig. 5.12(f)(g). Ion binding in RNA-GGC4 is more localized than in RNA-CGG4 because of the binding conformation in Fig. 5.12(g). This also explains why for DNA-GGC4, the cyan surfaces are more stretched in the direction of central axis than in DNA-CGG4. DNA also shows binding with lower density in the minor groove. For all but RNA-CGG4, ion binding reaches its highest density around the mismatches.

5.4 Discussion

As discussed in the introduction, sequences of the form d(CGG)·d(CCG) are overexpressed in the exons of the human genome. Expansions of CGG sequences lead to FXTAS in males [35], premature ovarian failure in females [36], and the inherited fragile X mental retardation syndrome [37]. CCG are related to FRAXE [38], Huntington's disease [39], and myotonic dystrophy type 1 [40]. In order to understand the mechanisms underlying sequence expansion, it is crucial to elucidate the secondary structure adopted by the the TR sequences both in DNA, where expansion originally occurs [41]; and in RNA, where the expansion leads to toxic RNA gain-of-function [6, 17–22]. Thus, a first step towards the understanding of these diseases involves the structural characterization of the atypical DNA and RNA structures.

The work presented here is part of our effort to achieve a unified and comparative description of the nucleic acid duplexes obtained from SSRs for both DNA and RNA, considering all the possible reading frames that result in CpG or GpC steps between the Watson-Crick base pairs, as shown in Fig. 6.1. The importance of the steps has been pointed out before. In the scheme introduced by Darlow and Leach [48, 49], hairpins were classified according to the alignment of the sides of the hairpins and the presence of an odd or even number of unpaired bases in the loop: "frame 1" corresponds to GpC steps between the Watson-Crick basepairs in the hairpin stem, while "frame 2" corresponds to CpG steps between the Watson-Crick basepairs in the stem (a "frame 3" presented not a single Watson-Crick basepair, which would therefore correspond to a considerably less stable structure). We have presented results for MD simulations and free energy calculations for both CCG and GGC trinucleotide repeats, with either CpG or GpC steps, for both RNA and DNA. This results in eight different non-equivalent helical duplexes.

Our main results are as follows.

1. The global minimum of the free energy maps associated with C·C mismatches in the four duplexes RNA/DNA CCG/GCC correspond to anti-anti conformations: anti *-ap* in the RNA duplexes and anti *-ac* in the DNA duplexes. In terms of the free energy, the next higher minimum corresponds to anti-syn conformations, while syn-syn conformations are even higher in energy. Anti-anti conformations are also observed in the extruded mismatches in the DNA e-motifs. Typical hydrogen bond conformations for the mismatches are given in Table II and Fig. 5.4 and Fig. 6.9.

2. The global minimum of the free energy maps associated with G·G mismatches in the four duplexes RNA/DNA GGC/CGG correspond to anti-syn conformations. In terms of the free energy, the next higher minimum corresponds to anti-anti conformations, while syn-syn conformations are even higher in free energy. The only exception for this is the phase diagram obtained for RNA-CGG, which includes only one repeat (sequence CGG1 in Fig. 6.1). In this case the anti-anti minimum is comparable to the anti-syn minima due to the stacking of three consecutive G's, rendered more stable by the considerable clamping exerted by the three G-C Watson-Crick basepairs on each side of the CGG sequence. Less constrained G-G mismatches in RNA-CGG4 exhibit the preferred anti-syn conformation during the 1μs regular MD. Typical hydrogen bond conformations for the mismatches are given in Table II and Fig. 5.4.

3. For DNA, the force fields BSC0, BSC1 and OL15 give similar free energy maps for the mismatch configurations. The three force fields predict the absolute minimum of the free energy maps in terms of the mismatch bases' χ angle to be anti-anti for the C·C mismatches and anti-syn for the G·G mismatches. The main difference is that the minima are deeper in BSC1 and OL15 providing for a more rigid DNA duplex with respect to that in BSC0. In addition, in the C-rich duplexes, the anti-syn minima are closer in depth to the anti-anti absolute minimum in BSC0 than in the other two fields; OL15 seems to be an intermediate case in this respect.

4. DNA duplexes in the GCC reading frame, with CpG steps between the Watson-Crick basepairs, exhibit the e-motif. In an e-motif, the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties towards the 5' direction in each strand. The phase diagrams based on the torsion angle χ are degenerate with respect to the intra- or extra-helical position of the mismatches.

Careful study of all the conformations obtained for both free energy maps and various MD simulations indicate that occasionally the C bases can be temporarily extruded as non-equilibrium duplexes (such as those started in syn-syn conformations) seek the global minimum. However, the only duplexes where the e-motif is stable under the three force fields [72] correspond to DNA GCC (paired) sequences. This corresponds to CpG steps between the Watson-Crick bonded basepairs and to *pseudo* GpC steps when the mismatches stack on the helix as the result of the C bases extrusion. The latter is the crucial factor in the stability of the e-motif: the pseudo GpC steps maximize helical stacking. The extruded C bases at position i further stabilize the helix by forming hydrogen bonds with Watson-Crick basepaired C bases in the 5' direction (position $(i-2)$ along the same strand for BSC0 and OL15, and position $(i-4)$ across strands for BSC1).

5. When mismatches are initially placed in non-equilibrium conformations, intra-helical C·C mismatches make the transition towards the global minimum faster than G·G mismatches. The G bases tend to form non-equilibrium stacking interactions that can considerably slow their evolution towards the equilibrium anti-syn conformation. On the other hand, the extrusion of the C mismatches in DNA GCC homoduplexes to form an e-motif can take from a few hundred of nanoseconds (BSC0) to microseconds or more (BSC1).

6. The mismatched duplexes exhibit characteristic sequence-dependent patterns where the disparities among the step twists are more pronounced in the G-rich sequences. Twist for all the helical duplexes are shown in Fig. 5.7. In the results section, we introduced the following notation: L=GpC=GC/GC; L_C=GC/CC=CC/GC and L_G=GC/GG=GG/GC (pseudo GpC step L containing either C or G mismatches); M=CpG=CG/CG; M_C=CG/CC=CC/CG and M_G=CG/GG=GG/CG (pseudo CpG step M containing either C or G mismatches). Thus, for steps (4-5-6-7-8-9-10) in Fig. 5.7, the step types are the following: $(CCG)_4$, L-M_C-M_C-L-M_C-M_C-L; $(GCC)_4$, M-L_C-L_C-M-L_C-L_C-M; $(CGG)_4$, L-M_G-M_G-L-M_G-M_G-L; and $(GGC)_4$, M-L_G-L_G-M-L_G-L_G-M. In the C-rich sequences, the twist is more uniform along the sequence. G-rich sequences, on the other hand, experience dramatic variation on the sequence-dependent twist, accompanied by some local unwinding. Both G-rich sequences for both DNA and RNA experience a considerable decrease of twist at mismatch steps 6 and 8, altough the decrease in CGG4 is more pronounced for the M_G steps surrounding the central GpC step than in GGC4 for the L_G step surrounding the central CpG step. This is agreement with the twist

behavior observed in DNA-$(CAG)_4$, where the most unwinding occurs in the mismatch steps surrounding the central GpC step [72]. The resulting unwinding of the DNA duplexes can also be observed in the handedness function shown in Fig. 5.6, where both CGG4 and GGC4 decrease their handedness with respect to the ideal B-DNA, but CGG4 becomes closer to A-DNA than GGC4.

Table 5.1 Main minima of the free energy maps for the single mismatch models.

A-A form		anti-anti	anti-syn		syn-syn
RNA-CCG1	approximate location (χ_5,χ_{14})	(-163,-163)	(-160,65)	(70,-163)	(63,60)
	relative free energy (kcal/mol)	0.0	5.9±0.1	4.6±0.1	10.4±0.1
	main H-bond	N3-H41,N4-H41	N3-H42		
DNA-CCG1	approximate location (χ_5,χ_{14})	(-122,-125)	(-128,70)	(70,-125)	(73,73)
	relative free energy (kcal/mol)	0.0	7.8±0.1	7.0±0.1	10.7±0.1
	main H-bond	N3-H41,N4-H41	N3-H42		
RNA-GCC1	approximate location (χ_5,χ_{14})	(-163,-163)	(-160,60)	(65,-160)	(63,61)
	relative free energy (kcal/mol)	0.0	5.8±0.3	5.0±0.1	9.0±0.6
	main H-bond	N3-H41,N4-H41	N3-H42		
DNA-GCC1	approximate location (χ_5,χ_{14})	(-122,-125)	(-142,64)	(64,-137)	(61,58)
	relative free energy (kcal/mol)	0.0	7.4±0.1	7.0±0.4	10.8±0.6
	main H-bond	N3-H41,N4-H41	N3-H42		
RNA-CGG1	approximate location (χ_5,χ_{14})	(-165,-165)	(-160,40)	(40,-158)	(45,45)
	relative free energy (kcal/mol)	0.0	0.2±1.0	0.0±0.7	9.8±0.2
	main H-bond	O6-H21	O6-H1,N7-H21,OP2-H22		
DNA-CGG1	approximate location (χ_5,χ_{14})	(-99,-105)	(-96,73)	(73,-96)	(67,67)
	relative free energy (kcal/mol)	2.7±0.1	0.0	1.4±0.3	5.6±0.2
	main H-bond	O6-H22	O6-H1,N7-H21		
RNA-GGC1	approximate location (χ_5,χ_{14})	(-165,-155)	(-160,40)	(40,-160)	(61,63)
	relative free energy (kcal/mol)	3.1±0.8	0.0	1.5±0.8	4.0±0.4
	main H-bond	O6-H21	O6-H1,N7-H21,OP2-H22		
DNA-GGC1	approximate location (χ_5,χ_{14})	(-116,-90)	(-113,70)	(70,-113)	(75,73)
	relative free energy (kcal/mol)	4.9±0.2	0.0	0.4±0.2	9.9±0.4
	main H-bond	O6-H21	O6-H1,N7-H21		

Table 5.2 H-bond percentage for the different conformations in the single mismatch sequences. AA stands for anti-anti and AS(SA) for anti-syn(syn-anti). The mismatch (Fig. 1) is B5-B14 (B is G or C base). For AA, the value represents the total H-bond. For AS or SA, the values in brackets correspond to B14(N3)-B5(H41) and outside the bracket to the complementarys B5(N3)-B14(H41). All the calculations use a 3.5 Å distance cutoff and a 135 degrees angle cutoff. Percentages less than 2% are not shown.

H-bond percentage(%)	RNA-CCG		RNA-GCC		DNA-CCG		DNA-GCC	
	AA	AS	AA	AS	AA	AS	AA	AS
N3-H41	50.5	-	47.8	-	95.5	-	87.9	-
O3-H41	28.5	-	6.2	-	12.7	-	9.0	-
N4-H41	20.7	-	35	(9.9)	16.1	(15)	33.2	(25.1)
N3-H42	-	7.3	-	20.0	-	43.4	-	67.6
O2-H42	-	-	-	-	-	8.2	-	17.1

H-bond percentage(%)	RNA-CGG			RNA-GGC			DNA-CGG			DNA-GGC		
	AA	AS	SA	AA	AS	SA	AA	AS	SA	AA	AS	SA
O6-H21	18.4	(28.9)	29.0	34.8	-	-	-	(14.8)	17.8	52.3	-	-
O6-H22	6.8	-	-	17.0	-	-	70.0	-	-	10.2	-	-
N2-H1	11.7	-	-	-	-	-	-	-	-	21.8	-	-
O6-H1	-	(86.5)	88.0	14.5	(65.4)	70.1	-	(87.0)	88.9	7.0	(79.1)	80.4
OP2-H22	-	(92.4)	92.0	-	(96.4)	96.3	-	(11.3)	8.8	-	(9.3)	10.1
N7-H21	-	(67.0)	67.8	-	(97.4)	97.4	-	(84.5)	81.4	-	(97.1)	96.7
N7-H1	-	-	-	-	(34.5)	31.1	-	(11.8)	9.1	-	(24.1)	23.0

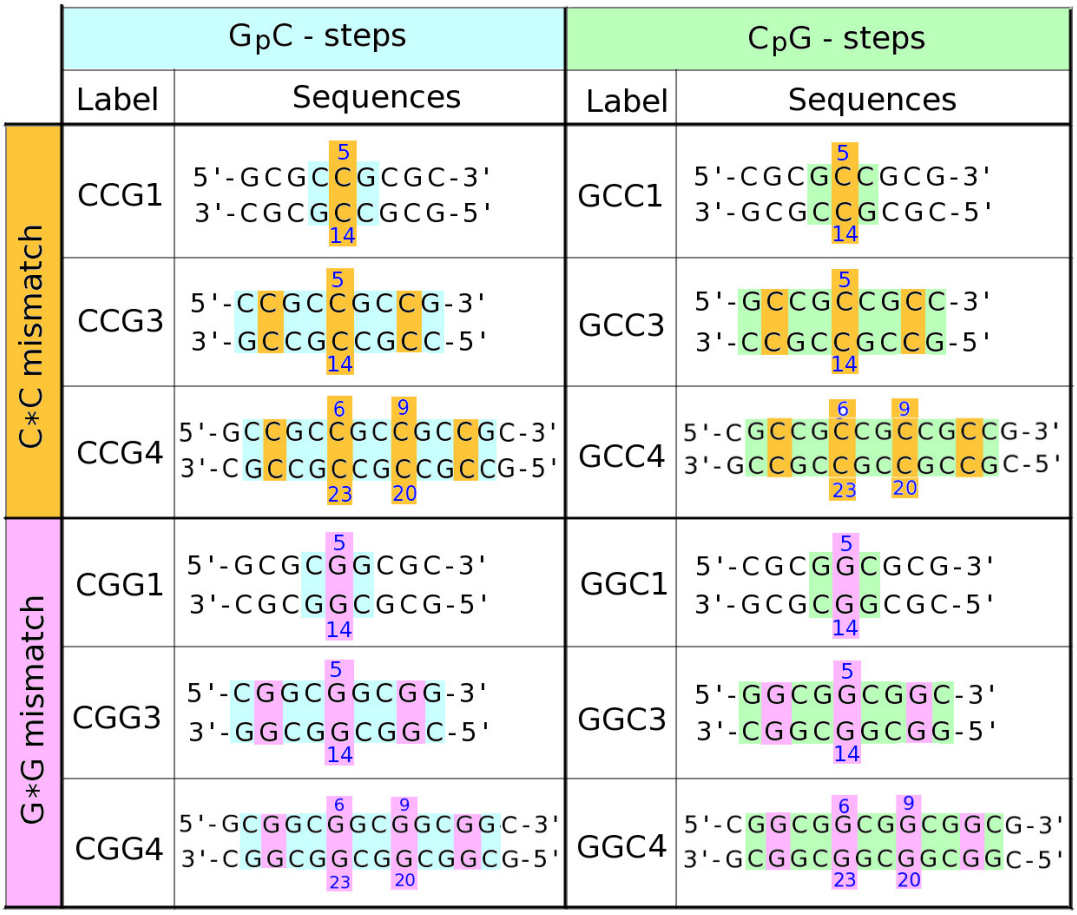


Figure 5.1 Nucleic acid sequences considered in this study.

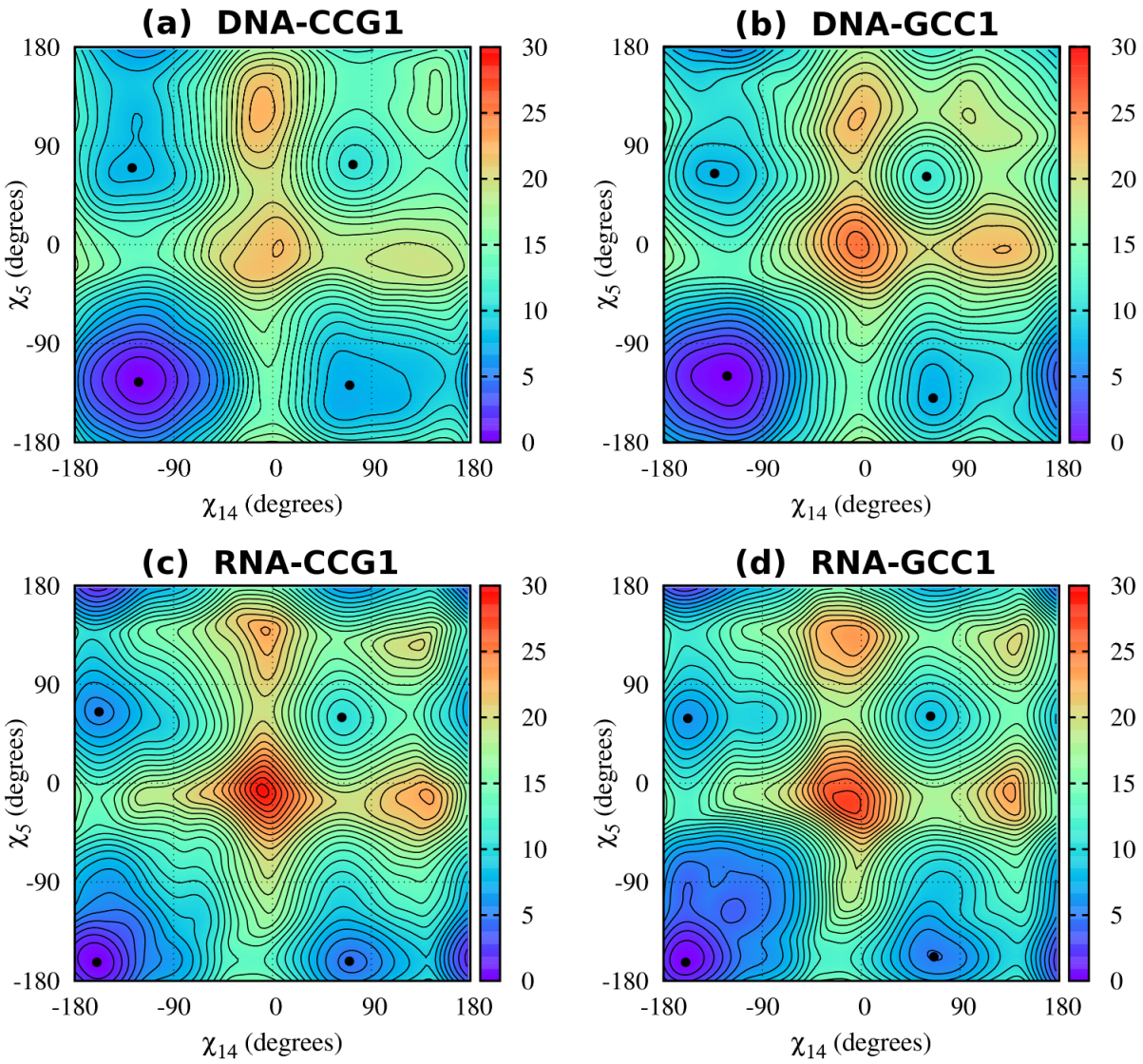


Figure 5.2 (χ_5, χ_{14}) free energy maps for single mismatches in DNA-CCG (a), DNA-GCC (b), RNA-CCG (c) and RNA-GCC (d).

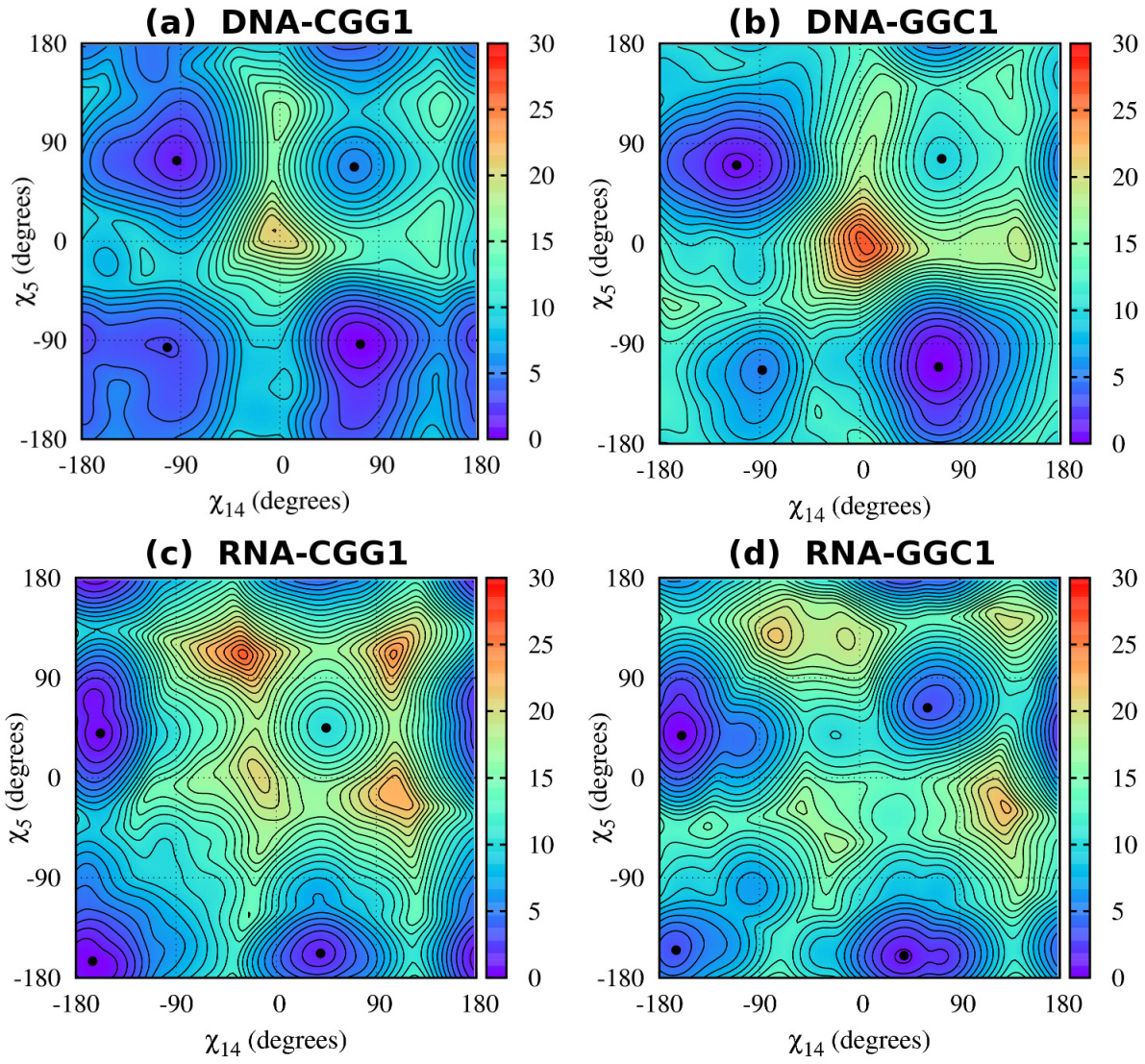


Figure 5.3 (χ_5, χ_{14}) free energy maps for single mismatches in DNA-CGG (a), DNA-GGC (b), RNA-CGG (c) and RNA-GGC (d).

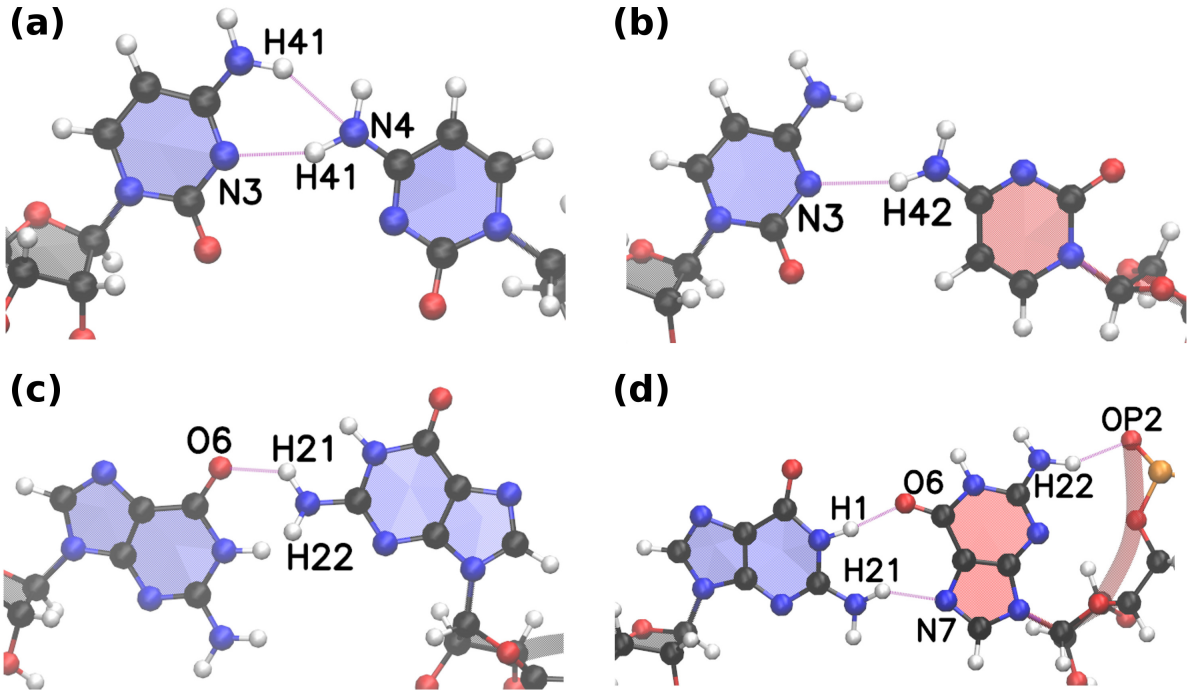


Figure 5.4 Atomic structures showing the main H-bonds for (a) CC mismatch (anti-anti) (b) CC mismatch (anti-syn) (c) GG mismatch (anti-anti) (d) GG mismatch (anti-syn). Anti bases are shown in blue and syn bases are shown in red.

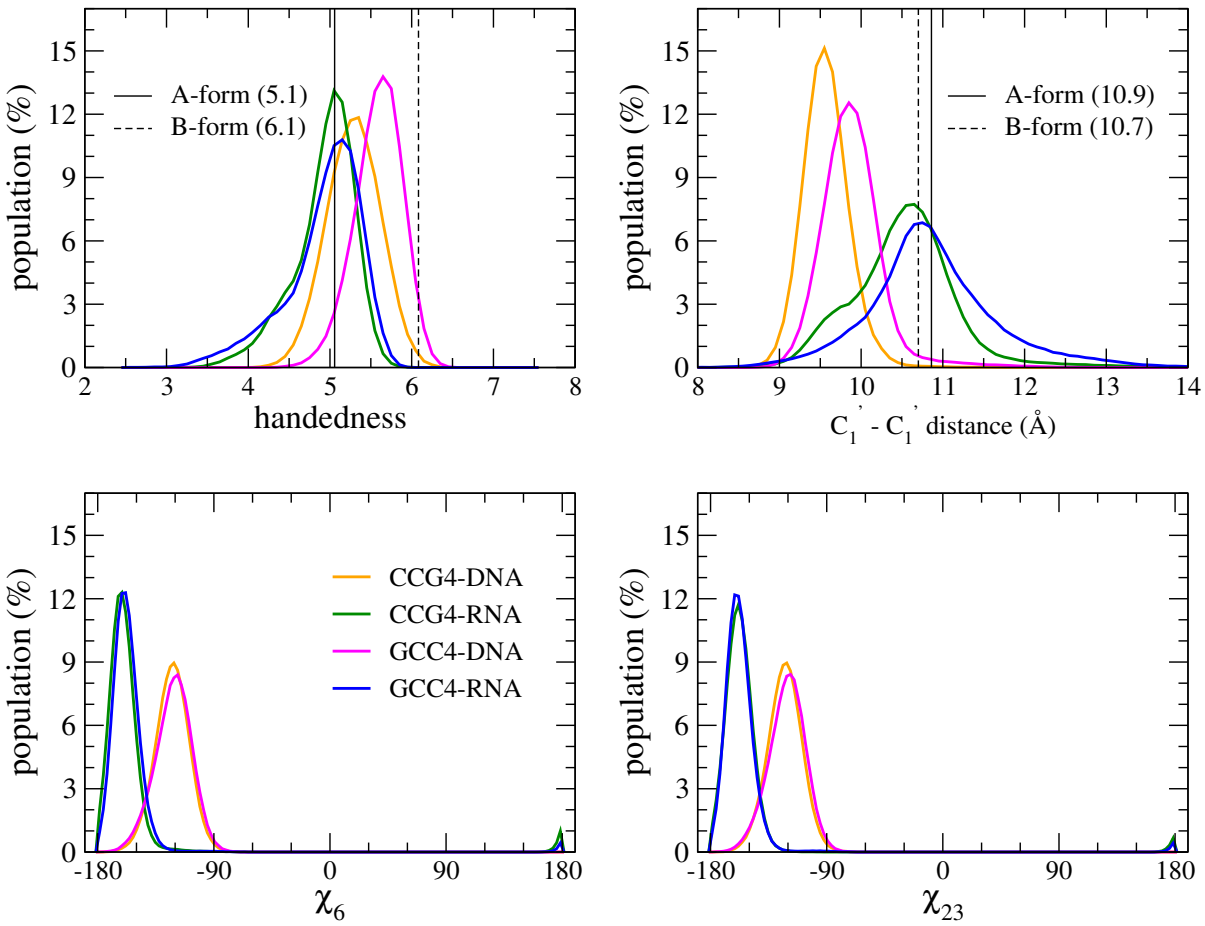


Figure 5.5 This graph shows the distribution of four TR helical duplexes CCG4 and GCC4 grouped by handedness, C_1'–C_1' distance, and χ_6 and χ_{20} dihedral angles (see Fig. 1). Handedness was calculated for the two central TRs. The curves are based on data from the last 800 ns of the two simulations for each sequence.

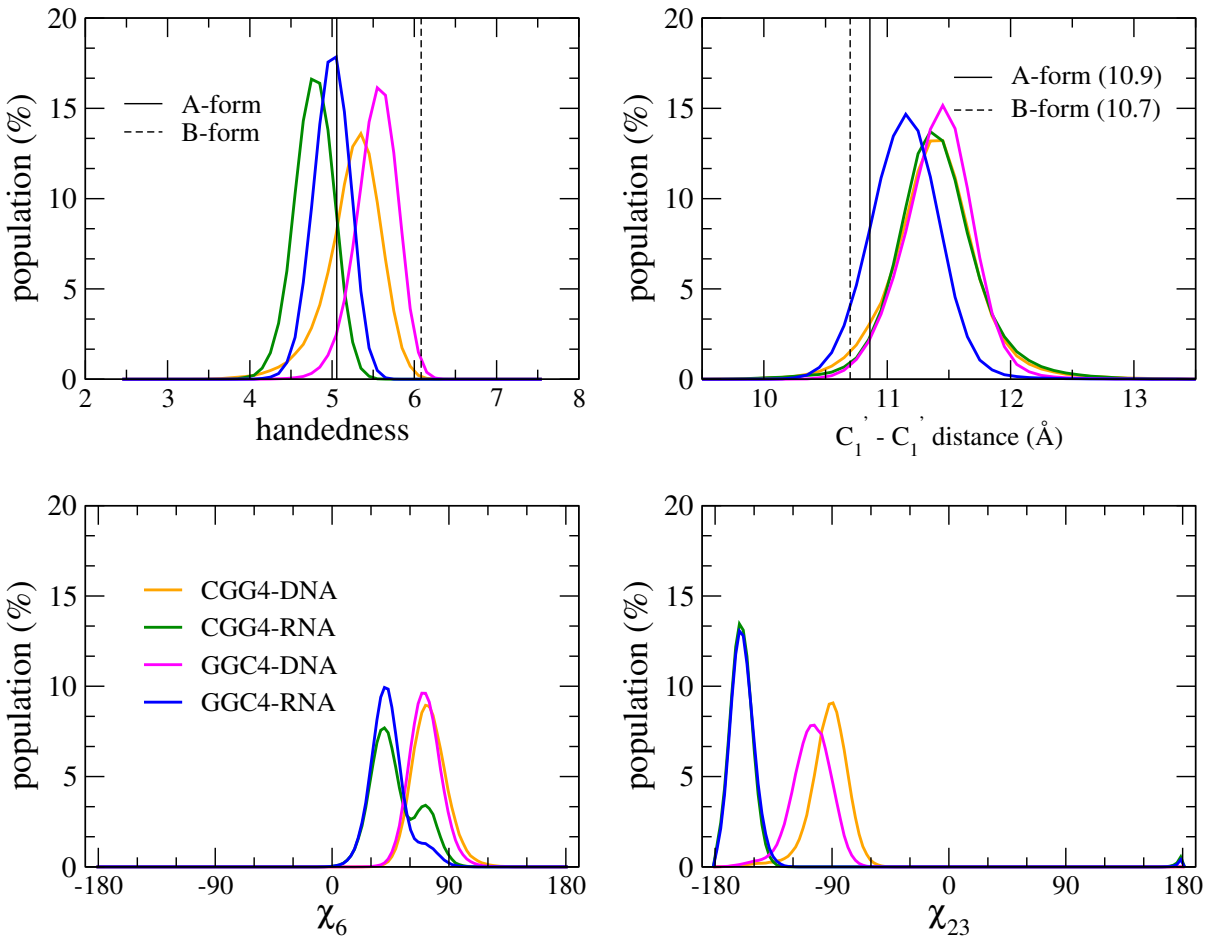


Figure 5.6 This graph shows the distribution of four TR helical duplexes CGG4 and GGC4 grouped by handedness, C_1'–C_1' distance, and χ_6 and χ_{20} dihedral angles (see Fig. 1). Handedness was calculated for the two central TRs. The curves are based on data from the last 800 ns of the two simulations for each sequence.

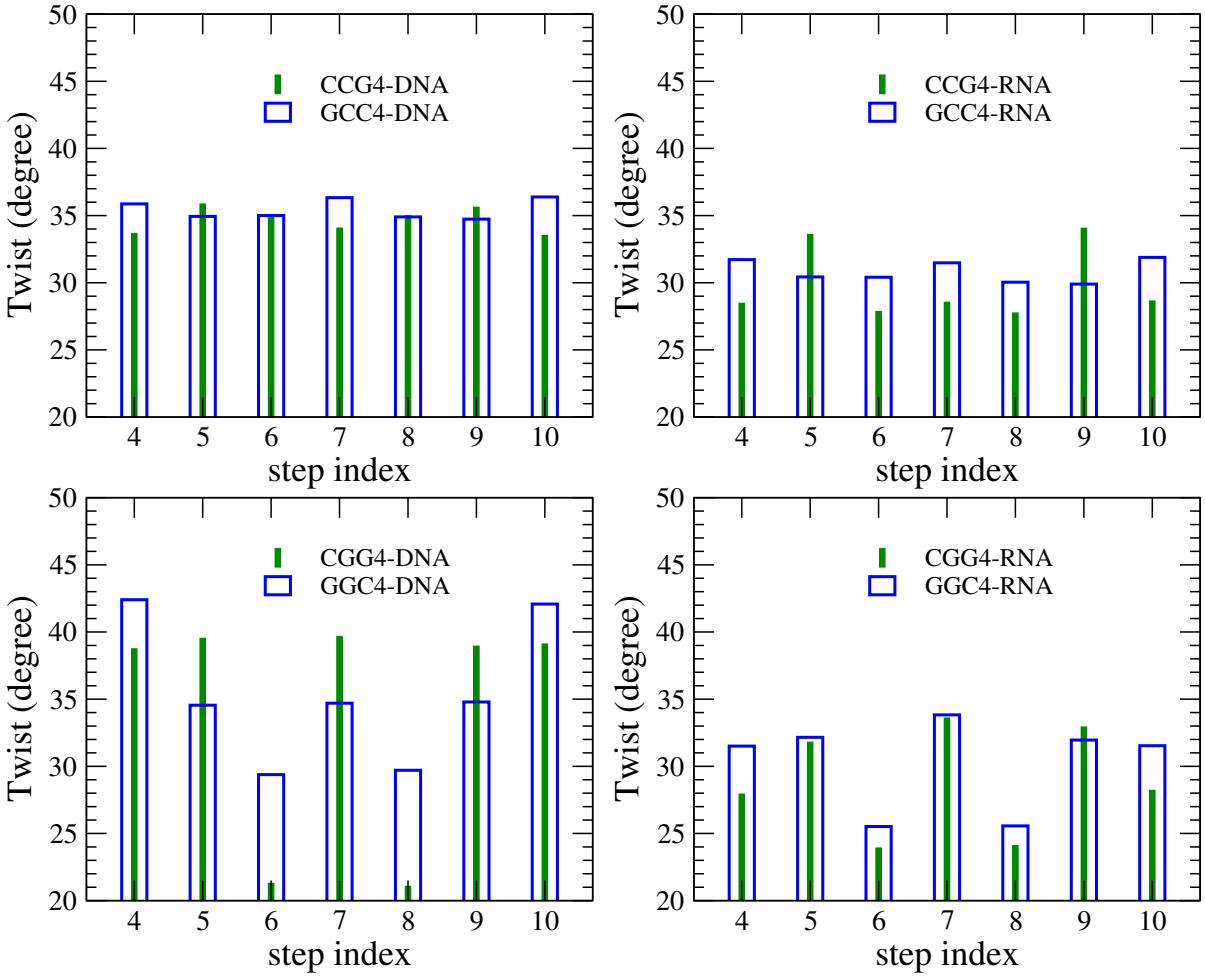


Figure 5.7 Simple twist in the four-mismatch homoduplexes. The data averaged over the last 800 ns of the two runs for each duplex.

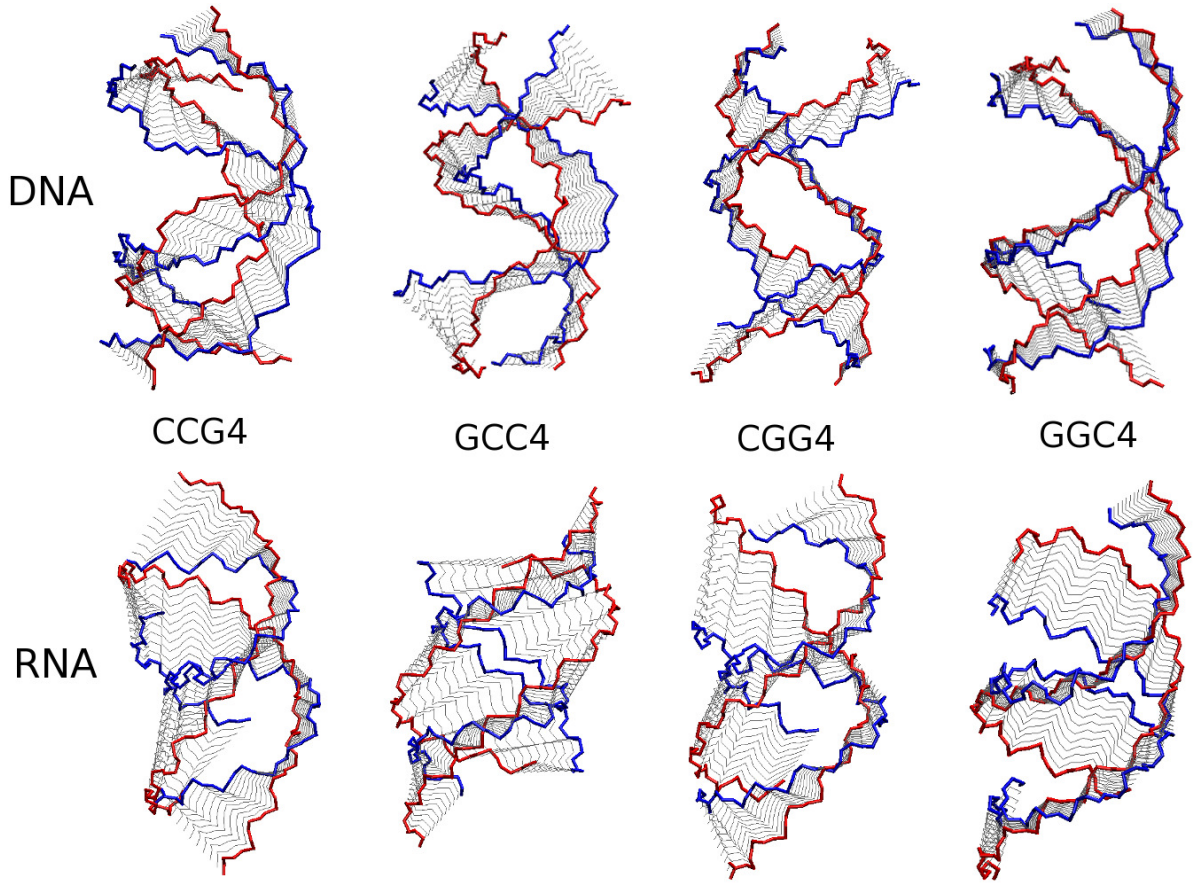


Figure 5.8 Conformational fluctuations around the first eigenvector direction based on the PCA analysis of the backbone of the four-mismatch homoduplexes.

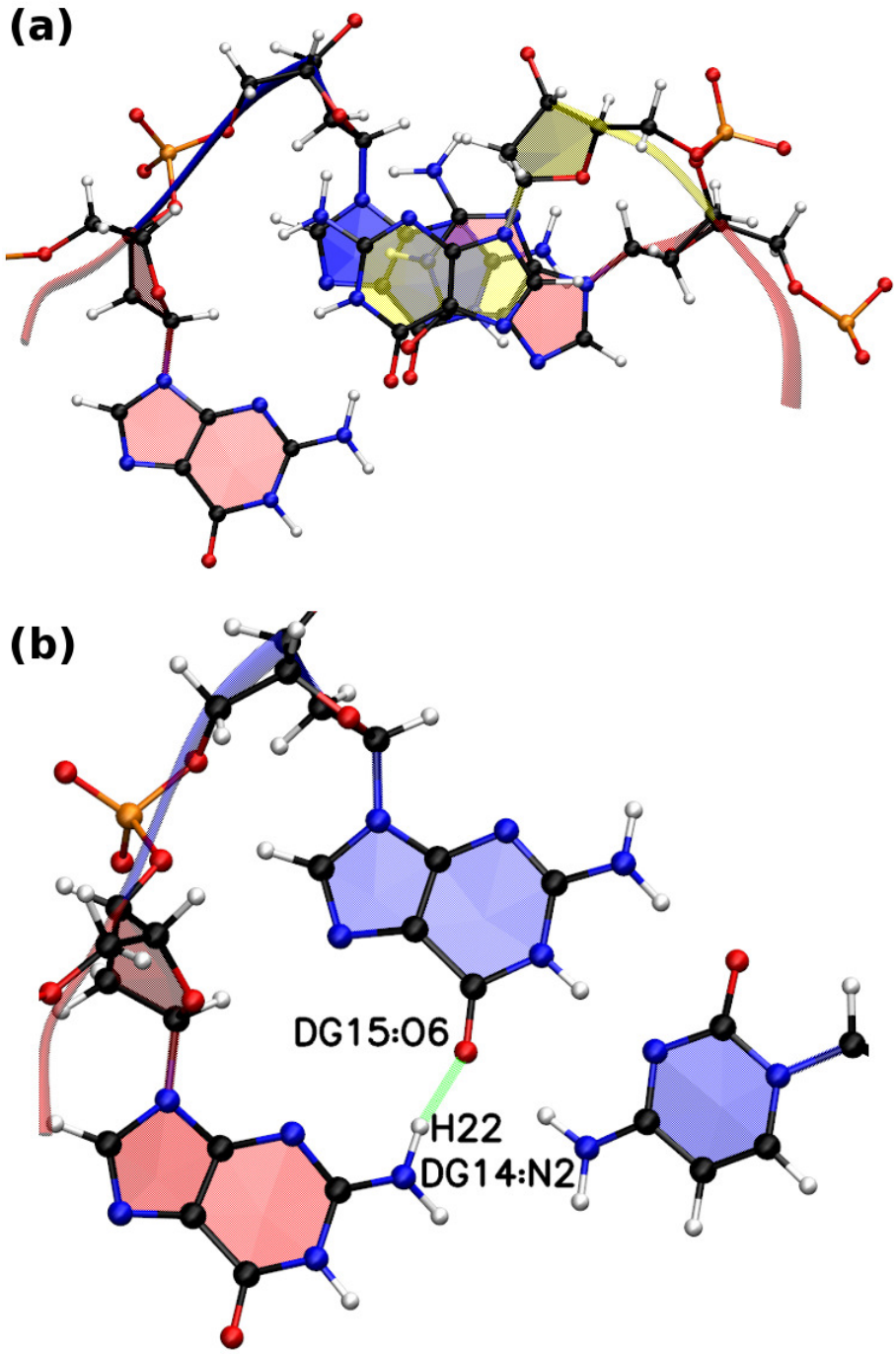


Figure 5.9 (a) Triple G stacking in DCGG(similar in RCGG). G5-G14 bases in red, G6 base in yellow and G15 base in blue. (b) Showing the H-bond between G14-N2 and G15-O6. G15-C4 WC pairs in blue and G14 base in red.

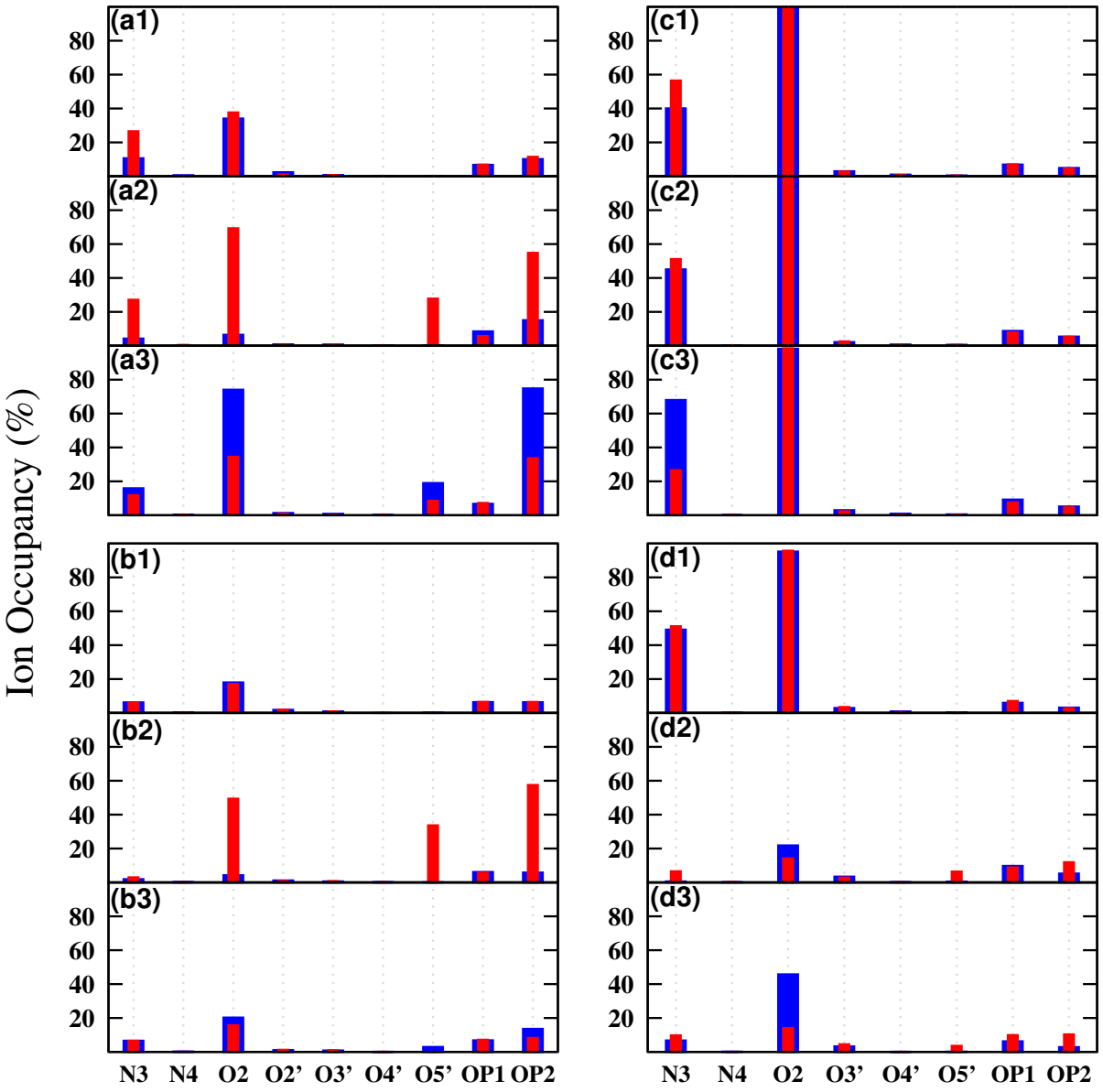


Figure 5.10 Ion occupancy around the single C-C mismatch in RNA and DNA. Blue: base C5. Red: base C14. RNA-CCG1: (a1) anti-anti; (a2) anti-syn (first 900ns); (a3) syn-syn. RNA-GCC1: (b1) anti-anti; (b2) anti-syn (first 700ns); (b3) syn-syn. DNA-CCG1: (c1) anti-anti; (c2) anti-syn; (c3) syn-syn. DNA-GCC1: (d1) anti-anti; (d2) anti-syn (first 150ns); (d3) syn-syn (first 450ns).

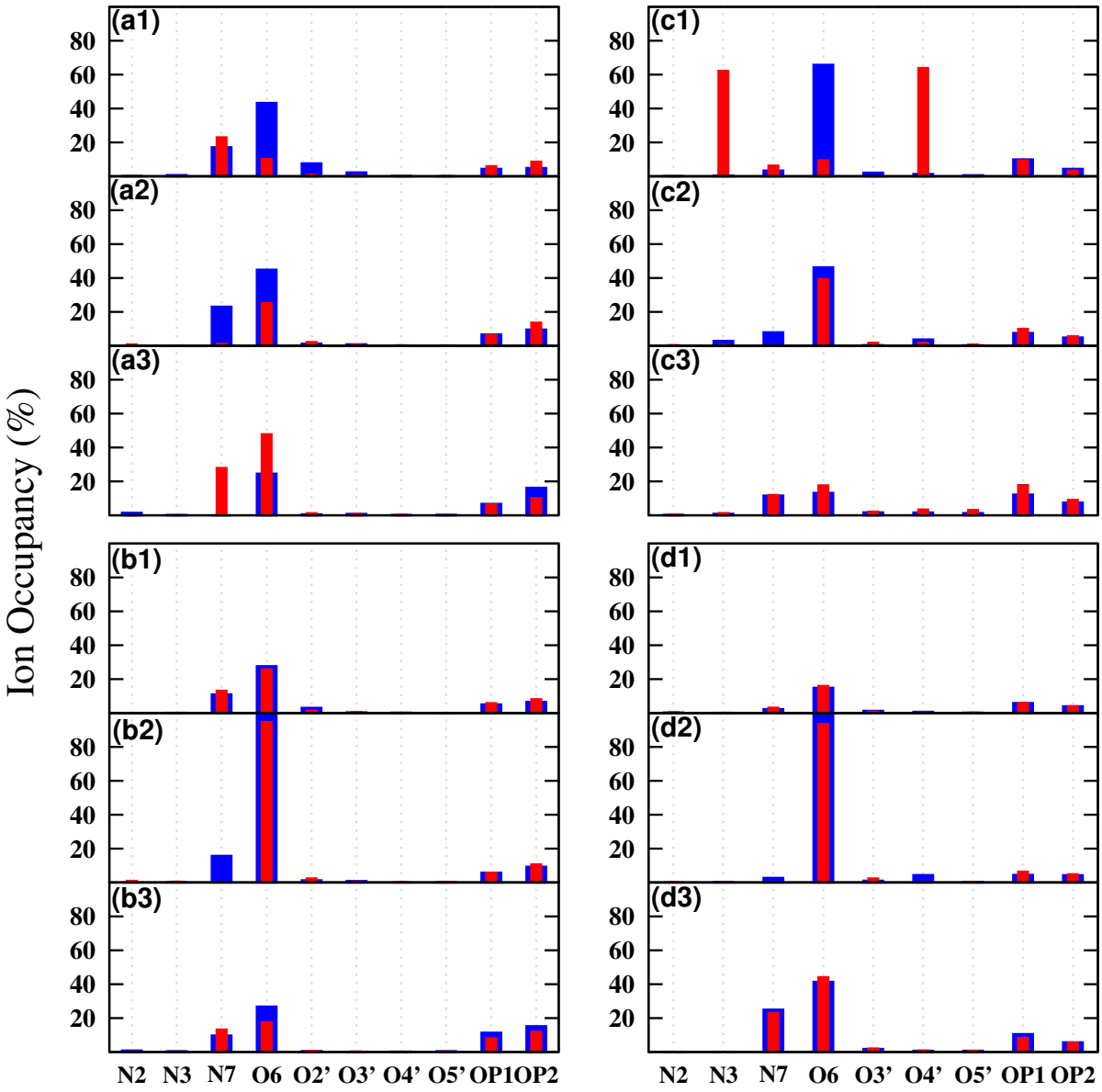


Figure 5.11 Ion occupancy around the single G-G mismatch in RNA and DNA. Blue: base G5. Red: base G14. RNA-CGG1: (a1) anti-anti; (a2) anti-syn; (a3) syn-syn. RNA-GGC1: (b1) anti-anti; (b2) anti-syn; (b3) syn-syn. DNA-CGG1: (c1) anti-anti; (c2) anti-syn; (c3) syn-syn (first 950 ns). DNA-GGC1: (d1) anti-anti; (d2) anti-syn; (d3) syn-syn.

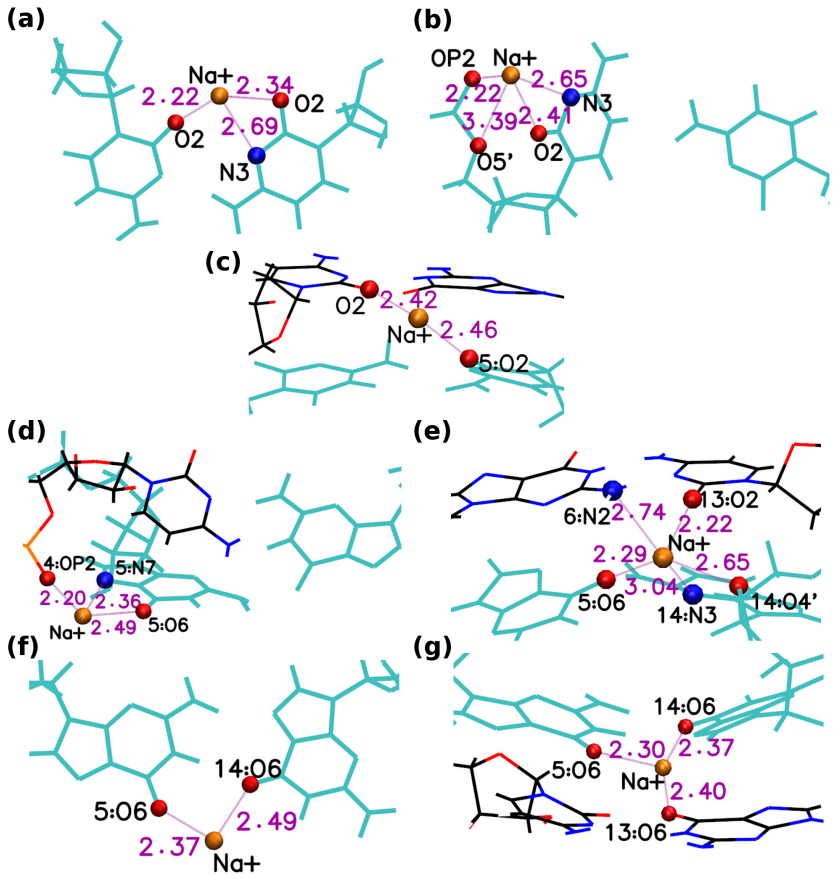


Figure 5.12 Some typical Na^+ ion binding sites. C-C or G-G mismatches are highlighted in cyan color and Na^+ ions are represented by orange spheres. (a) Binding to O2 and N3 atoms in minor groove for a C·C mismatch in anti-anti conformation, for both RNA and DNA. (b) Binding to O2, N3, O5' and OP2 of C-base(syn) in the major groove of RNA-CCG in anti-syn conformation. (c) Typical binding for DNA-CCG in anti-syn, that occurs in the minor groove. It involves the O2 atom of a mismatched C base(anti) and the neighboring O2 atom of a Watson-Crick C base. (d) For RNA-CGG and RNA-GGC, Na^+ binds to the N7 and O6 atoms in the major groove and the OP2 backbone atoms. (e) Binding to N3, O6 and O4' in the minor groove of DNA-CGG in anti-anti conformation. Binding also involves the O2 and N2 atoms of the neighboring Watson-Crick basepair. (f) Binding to O6 atoms in the major groove for both RNA-CGG and DNA-CGG in anti-syn. (g) Similar binding to (f), but in GGC. The binding occupancy in GGC is much higher because Na^+ also binds a third G-O6 atom.

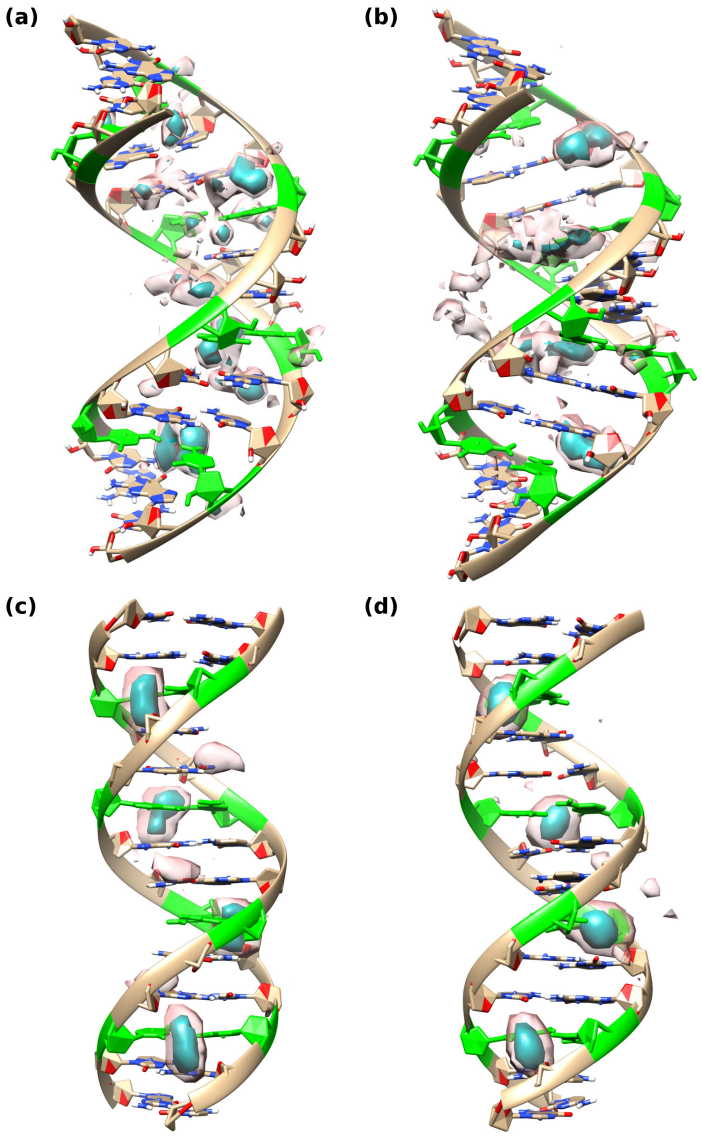


Figure 5.13 Ion cloud densities around the C·C mismatched duplexes. (a) RNA-CCG4; (b) RNA-GCC4; (c) DNA-CCG4; (d) DNA-GCC4. All the C·C mismatches (shown in green) are in anti-anti form. The cyan surface shows a high ion density and the pink surface shows a low ion density.

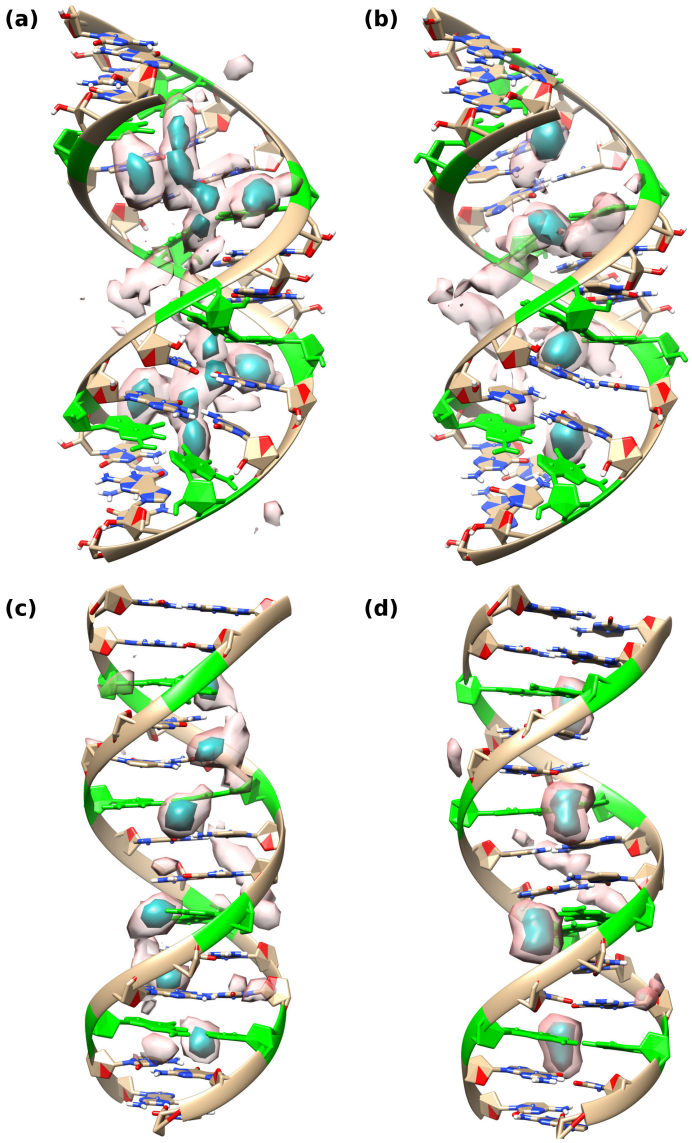


Figure 5.14 Ion cloud densities around the G·G mismatched duplexes. (a) RNA-CGG4; (b) RNA-GGC4; (c) DNA-CGG4; (d) DNA-GGC4. All G·G mismatches (shown in green) are in anti-syn form. The cyan surface shows a high ion density and the pink surface shows a low ion density.

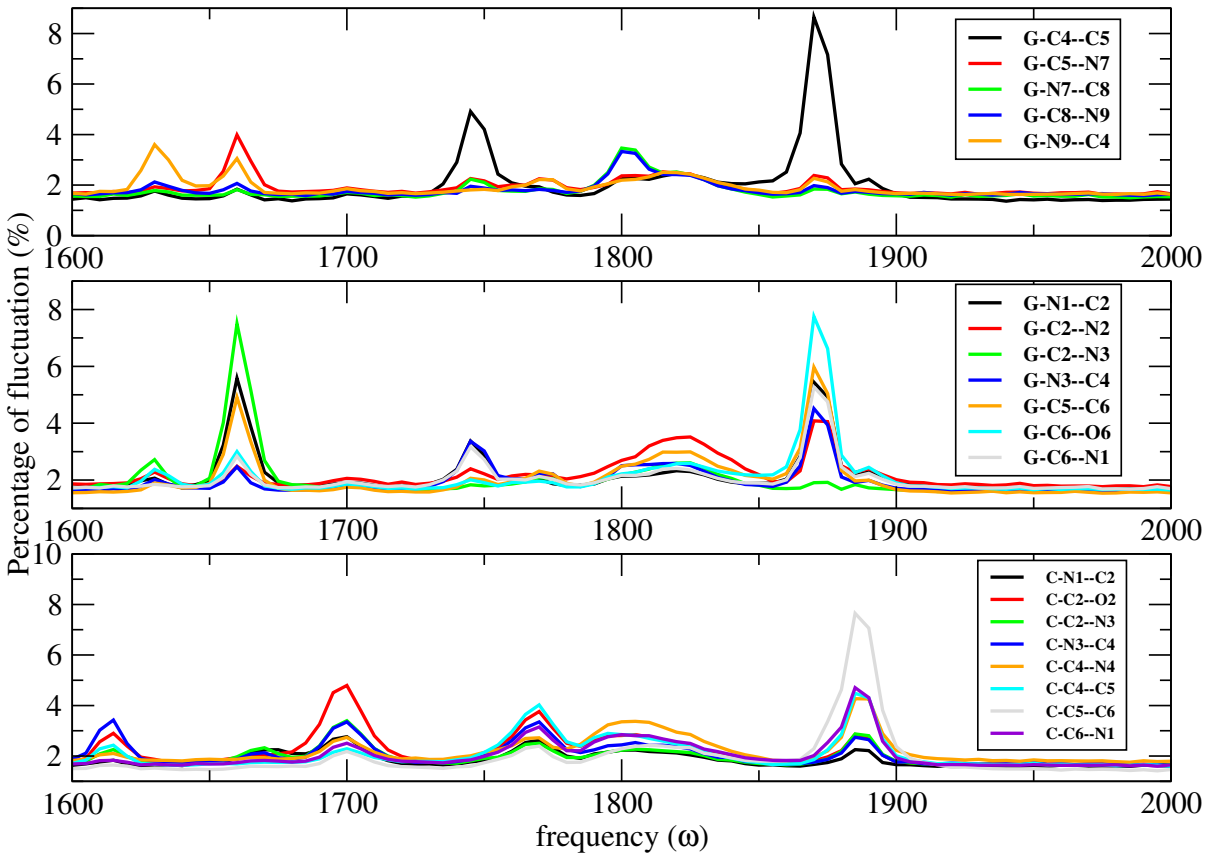


Figure 5.15 The scanning of frequency over different bonds in C-base and G-base of DNA-CCG. This shows the frequency of 1870 Hz gives a largest fluctuation of G-base bonds as well as a medium fluctuation to C-base and was chosen to be the one to do laser-melting.

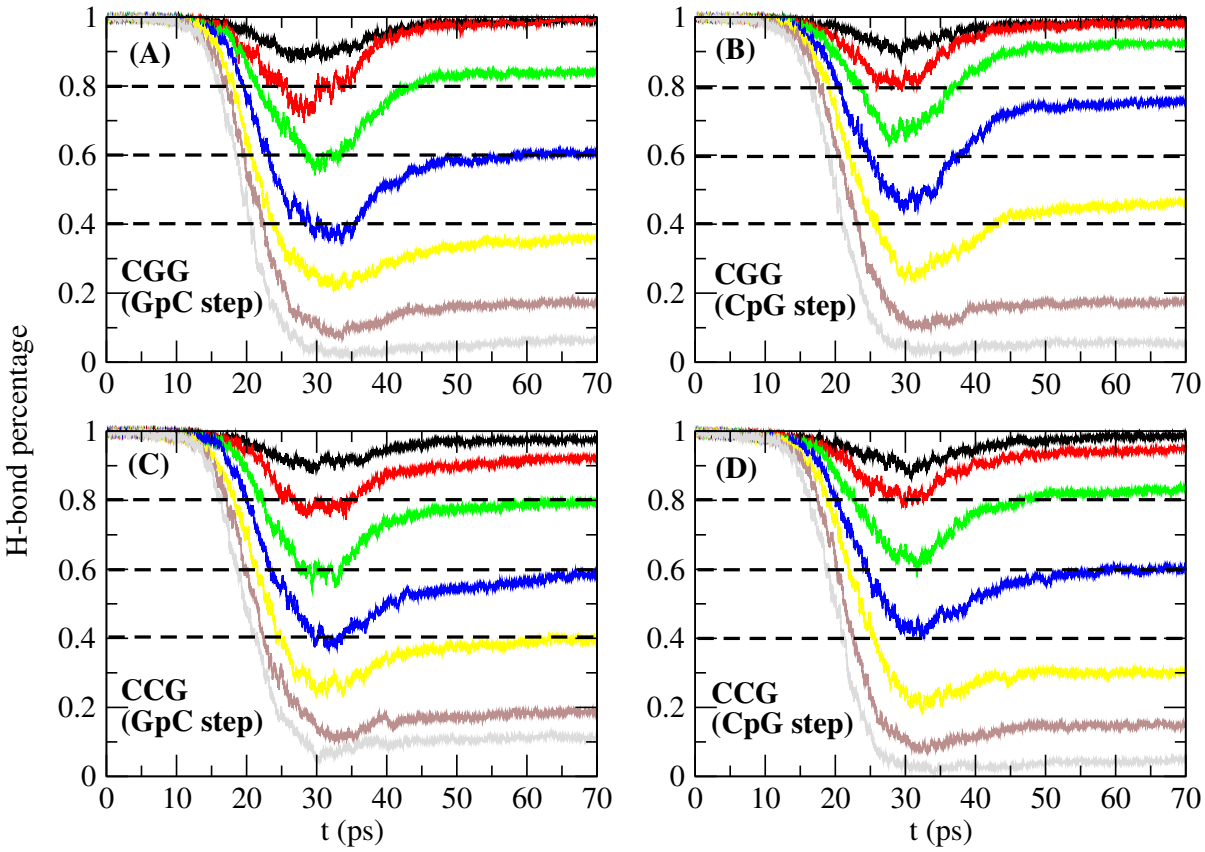


Figure 5.16 The H-bond percentage versus time in laser-melting simulation for all DNA sequences, this was carried out at force field ff99SBbsc0. CC is in anti-anti and GG is in anti-syn. This shows the relative stability of the lowest minima for all sequences, where DNA-GGC gives the most stable structure.

Acknowledgements

The work was supported by the National Institute of Health [NIH-R01GM118508]; the National Science Foundation (NSF) [SI2-SEE-1534941]; and the Extreme Science and Engineering Discovery Environment (XSEDE) [TG-MCB160064]. We thank the NC State HPC Center for computational support.

References

[1] H Ellegren. Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.*, 5:435–445, 2004.

[2] I Oberle, F. Rouseau, D. Heitz, D. Devys, S. Zengerling, and J.L. Mandel. Molecular-basis of the fragile-X syndrome and diagnostic applications. *Am. J. of Human Genet.*, 49:76, 1991.

[3] P Giunti, MG Sweeney, M Spadaro, C Jodice, A Novelletto, P Malaspina, M Frontali, and Harding AE. The trinucleotide repeat expansion on chromosome 6p (sca1) in autosomal dominant cerebellar ataxias. *Brain*, 117:645–649, 1994.

[4] V Campuzano, L Montermini, MD Molto, L Pianese, M Cossee, F Cavalcanti, E Monros, F Rodius, F Duclos, A Monticelli, F Zara, J Canizares, H Koutnikova, SI Bidichandani, C Gellera, A Brice, P Trouillas, G DeMichele, A Filla, R DeFrutos, F Palau, PI Patel, S DiDonato, JL Mandel, S Cocozza, M Koenig, and M Pandolfo. Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, 271:1423–1427, 1996.

[5] Sergei M. Mirkin. DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. in Struct. Biol.*, 16:351–358, 2006.

[6] Mirkin, S. Expandable DNA repeats and human disease. *Nature*, 447:932, 2007.

[7] CE Pearson, KN Edamura, and JD Cleary. Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.*, 6:729–742, 2005.

[8] Wells, R.D. and Warren, S. *Genetic instabilities and neurological diseases*. Academic Press, San Diego, CA, Elsevier, 1998.

[9] Orr, H. and Zoghbi, H. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, 30:575, 2007.

[10] CE Pearson and RR Sinden. Slipped strand DNA (S-DNA and SI-DNA), trinucleotide repeat instability and mismatch repair: A short review. In Sarma, RH and Sarma,

MH, editor, *Structure, Motion, Interaction and Expression of Biological Macromolecules, Vol 2*, pages 191–207. US NIH, 1998. 10th Conversation in Biomolecular Stereodynamics Conference, SUNY Albany, JUN 17-21, 1997.

[11] RD Wells, R Dere, ML Hebert, M Napierala, and LS Son. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucl. Acids Res.*, 33:3785–3798, 2005.

[12] Jane C. Kim and Sergei M. Mirkin. The balancing act of DNA repeat expansions. *Curr. Opin. in Genet. & Devel.*, 23:280–288, 2013.

[13] JP Cleary, DM Walsh, JJ Hofmeister, GM Shankar, MA Kuskowski, DJ Selkoe, and KH Ashe. Natural oligomers of the amyloid-protein specifically disrupt cognitive function. *Nat. Neurosci.*, 8:79–84, 2005.

[14] Vincent Dion and John H. Wilson. Instability and chromatin structure of expanded trinucleotide repeats. *Trends in Genet.*, 25:288–297, 2009.

[15] Cynthia T. McMurray. Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair*, 7:1121–1134, 2008.

[16] Yunfu Lin and John H Wilson. Transcription-induced DNA toxicity at trinucleotide repeats: double bubble is trouble. *Cell Cycle*, 10:611–618, 2011.

[17] Laura P. W. Ranum and Thomas A. Cooper. RNA-mediated neuromuscular disorders. *Ann. Rev. of Neuroscience*, 6:259–277, 2006.

[18] Ling-Bo Li and Nancy M. Bonini. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends in Neurosciences*, 33:292–298, 2010.

[19] P Jin, DC Zarnescu, FP Zhang, CE Pearson, JC Lucchesi, K Moses, and ST Warren. RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in Drosophila. *Neuron*, 39:739–747, 2003.

[20] H Jiang, A Mankodi, MS Swanson, RT Moxley, and CA Thornton. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum. Mol. Genet.*, 13:3079–3088, 2004.

[21] R. Daughters, D. Tuttle, W. Gao, Y. Ikeda, M. Moseley, T. Ebner, M. Swanson, and L. Ranum. RNA Gain-of-Function in Spinocerebellar Ataxia Type 8. *PLoS Genet.*, 5:e1000600, 2009.

[22] W. Krzyzosiak, K. Sobczak, M. Wojciechowska, A. Fiszer, A. Mykowska, and P. Kozlowski. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nuc. Acids Res.*, 40:11–26, 2012.

[23] V Campuzano, L Montermini, Y Lutz, L Cova, C Hindelang, S Jiralerspong, Y Trottier, SJ Kish, B Faucheux, P Trouillas, FJ Authier, A Durr, JL Mandel, A Vescovi, M Pandolfo, and M Koenig. Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum. Mol. Genet.*, 6:1771–1780, 1997.

[24] E. Kim, M. Napierala, and S. Dent. Hyperexpansion of GAA repeats affects post-initiation steps of FXN transcription in Friedreich's ataxia. *Nucl. Acids Res.* , 39:8366–8377, 2011.

[25] D. Kumari, R. Biacsi, and K. Usdin. Repeat expansion in intron 1 of the Frataxin gene reduces transcription initiation in Friedreich ataxia. *Faseb J.*, 25:895, 2011. Experimental Biology Meeting 2011, Washington, DC, APR 09-13, 2011.

[26] T. Punga and M. Buehler. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *Embo Mol. Medicine*, 2:120–129, 2010.

[27] Jason R. O'Rourke and Maurice S. Swanson. Mechanisms of RNA-mediated Disease. *J. Biol. Chem.*, 284(12):7419–7423, 2009.

[28] Peter K. Todd and Henry L. Paulson. RNA-mediated neurodegeneration in repeat expansion disorders. *Ann. Neurol.*, 67(3):291–300, 2010.

[29] Gloria V. Echeverria and Thomas A. Cooper. RNA-binding proteins in microsatellite expansion disorders: mediators of RNA toxicity. *Brain Res.*, 1462:100–111, 2012.

[30] Marzena Wojciechowska and Wlodzimierz J. Krzyzosiak. Cellular toxicity of expanded RNA repeats: focus on RNA foci. *Hum. Mol. Genet.*, 20(19):3811–3821, 2011.

[31] Tao Zu, Brian Gibbens, Noelle S. Doty, Mario Gomes-Pereira, Aline Huguet, Matthew D. Stone, Jamie Margolis, Mark Peterson, Todd W. Markowski, Melissa A. C. Ingram, Zhenhong Nan, Colleen Forster, Walter C. Low, Benedikt Schoser, Nikunj V. Somia, H. Brent Clark, Stephen Schmechel, Peter B. Bitterman, Genevieve Gourdon, Maurica S. Swanson, Melinda Moseley, and Laura P. W. Ranum. Non-atg–initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci USA*, 108(1):260–265, 2011.

[32] Ying-Hui Fu, Derek P.A. Kuhl, Antonio Pizzuti, Maura Pieretti, James S. Sutcliffe, Stephen Richards, Annemieke J.M.H. Verkert, Jeanette J.A. Holden, Raymond G. Fenwick Jr., Stephen T. Warren, Ben A. Oostra, David L. Nelson, and C.Thomas Caskey. Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. *Cell*, 67(6):1047–1058, 1991.

[33] Nan Zhong, Weina Ju, James Pietrofesa, Daowen Wang, Carl Dobkin, and W. Ted Brown. Fragile X "gray zone" alleles: AGG patterns, expansion risks, and associated haplotypes. *Am J Med Genet.*, 64(2):261–5, 1996.

[34] C. Dombrowski, S. Lévesque, M. L. Morel, P. Rouillard, K. Morgan, and F. Rousseau. Premutation and intermediate-size FMR1 alleles in 10 572 males from the general population: loss of an AGG interruption is a late event in the generation of fragile X syndrome alleles. *Hum Mol Genet.*, 11(4):371–378, 2002.

[35] RJ Hagerman, M Leehey, W Heinrichs, F Tassone, R Wilson, J Hills, J Grigsby, B Gage, and PJ Hagerman. Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X. *Neurology*, 57(1):127–30, 2001.

[36] Stephanie L. Sherman. Premature Ovarian Failure among Fragile X Premutation Carriers: Parent-of-Origin Effect? *Am J Hum Genet.*, 67(1):11–3, 2000.

[37] I.A. Glass. X linked mental retardation. *J. Med. Genet.*, 28:361–371, 1991.

[38] Yanghong Gu, Ying Shen, Richard A. Gibbs, and David L. Nelson. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nat. Genet.*, 13(1):109–113, 1996.

[39] Baorong Zhang, Jun Tian, Yaping Yan, Xinzhen Yin, Guohua Zhao, Zhiying Wu, Weihong Gu, Kun Xia, and Beisha Tang. CCG polymorphisms in the huntingtin gene have no effect on the pathogenesis of patients with Huntington's disease in mainland Chinese families. *J. Neurol. Sci.*, 312(1-2):92–96, 2012.

[40] Claudia Braida, Rhoda K.A. Stefanatos, Berit Adam, Navdeep Mahajan, Hubert J.M. Smeets, Florence Niel, Cyril Goizet, Benoit Arveiler, Michel Koenig, Clotilde Lagier-Tourenne, Jean-Louis Mandel, Catharina G. Faber, Christine E.M. de Die-Smulders, Frank Spaans, and Darren G. Monckton. Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum. Mol. Genet.*, 19(1-2):1399–1412, 2010.

[41] CT McMurray. DNA secondary structure: A common and causative factor for expansion in human disease. *Proc. Natl. Acad. Sci. USA*, 96:1823–1825, 1999.

[42] M Mitas. Trinucleotide repeats associated with human disease. *Nuc. Acids Res.*, 25:2245–2253, 1997.

[43] Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak, and Wojciech Rypniewski. Crystal structures of CGG RNA repeats with implications for fragile X-associated tremor ataxia syndrome. *Nuc. Acids Res.*, 39:7308–7315, 2011.

[44] Amit Kumar, Pengfei Fang, Hajeung Park, Min Guo, Kendall W. Nettles, and Matthew D. Disney. A crystal structure of a model of the repeating r(cgg) transcript found in fragile syndrome. *Chembiochem.*, 12(14):2140–2142, Sep. 2011.

[45] Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak, and Wojciech Rypniewski. Crystallographic characterization of CCG repeats. *Nucl. Acids Res.*, 40:8155–8162, 2012.

[46] XL Gao, XN Huang, GK Smith, MX Zheng, and HY Liu. New antiparallel duplex motif of DNA CCG repeats that is stabilized by extrahelical basis symmetrically located in the minor-groove. *J. Am. Chem. Soc.*, 117:8883–8884, 1995.

[47] MX Zheng, XN Huang, GK Smith, XY Yang, and XL Gao. Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study. *J. of Mol. Biol.*, 264:323–336, 1996.

[48] JM Darlow and DRF Leach. Secondary structures in d(CGG) and d(CCG) repeat tracts. *J. Mol. Biol.*, 275:3–16, 1998.

[49] JM Darlow and DRF Leach. Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts in vivo. *J. Mol. Biol.*, 275:17–23, 1998.

[50] Daniel Svozil, Pavel Hobza, and Jiří Šponer. Comparison of intrinsic stacking energies of ten unique dinucleotide steps in A-RNA and B-DNA duplexes. Can we determine correct order of stability by quantum-chemical calculations? *J. Phys. Chem. B*, 114(2):1191–203, 2010.

[51] F. Pan and V. Man and C. Roland and C. Sagui. Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats. *Biophys. J.*, 113:19–36, 2017.

[52] Y. Zhang, C. Roland, and C. Sagui. Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases. *ACS Chemical Neuroscience*, 8:578–591, 2016.

[53] Huda Y. Zoghbi and Harry T. Orr. Glutamine repeats and neurodegeneration. *Annual Review of Neuroscience*, 23(1):217–247, 2000.

[54] E. Delot, L. M. King, M. D. Briggs, W. R. Wilcox, and D. H. Cohn. Trinucleotide Expansion Mutations in the Cartilage Oligomeric Matrix Protein (Comp) Gene. *Hum Mol Genet*, 8:123–128, 1999.

[55] Mariely DeJesus-Hernandez, Ian R. Machkenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, NiCole A. Finch, Heather

Flynn, Jennifer Adamson, Naomi Kouri, Aleksandra Wojtas, Pheth Sengdy, Ging-Yuek R. Hsiung, Anna Karydas, William W. Seeley, Keith A. Josephs, Giovanni Coppola, Daniel H. Geschwind, Zbigniew K. Wszolek, Howard Feldman, David S. Knopman, Ronald C. Petersen, Bruce L. Miller, and Dennis W. Dickson. Expanded ggggcc hexanucleotide repeat in noncoding region of c9orf72 causes chromosome 9p-linked ftd and als. *Neuron*, 72(2):245–256, Oct. 2011.

[56] Alan E. Renton, Elisa Majounie, Adrian Waite, Javier Simon-Sanchez, Sara Rollinson, J. Raphael Gibbs, Jennifer C. Schymick, Hannu Laaksovirta, John C. van Swieten, Liisa Myllykangas, Hannu Kalimo, Anders Paetau, Yevgeniya Abramzon, Anne M. Remes, Alice Kaganovich, Sanja W. Scholz, Jamie Duckworth, Jinhui Ding, Daniel W. Harmer, Dena G. Hernandez, Janel O. Johnson, Kin Mok, Mina Ryten, Danyah Trabzuni, and Rita J. Guerreiro. A hexanucleotide repeat expansion in c9orf72 is the cause of chromosome 9p21-linked als-ftd. *Neuron*, 72(2):257–268, Oct. 2011.

[57] D. A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T.Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *"AMBER 16"*. University of California, San Francisco, 2016.

[58] I. Ivani, P. Dans, A. Noy, A. Prez, I. Faustino, A. Hopsital, J. Walther, P. Andrio, R. Goni, A. Balaceanu, G. Portella, F. Battistini, J. GelpA, C. Gonzlez, M. Vendruscolo, C. Laughton, S. Harris, D. Case, and M. Orozco. Parmbsc1: A refined force field for DNA simulations. *Nature Meth.*, 13:55–58, 2016.

[59] A. Perez, I. Marchan, 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.*, 92:3817–3829, 2007.

[60] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham, and P. Jurecka. 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.

[61] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham III, Rodrigo Galindo-Murillo, and Petr 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.*, 11(12):5723–36, 2015.

[62] W. L. Jorgensen, J. Chandrasekhar, J. Madura, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926 – 935, 1983.

[63] I. S. Joung and T. E. Cheatham. Determination of Alkali and Halide Monovalent Ion Parameters for use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B*, 112:9020–9041, 2008.

[64] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577 – 8593, 1995.

[65] V. Babin, C. Roland, and C. Sagui. Adaptively Biased Molecular Dynamics for free energy calculations. *J. Chem. Phys.*, 128:134101, 2008.

[66] V Babin, V Karpusenka, M Moradi, C Roland, and C Sagui. Adaptively Biased Molecular Dynamics: An umbrella sampling method with a time-dependent potential. *Int. J. Quantum Chem.*, 109:3666–3678, 2009.

[67] Paolo Raiteri, Alessandro Laio, Francesco Luigi Gervasio, Cristian Micheletti, and Michele Parrinello. Efficient reconstruction of complex free energy landscapes by multiple walkers Metadynamics. *J. Phys. Chem.*, 110:3533 – 3539, 2006.

[68] K. Minoukadeh and Ch. Chipot and T. Lelievre. Potential of Mean Force Calculations: A multiple-walker adaptive biasing force technique. *J. Chem. Theor. and Comput.*, 6:1008, 2010.

[69] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141, 1999.

[70] A. Barducci, G. Bussi, and M. Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free energy method. *Phys. Rev. Lett.*, 100:020603, 2008.

[71] F. Pan and C. Roland and C. Sagui. Ion distribution around left- and right-handed DNA and RNA duplexes: a comparative study. *Nucl. Acids Res.*, 42:13981–96, 2014.

[72] F. Pan, Y. Zhang, V. Man, C. Roland, and C. Sagui. E-motif formed by extrahelical cytosine bases in DNA homoduplexes of trinucleotide and hexanucleotide repeats. *Nucleic Acids Res.*, Revised manuscript submitted, 2017.

[73] A. Amadei, A. B.M. Linssen, and H. J.C. Berendsen. Essential Dynamics of Proteins. *PROTEINS: Structure, Function, and Genetics*, 17:412–425, 1993.

Chapter 6

E-motif formed by extrahelical cytosine bases in DNA homoduplexes of trinucleotide and hexanucleotide repeats

Feng Pan, Yuan Zhang, Viet Hoang Man, Christopher Roland and Celeste Sagui. (2017) E-motif formed by extrahelical cytosine bases in DNA homoduplexes of trinucleotide and hexanucleotide repeats. *Nucleic Acids Research*, Revised manuscript submitted.

ABSTRACT

Atypical DNA secondary structures play an important role in expandable trinucleotide repeat (TR) and hexanucleotide repeat (HR) diseases. The cytosine mismatches in C-rich homoduplexes and hairpin stems are weakly bonded; experiments show that for certain sequences these may flip out of the helix core, forming an unusual structure termed an "e-motif". We have performed molecular dynamics simulations of C-rich TR and HR DNA homoduplexes in order to characterize the conformations, stability and dynamics of formation of the e-motif, where the mismatched cytosines symmetrically flip out in the minor groove, pointing their base moieties towards the 5'-direction in each strand. TRs have two non-equivalent reading frames, $(GCC)_n$ and $(CCG)_n$; while HRs have three: $(CCCGGC)_n$, $(CGGCCC)_n$, $(CCCCGG)_n$. We define three types of pseudo basepair steps related to the mismatches and show that the e-motif is only stable in $(GCC)_n$ and $(CCCGGC)_n$ homoduplexes due to the favorable stacking of pseudo GpC steps (whose nature depends on whether TRs or HRs are involved) and the formation of hydrogen bonds between the mismatched cytosine at position i and the cytosine (TRs) or guanine (HRs) at position $i-2$ along the same strand. We also characterize the extended e-motif, where all mismatched cytosines are extruded, their extra-helical stacking additionally stabilizing the homoduplexes.

6.1 Introduction

Atypical DNA secondary structures have been identified as a common and causative factor for expansion in trinucleotide and hexanucleotide repeat sequences that underlie approximately 30 DNA expandable simple sequence repeat (SSR) diseases [1–3]. SSRs exhibit "dynamic mutations" that do not follow Mendelian inheritance: intergenerational expansion of SSRs is behind inherited neurodegenerative and neuromuscular disorders known as "anticipation diseases", where the age of the onset of the disease decreases and its severity increases with each successive generation [4–10]. After a certain threshold in the length of the repeated sequence, the probability of further expansion and the severity of the disease increase with the length of the repeat. The expansion is believed to be primarily caused by some sort of slippage during DNA replication, repair, recombination or transcription [2,3,7–15], that involves transient separation of complementary DNA strands

or exposure of a single DNA strand. This, in turn, can lead to the formation of hairpins and other secondary structures in the exposed strand. Cell toxicity and death have been linked to the atypical conformation and functional changes of the RNA transcripts, of DNA itself [3,16] and, when TRs are present in exons, of the translated proteins [3,17–26].

In particular, sequences of the form d(CGG)·d(CCG) are overexpressed in the exons of the human genome: CGG SSRs are found in the 5'-untranslated region (5'-UTR) of the fragile X mental retardation gene (FMR1) [27], while CCGs are found both in the 5'-UTR and translated regions of more than one gene. The normal range of the CGG SSRs in the population is 5-54, with the last ten repeats increasing the probability of disease in descendants [28,29]. SSRs of 55-200 CGGs constitute premutations associated with fragile X-associated tremor ataxia syndrome (FXTAS) in males [30] and premature ovarian failure in females [31]. SSRs longer than 200 CGG cause the inherited fragile X mental retardation syndrome [32]. CCG SSRs are related to three SSR diseases: the longest expansion occurs in the FRM2 gene giving rise to chromosome X-linked mental retardation (FRAXE) [33], and they also seem to play a role in Huntington's disease [34], and myotonic dystrophy type 1 [35]. More recently, it has been found that a d(GGGGCC)·d(GGCCCC) SSR in the first intron of the C9ORF72 gene leads to a hexanucleotide repeat expansion identified as the major cause behind frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) [36,37]. While the unaffected population carries fewer than 20 repeats (generally no more than a couple), large expansions greater than 70 repeats and usually encompassing 250-1600 repeats have been found in C9FTD and ALS patients.

In this work, we are interested in the C-rich repeat sequences. In order to understand the mechanisms underlying sequence expansion, gene hypermethylation, and folate-induced chromosomal fragile sites, it is crucial to elucidate the secondary structure adopted by the C-rich sequences $d(CCG)_n$ of various repeat length n. These sequences attracted considerable interest over 20 years ago, when it was found that the homoduplexes d(CCG)·d(CCG) (i.e., duplexes formed by the same CCG SSR strands) exhibited an unusual DNA secondary structure termed the "e-motif" [38,39]. This motif was seen in a solution NMR DNA antiparallel duplex where each strand consisted of two repeats, 5'-$(CCG)_2$-3' (PDB ID 1NOQ). In this helical duplex, the slipping of the strands leaves

the two 5'-C terminal unpaired, and a single central C·C mismatch surrounded by two Watson-Crick pairs. The central mismatch gave rise to the "e-motif", where the C bases in the mismatch symmetrically flip out in the minor groove, pointing their base moieties towards the 5' direction in each strand.

Remarkably, since the initial publication of the NMR DNA 5'-$(CCG)_2$-3'duplex results in 1995, there has been no other direct structural observation of the e-motif, reflecting the difficulty of experimental observation of flexible DNA duplexes, made probably more labile by the presence of the mismatches. However, there has been indirect observations that support the presence of e-motifs in DNA homoduplexes and hairpins of various lengths. The two most important results in this direction were obtained by chemical modification of the bases followed by subsequent cleavage [40, 41]. These studies also provided indirect evidence to the proposition that $d(GCC)_n$ homoduplexes or hairpin stems exhibit an *extended* e-motif formed by consecutive extrahelical C·C mismatches. Finally, notice that the GCC alignment in these duplexes is different from the CCG alignment in the NMR DNA duplex. However, since the short strands slip with respect to each other in the two-repeat duplex, the NMR structure also exhibits CpG steps between the Watson-Crick pairs.

In this work we present results from molecular dynamics simulations that provide a detailed structural and dynamical characterization of the e-motif. We first encountered an e-motif in our study of the hexanucleotide repeats behind ALS and C9FTD diseases [42]. Here, we extend this study and add the C-rich trinucleotide repeats in order to determine which sequences give rise to the e-motif, how stable they are, and what are the transistion mechanisms which transform the internal C·C mismatch to an e-motif.

6.2 Materials and Methods

Molecular Dynamics (MD). The sequences employed in this work are shown in Fig. 6.1. The simulations were carried out using the PMEMD module of the AMBER v.16 [43] software package with force fields ff99 BSC1 [44], ff99 BSC0 [45], and OL15 [46] used in different cases. A summary of the sequences and force fields used is presented in Table 1. All the C·C mismatches are initially chosen in the *anti-anti* conformation, which represents the minimum free energy conformation of the mismatches in the phase space mapped out by the torsion angle conformations.

The TIP3P model [47] was used for the water molecules, along with the standard parameters for ions in the AMBER force fields [48]. The long-range Coulomb interaction was evaluated by means of the Particle-Mesh Ewald (PME) method [49] with a 9 Å cutoff and an Ewald coefficient of 0.30768. Similarly, the van der Waals interaction were calculated by means of a 9 Å atom-based nonbonded list, with a continuous correction applied to the long-range part. The production runs for MD were generated using the leap-frog algorithm with a 2 fs timestep with Langevin dynamics with a collision frequency of 1 ps^{-1}. Conformations were saved every picosecond. The SHAKE algorithm was applied to all bonds involving hydrogen atoms. Regular MD was run for all sequences, for times that vary between 1 μs and 2 μs. For completion, we discuss results related to hexanucleotides repeats DC-1, DC-2 and SC-3, previously presented in Ref. [42], which are revisited in order to find the common denominator behind the e-motif formation. These simulations are extended in the present work in order to include DC-1-MUT, and DC-2$_{emotif}$ as specified in Table 1. DC-1-MUT is obtained from sequence DC-1 by mutating bases G5, G16 to C's and bases C9, C20 to G's. These changes enable one to probe the stability of e-motif structures in DC-1 HRs when the mismatched cytosines form hydrogen bonds between C's rather than G's. In addition, the DC-1-MUT and DC-2$_{emotif}$ structures were constructed with an initial e-motif as shown in Fig.1.

6.3 Results

The duplexes considered in this study are shown in Fig. 6.1, and the main features of each duplex simulation are listed in Table 1. An important issue when considering possible SSR conformations is the nature of the Watson-Crick pairs that surround the mismatches: sequences of the form 5'-(CCG)$_n$-3' and 5'-(GCC)$_n$-3' (without slipping such that strand ends are paired) exhibit Watson-Crick base pairs with GpC and CpG steps, respectively. In order to facilitate the discussion of our results, we have labeled the standard steps as L=GpC=GC/GC, M=CpG=CG/CG, and N=GG/CC; and defined three classes of pseudo steps, as listed in Table 2. With this notation, the step patterns, before and after e-motif formation, are given in Table 3. In the following we present our main results.

A. Spontaneous formation of e-motif in regular molecular dynamics. We carried out regular, unconstrained MD simulations for sequences GCC4 and CCG4 shown in Fig. 6.1(a), (b) with force fields BSC0 and BSC1 for 1 μs. Two initial conformations were

chosen: ideal A-DNA and ideal B-DNA. For each force field, the two initial conformations quickly converge. For the GCC4 sequence (but not for CCG4), spontaneous formation of an e-motif occurs in the BSC0 force field. With respect to the hexanucleotides, we previously presented results corresponding to two sets of 1 μs regular MD with the BSC0 force field for the sequences DC-1, DC-2, SC-3 , Fig. 6.1 (g,h,i). The DC-1 CCCGGC sequence showed spontaneous formation of the e-motif. These transitions are shown in Movie S1 for GCC4 (residues 10^{th} to 14^{th}, and complementary) where the e-motif forms at approximately 600 ns, and Movie S2 for CCCGGC in DC-1 (residues 4^{th} to 9^{th}, and complementary) where the e-motif forms around 300 ns and is stable for the remaining 700ns of the simulation.

B. Description and characterization of the e-motif. In the e-motif, the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the $i-2$ residue (i.e., towards the 5' direction in each strand). Figure 6.2 shows initial and late conformations for GCC4 and DC-1, which form an e-motif, and CCG4 and DC-2, which do not form an e-motif. Several quantities can be defined to clearly describe the transition from an intra-helical C·C mismatch to an e-motif. In Fig. 6.3 we show some of these quantities for the formation of an e-motif in GCC4 (compared to CCG4, which does not exhibit an e-motif). In the figure we can see clear transitions for the mismatch in GCC4, involving some intermediate transition states, between well defined initial and final average values. Shown are the partial handedness [50] of the C_{12}-C_{17} mismatch (from 0.5 to -0.5); the pseudodihedral angle (Ω_{12}) that describes the base unstacking of C_{12} with respect to the helical axis [51] (from $\sim 60^\circ$ to $\sim -100^\circ$); the center-of-mass distance (ep-distance) between basepairs adjacent to the mismatch, in this case basepairs 11-16 and 13-18 (from 7 Å to 4 Å); and the "e-motif distance" (ec-distance), defined as the distance between the N4 atom of C_{12} and the O2 atom of C_{10} in GCC4 or the N3 atom of G_{10} in CCG4 (from 7.5 Å to 4 Å). Since the spontaneous transition did not happen in the BSC1 and OL15 force fields in this time scale, the values of these quantities stay consistently near the initial values, as shown for instance in Fig. S1. In our work with the hexanucleotides [42] we showed that the backbone torsion angles α and γ (that normally display an anticorrelation such that their sum stays constant) behave such that $\alpha+\gamma$ corresponding to a mismatched base decreases approximately by 100° when the base flips into the minor groove. Thus, intrahelical C mismatches have $\alpha+\gamma \simeq 340^\circ$ while C mismatches flipped out into the minor groove have $\alpha+\gamma \simeq 240^\circ$,

as shown in Fig. 6.5 and Fig. 6.8. Hydrogen bonding for the e-motif is discussed below.

C. The e-motif is stable under the three force fields. Since the time scale for the spontaneous formation of the e-motif under the force fields BSC1 and OL15 can potentially be rather large, we decided to check the stability of an initial, built-in e-motif. Thus, we have built a $(GCC)_5$ duplex with an internal e-motif Fig. 6.1 (c) and checked for stability. Results for $GCC5_{emotif}$ under the three force fields are shown in Fig. S3 and Fig. S4 in the SI. Fig. S3 shows the RMSD of the middle 14 residues around the e-motif (R5-R11 and R20-R26), while Fig. S4 shows the RMSD of the bases participating in the pseudo GpC step (bases G7, C9, G22, and C24). Up to 1 μs, the e-motif is stable in the three force fields; the only change occurs in BSC0 where there are fluctuations in the e-motif in the interval 200-300 ns, after which the e-motif stabilizes again. One final observation: the e-motif in our simulations is stable for four and five trinucleotide repeats. The NMR study showed that two repeats can also form an e-motif, so we additionally ran 1 μs simulations of a single trinucleotide surrounded by one G-C Watson-Crick base pair at the end. For the single trinucleotide, the e-motif unraveled in our simulations.

D. A single e-motif is partially stabilized by the formation of hydrogen bonds between the C bases (i residue) in a mismatch and the $i-2$ bases: these are C bases in the case of (GCC), and G bases in the case of (CCCGGC) in DC-1. A hydrogen bond analysis for the e-motif in $GCC5_{emotif}$ and $CCG5_{emotif}$ is shown in Fig. 6.4. The most important hydrogen bonds stabilizing the extruded bases in $GCC5_{emotif}$ are C8(N4)-C6(O2) and its equivalent C23(N4)-C21(O2) for both BSC0 and OL15; and C8(N4)-C26(O3') and its equivalent C23(N4)-C11(O3') for BSC1, which –unlike the previous bond– represents hydrogen bonding across strands. We notice that in the experimental NMR duplex (PDB ID 1NOQ) the hydrogen bonds are of the type C8(N4)-C6(O2) and its equivalent C23(N4)-C21(O2), validating the results for BSC0 and OL15. For $CCG5_{emotif}$ the only hydrogen bonds that have any measurable presence are C8(N4)-G6(N3) and its equivalent on the other strand, C23(N4)-G21(N3). Now we turn to the hexanucleotides. Figure 6.7 shows the time evolution of the number of hydrogen bonds for the hexanucleotides. The top panel shows the data for DC-1, where the initial mismatches are all intra-helical. Before 100 ns, the mismatch C7-C18 is still intrahelical. After the flipping out of the mismatched bases C7 and C18 into the minor groove is completed, C7 forms hydrogen bonds with G5 and C18 does so with G16. Of these bonds, the most important are C7(N4)-G5(N3) and its equivalent on the other strand,

C18(N4)-G16(N3).

E. The e-motif occurs in paired-end homoduplexes of (GCC) and (CCCGGC) SSRs, but not in the other reading frames. First, we discuss the trinucleotide repeat results. For none of the three force fields did the CCG4 duplexes spontaneously formed an e-motif. To further check the stability of the e-motif in CCG sequences, we built a $(CCG)_5$ duplex with an internal e-motif Fig. 6.1 (d) and probed its stability. Results for the $CCG5_{emotif}$ under the three force fields are shown in Fig. S5 and Fig. S6 in the SI. Figure S5 shows the RMSD of the middle 14 residues around the e-motif (R5-R11 and R20-R26), while Fig. S6 shows the RMSD of the bases participating in the pseudo CpG step (bases C7, G9, C22, and G24). The e-motif quickly unravels and the mismatched C bases become intrahelical for both BSC1 and OL15. This transition is shown in Movie S3 (residues 6^{th} to 10^{th}, and complementary) for force field BSC1. The transition for these two force fields can be identified by the sum of the backbone torsion angles $\alpha + \gamma$, as shown in Fig. 6.5 which goes from $\simeq 240°$ to $\simeq 340°$ as the mismatched bases become intrahelical. For BSC0 the mismatched bases continue being extrahelical in the 1 μs of the simulation, but their characteristic hydrogen bond pattern decays with time as shown in Fig. 6.6.

Now we turn to the hexanucleotides. In previous work [42], we showed that the sequence in DC-1 led to the formation of an e-motif, that forms around 300 ns and is stable for the remaining 700 ns of the simulation. By contrast, sequences in DC-2 and SC-3 did not form an e-motif. In DC-2, the bases of a mismatch alternate between the minor and major grooves, while SC-3 is unstable and either unfolds or converts to the more stable DC-1 duplex. To dispel the possibility that DC-2 may not have formed an e-motif because the simulation was not long enough, we have extended this work by choosing as initial conformation for DC-2 one with an e-motif. A movie showing the time evolution of this duplex is shown in Movie S4. The initial e-motif unravels at different times for each of the force fields (Table 1), lasting longer for BSC0, where it is stable for about 350 ns. However after that, the bases turn back into the helix. They also push the bases in the flanking mismatch into the major and minor grooves occasionally, but none of the mismatches formed an e-motif again. In fact, the dynamical configurations are the same as those observed for DC-2 in our previous work. Finally, the time evolution of the mutated case in DC-1-MUT (which does not belong to any SSR) is shown in Movie S5. The initial e-motif also unravels at different times for each of the force fields (Table

1), lasting longer for BSC0, where it is stable for about 250 ns, but then the bases turn towards the inner helix, with the base C7 flipping in and out of the helix, and affecting with its motion the base C6 in the flanking mismatch. The unraveling of the e-motif is quantified in the middle and bottom panels of Fig. 6.7 and in Fig. 6.8, and in Fig. 6.10. In Fig. 6.7, duplex DC-1 starts with no e-motif but forms one at mismatch C7-C18 at about 300 ns. Both DC-1-MUT and DC-2_{emotif} start with an e-motif at C7-C18 and hydrogen bonds between mismatched bases at position i and those at $i-2$, i.e., C7-C5 and C18-C16, which disappear as the system evolves. Figure 6.8 shows the $\alpha+\gamma$ jump of 100° as the e-motif is formed in DC-1 (negative jump), and as the initial e-motif disappears in DC-1-MUT and DC-2_{emotif} (positive jump).

F. Creation of the e-motif is favored by the formation of pseudo GpC steps when the bases in the C·C mismatches are extruded. Figure 6.9 shows the G-G stacking that occurs in a pseudo GpC step after the C mismatches have been extruded (L_L in our notation) in a GCC sequence, whose consequence is a better overall stacking of the helix. This fact explains why in homoduplexes with paired ends, the e-motif occurs in GCC sequences (formation of L_L steps after extrusion) but not in CCG sequences (formation of M_M steps after extrusion). Fig. S7 shows the overlap areas of the basepair ring atoms of the pseudo GpC step for for GCC5_{emotif}; the distribution functions have a peak at around 2.6 Å^2. For the hexanucleotide repeats, the step nature becomes slightly more complicated. In DC-1, the extrusion of the bases in an e-motif leads to a pseudo GpC (L_{LC}) step, such that the new step pattern around the intra-helical mismatch (the one that was not extruded) is N-L_C-L_{LC}-N. By contrast, flipping of the bases of the C7-C18 mismatch in DC-2 results in a pseudo CpG (M_{MC}) step, such that the new step pattern around the intra-helical mismatch is N-M_C-M_{MC}-N, and therefore the e-motif is not favored. Figure S8 shows the overlapping for steps L_{LC} and M_{MC} for DC-1 and DC-2 respectively, for an almost perfect e-motif in C7-C18. While there is good overlap in L_{LC} in DC-1, there is almost no overlap for the M_{MC} step in DC-2. This trend is reinforced by the previous steps (not shown): L_C in DC-1 has good overlap, but M_C in DC-2 does not. Finally, in the mutated case DC-1-MUT, flipping of the C7-C18 bases leads to a step pattern L-M_C-L_{LC}-M, which cannot completely stabilize the e-motif. Figure 6.10 shows the overlap areas of the basepair ring atoms of the pseudo GpC steps (L_{LC}) of DC-1 and DC-1-MUT, as well as the overlap areas of the pseudo CpG steps (M_{MC}) of DC-2_{emotif} as a function of time for the 1 μs run.

G. The extended e-motif is stabilized by highly cooperative interactions. Finally, we have considered the stability and structural characteristics of the extended e-motif, when all the C·C mismatches are extruded in e-motifs. In order to characterize this motif, we have considered the duplexes GCC4$_{extended}$ and CCG4$_{extended}$ shown in Fig. 6.1 (e) and (f), with four consecutive e-motifs. In addition to the favorable stacking afforded by pseudo GpC steps, the extended e-motif is further stabilized by the stacking of the extruded C bases themselves (see Figs. 6.13 and 6.14). Figure 6.11 shows the RMSD of the central section (residues 4-12,18-26) of the duplexes GCC4$_{extended}$ and CCG4$_{extended}$ with respect to the initial frame in a 2 μs MD simulation for the three force fields. These figures suggest that GCC4$_{extended}$ is stable in the 2 μs time scale, while CCG4$_{extended}$ start to deviate from the initial structure at late times: the duplex in BSC1 shows a considerable increase of RMSD at approximately 1.6 μs, the OL15 RMSD shows a smaller jump at approximately the same time, while the average RMSDs for the BSC0 duplexes increases slowly and monotonically from an "early" value of 1.4 Å at 10 ns to a value of 2.7 Å at 2 μs. Figure S9 shows the time behavior of the first principal component of internal nucleotides (nucleotides 4-12, 18-26) as obtained from a principal component analysis (PCA). GCC4$_{extended}$ is characterized by regular fluctuations in the three force fields, while CCG4$_{extended}$ displays a breaking of symmetry around the zero eigenvalue in both BSC1 and OL15, signaling a conformational transition. Figure 6.12 shows the hydrogen bonds with highest percentage in GCC4$_{extended}$ associated with the extruded C6, C9, C12 bases and the symmetric ones on the other strand. Notice that while OL15 displays consistently intra-strand C$_i$(N4)-C$_{(i-2)}$($O2$) bonding, BSC0 shows two of these (for C6 and C12) and one inter-strand bonding [C9(N4)-C24(O3')], and BSC1 shows three inter-strand bondings C$_i$(N4)-C$_{(33-i)}$(O4'). Finally, Figs. 6.13 and 6.14 show the stacking of the extruded C bases themselves for the BSC1 and OL15 force fields, respectively. The inter-strand hydrogen bonding in BSC1 leads to better C-C stacking than the inter-strand bonding in OL15.

6.4 Discussion

In this work, we have presented results from MD simulations that provide a detailed structural and dynamical characterization of the e-motif, along with the factors that stabilize it. The initial duplex with an e-motif as revealed by an NMR study by Gao *et al.*

in 1995 [38] supplied the first evidence that the C·C mismatch pairs were flexible enough to produce a significant conformational change within a DNA double helix. After that, two important studies [40,41] provided indirect evidence of the presence of e-motifs. These studies employed chemical modification of the bases followed by subsequent cleavage. The modifications involved guanine and cytosine chemical modifications in DNA hairpins with DMS and hydroxylamine respectively [40], and by mechlorethamine crosslinking reaction in DNA homoduplexes of the form d(GCC)$_n$·d(GCC)$_n$ [41]. Both studies found that the helical part (standard duplex or hairpin stem) contains CpG steps between the Watson-Crick pairs and that the C·C mismatches are extrahelical. It is important to note that although the mechanisms of chemical modification probably occurred after the extrusion of the mismatched C bases, one cannot ultimately exclude the possibility that the chemical process itself could induce the extrahelical cytosine conformations. Given that the C·C mismatch is the least stable mismatch pair, these can easily become unstacked from the core helix, flipping outside the helix depending on their local environment. These studies also provided indirect evidence to the proposition that d(GCC)$_n$ homoduplexes or hairpin stems exhibit an *extended* e-motif formed by consecutive extrahelical C·C mismatches, something that could not be achieved in the short sequence employed by the NMR study. Based on their indirect evidence, Yu *et al.* [40] proposed a schematic of an extended e-motif that is in remarkable agreement with the atomic structures presented in this work.

An important issue when considering possible SSR conformations is the nature of the Watson-Crick pairs that surround the mismatches: sequences of the form 5'-(CCG)$_n$-3' and 5'-(GCC)$_n$-3' (without slipping such that strand ends are paired) exhibit Watson-Crick base pairs with GpC and CpG steps, respectively. The slipping of strands with respect to each other in the (CCG) DNA NMR structure 5'-(CCG)$_2$-3' (PDB ID 1NOQ) [38] results in CpG steps (as opposed to the GpC steps that would result if the DNA strands were paired at the ends). The importance of the steps has been pointed out before. In the scheme introduced by Darlow and Leach [52,53], hairpins were classified according to the alignment of the sides of the hairpins and the presence of an odd or even number of unpaired bases in the loop: "frame 1" corresponds to GpC steps between the Watson-Crick basepairs in the hairpin stem, while "frame 2" corresponds to CpG steps between the Watson-Crick basepairs in the stem (a "frame 3" corresponds to alignment CGC that lacks Watson-Crick basepairs, and therefore corresponds to a considerably less stable structure).

For the hexanucleotides, there are three possible alignments for the C-rich sequences: DC-1 and DC-2 combine neighboring Double C·C mismatches separated by four Watson-Crick basepairs; DC-1 combines CC/GG and CpG steps; DC-2 combines GG/CC and GpC steps. The third alignment is SC-3, that combines Single C·C mismatches separated by two Watson-Crick basepairs. This duplex only contains CC/GG steps. In order to facilitate the discussion, we introduced a notation for pseudo steps in Tables 2 and 3. In the following we discuss our main results.

A. The e-motif is stable under the three force fields. In a related context, we have calculated the various free energy maps for the mismatch conformations in CCG, GCC, GGC and CGG trinucleotide repeats, and show that the force fields BSC0, BSC1 and OL15 all share the same minima for the mismatch conformations. The main difference is that barriers between these minima are lowest for BSC0 and largest for BSC1, with OL15 providing for intermediate barriers. Thus, transitions between mismatch conformations corresponding to different global and relative minima statistically will happen faster in BSC0 than in BSC1. Indeed, we see spontaneous formation of e-motifs during regular MD in a few hundreds of nanoseconds both in trinucleotide repeats (GCC4) and hexanucleotide repeats (DC-1) under the BSC0 force field. The slower transitions in BSC1 could easily put the e-motif formation completely out of range for the current computer capabilities available. Instead, we built a $(GCC)_5$ duplex with an internal e-motif Fig. 6.1 (c) and checked its stability. Up to 1 μs, the e-motif is stable in the three force fields. This is of course no proof that the e-motif is a minimum in the free energy map, but strongly supports this when considered with the rest of the results and the experimental data. Incidentally, our study for the single e-motif also seems to indicate that OL15 is perhaps most suited for the description of mismatches. Both BSC1 and OL15 were created as an attempt to correct deficiencies in BSC0. The e-motif is formed readily under the BSC0 force field. BSC1 has relatively high free energy barriers between relative minima that correspond to labile mismatch conformations observed experimentally (i.e., BSC1 seems to be too rigid for mismatches) and it does not reproduce the hydrogen bond pattern of the NMR structure 1NOQ. OL15, on the other hand, behaves properly as far as formation of the e-motif and hydrogen bond patterns (compared to the only experimental structure that provides atomic detail). However, for the extended e-motif, for which there is no experimental structural data, the stacking of the extruded C bases is optimized for BSC1 (across strands).

B. Description and characterization of the e-motif. In the e-motif, the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the $i-2$ residue (i.e., towards the 5' direction in each strand). This is seen in Fig. 6.2 for GCC4 and DC-1. The transition from an intrahelical C·C mismatch to an e-motif can be described quite distinctly by several quantities that involve some intermediate transition states between well defined initial and final average values that are very different, as described in the Results section. These quantities (shown for trinucleotides in Figs. 6.3, 6.5 and 6.8) include (i) partial handedness of the e-motif; (ii) pseudodihedral angles describing the mismatched base unstacking with respect to the helical axis; (iii) the center-of-mass distance (ep-distance) between the basepairs surrounding the mismatch; (iv) the "e-motif distance" (ec-distance), defined as the distance between the N4 atom of a mismatched C base at position i and the O2 atom of the C base at position $i-2$ in GCC4 or the N3 atom of G base at position $i-2$ in CCG4; and (v) the sum of backbone torsion angles $\alpha+\gamma$ that decreases approximately by 100° when the mismatched base flips into the minor groove.

C. Creation of the e-motif is favored by the formation of pseudo GpC steps when the bases in the C·C mismatches are extruded. Consequently, the e-motif is stable in paired-end homoduplexes of (GCC) and (CCCGGC) SSRs, but not in the other reading frames. In trinucleotide repeats, the extrusion of the C mismatches results in a pseudo GpC step L_L in a pair-ended GCC sequence, which leads to G-G stacking in the adjacent basepairs (Fig. 6.9) and a better overall stacking of the helix. This fact explains why in homoduplexes with paired ends, the e-motif is favored in GCC sequences (formation of L_L pseudo steps after extrusion) but not in CCG sequences (formation of M_M pseudo steps after extrusion, see Table 3). Indeed, the simulations presented here show spontaneous formation of e-motif in GCC4 but not in CCG4. They also show stability in the three force fields of the initial e-motif in $GCC5_{emotif}$ (Figs. 6.4, and S2 and S3 in the SI), but not for $CCG5_{emotif}$ (Figs. 6.5, and S4 and S5 in the SI). In the latter case, the e-motif quickly unravels and the mismatched C bases become intrahelical for both BSC1 and OL15. For BSC0 the mismatched bases continue being extrahelical in the 1 μs of the simulation, but their characteristic hydrogen bond pattern decays with time as shown in Fig. 6.6.

For the hexanucleotide repeats, this argument still holds, even though the presence of two mismatches makes the helix less stable and introduces some additional nuances.

First, we notice that due to the symmetry of the DC-1 and DC-2 sequences, the bases of either of the two mismatches, C7-C18 or C6-C19, can be extruded to form equivalent e-motifs. In DC-1, the extrusion of the bases in an e-motif leads to a pseudo GpC (L_{LC}) step, such that the new step pattern around the intra-helical mismatch (the one that was not extruded) is N-L_C-L_{LC}-N. By contrast, flipping of the bases of the C7-C18 mismatch in DC-2 results in a pseudo CpG (M_{MC}) step, such that the new step pattern around the intra-helical mismatch is N-M_C-M_{MC}-N, and therefore the e-motif is not favored. Finally, in the mutated case DC-1-MUT, flipping of the C7-C18 bases leads to a step pattern L-M_C-L_{LC}-M, which cannot completely stabilize the e-motif. The overlap areas of the pseudo steps (Figs. S8 and 6.10) reflect the stability of the helical stacking as DC-1 forms an e-motif and DC-1-MUT and DC-2$_{emotif}$ lose their initial e-motif. Notice that the extruded bases in DC-1 at position i form hydrogen bonds C(N4)-G(N3) with the G bases at position $i-2$. Instead, both DC-1-MUT and DC-2 form C(N4)-C(O2) hydrogen bonds with C bases at position $i-2$. Given that the O2-H-N4 hydrogen bonds are in principle stronger than the N3-H-N4 hydrogen bonds, but that C(N4)-C(O2) cannot stabilize either DC-1-MUT or DC-2$_{emotif}$, it is clear that the favorable stacking of the overall helix afforded by the GpC pseudo steps is the predominant factor for the formation of e-motifs.

D. The single e-motif is partially stabilized by the formation of hydrogen bonds between the C bases (i residue) in a mismatch and the $i-2$ bases: these are C bases in the case of (GCC), and G bases in the case of (CCCGGC) in DC-1. The most important hydrogen bonds stabilizing the extruded bases in GCC sequences are $C_{i,mismatch}$(N4)–$C_{(i-2),WatsonCrick}$(O2) for both BSC0 and OL15, as well as for the experimental NMR duplex in 1NOQ (Fig. 6.4). On the other hand, BSC1 finds an inter-strand hydrogen bond between the N4 atom of the mismatched C base in one strand, and the O3' atom of the next mismatched C in the opposite strand. For the (CCCGGC) hexanucleotide the most important hydrogen bond is C_i(N4)–$G_{(i-2)}$(N3) (Fig. 6.7).

E. The mismatched C bases in an e-motif are always in the minor groove. This property is directly linked to point (C) above: from Fig. 6.1 it is clear that a GpC pseudo step L_L in a GCC trinucleotide repeat means that the extruded C basis is preceded (in the 5' direction) by a G basis, while in a CpG pseudo step M_M in a CGG trinucleotide repeat it is followed (in the 3' direction) by a G basis. The step arrangements have immediate consequences on the rotation paths followed by the extruded bases. A mismatched C

preceded by a G in the 5' direction favors a rotation path towards the minor groove such that the sum of backbone torsion angles $\alpha + \gamma$ decreases approximately by 100° when the base flips into the minor groove. On the other hand, a mismatched C followed by a G in the 3' direction favors a rotation path towards the major groove such that the difference of backbone torsion angles $\epsilon - \zeta$ increases approximately by 290° when the basis flips into the major groove, as we have shown in our previous work [42]. Once the base has flipped into the minor groove, it finds it easier to form hydrogen bonds with bases in the 5' direction due to the narrower space. Instead bases extruded into the major groove find themselves in a wider space and flip back and forth, unable to stably anchor themselves to another base. The same argument applies to the hexanucleotides, except that in this case it is the double mismatches that must be preceded by a G in the 5' direction in order for one of them to flip into the minor groove (as stated before, both mismatches are completely equivalent). Thus DC-1 favors e-motifs but DC-2 does not. SC-3 is a special case as it is less stable than the other two. Topologically it is also more different as single mismatches are intercalated every two Watson-Crick basepairs. In our previous work, we found that in 1μs run, the duplex unraveled and in another 1μs run, one strand slipped and the duplex adopted a DC-1 conformation. However longer repeats might be more stable, in which case mismatches like C7-C18 would favor e-motif which may help to stabilize the helix (but not C4-C22 or C10-C16).

F. The extended e-motif is stabilized by highly cooperative interactions. In addition to the favorable stacking afforded by pseudo GpC steps, either L_L in trinucleotides or L_C-L_{LC} in hexanucleotides, and the hydrogen bonds between the mismatched bases and other nucleotides, the extended e-motif is further stabilized by the stacking of the extruded C bases themselves (Figs. 6.13 and 6.14). The net result is a very stable anomalous secondary structure. Our simulations suggest that $GCC4_{extended}$ is stable in the 2 μs time scale, while $CCG4_{extended}$ start to deviate from the initial structure at late times. The pattern of stabilizing hydrogen bonds for $GCC4_{extended}$ depend on the force field: OL15 displays consistently intra-strand C_i(N4)-$C_{(i-2)}(O2)$ bonding, BSC0 shows two of these (for C6 and C12) and one inter-strand bonding [C9(N4)-C24(O3')], and BSC1 shows three inter-strand bondings between the N4 atom of the C_i mismatched base in one strand and the O4' atom of the next Watson-Crick paired C in the opposite strand. It is clear that the additional cooperativity provided by the C-stacking enormously extends the time scale to probe the stability of the extended e-motif. The results presented

here are only indicative that the e-motifs in $GCC4_{extended}$ are stable and that those in $CCG4_{extended}$ may eventually unravel and become intra-helical mismatches. Finally, there is the question of whether cytosines may be protonated and how that might affect our results. Experimentally, C protonation seems to depend (not surprisingly) on the environment. Of the two studies that proposed an extended e-motif, one [40] reported that some C mismatches are protonated, but the other [41] did not, mainly the N3 are used for crosslinking with mechlorethamine (see Fig. 2B in Ref. [41]). Moreover, the NMR structure by Gao et al. does not contain protonated cytosines (but the sequence is very short with only one e-motif). Our simulations were carried out with unprotonated Cs, but we make the following observations. If the Cs were initially protonated (before they form e-motif) they would tend to stabilize intrahelical mismatches by an additional hydrogen bond, as has been reported, for instance, in parallel DNA helices, and they would not favor the formation of the e-motif. Once the C bases are extruded, they can probably protonate. However, that would not make an important difference in the results presented here: the main driving force behind the formation of the e-motif is the stacking provided by the pseudo GpC steps of various forms, because it stabilizes the helical duplex, both for single an extended e-motif. In addition, observation of the extended e-motif indicates that due to spatial constraints, the stacking of extruded C bases can only take two forms: either intra-strand as obtained with OL15, or inter-strand as obtained with BSC1 (of course, along the duplex there could be an assortment of these, as shown by BSC0). Thus protonated Cs will not find a "third" form of extruded-cytosine stacking, although they might favor one form versus the other.

G. Biological implications. As discussed in the introduction, the first step in the expansion of SSR is the formation of atypical secondary structures in single-stranded DNA. To date, there is no complete understanding of how exactly this happens or why there is a critical threshold length that makes the repeating tract unstable and triggers the onset of pathology. The initial intuitive explanation [3, 14, 54–58] was that it is the minimal length at which the DNA atypical secondary structure becomes stable. However, as discussed by Lee and McMurray [59], small-sized loops can be stable. Instead, in single strand breaks or on Okazaki fragments during replication, there is a competition between duplex reconstitution (no mutation) and secondary structure formation in the single strand (leading to a premutation). If the gap filling synthesis cannot prevent or runs past a relatively stable self-pairing in the strand, then the excess bases initiate

folding into secondary structure in the SSR strand. The details of this atypical DNA secondary structure are important for a complete understanding of sequence expansion; gene hypermethylation; interactions with proteins involved in transcription coupled repair (TCR) nucleotide excision repair (NER), flap endonuclease 1 (FEN1), DNA mismatch repair (MMR, especially MutSβ, whose abnormal binding to SSR hairpins has been linked to SSR expansion), and others.

Table 6.1 Summary of molecular dynamics simulation results for the different DNA helical duplexes considered.

Label	Sequence	Initial E-motif	Force Field	Time (ns)	E-motif Status
GCC4	C-$(GCC)_4$-G	no	BSC0, BSC1	1000	e-motif formation at 600 ns in BSC0
CCG4	G-$(CCG)_4$-C	no	BSC0, BSC1	1000	no e-motif transition
DC-1	$(CCCGGC)_2$	no	BSC0, BSC1, OL15	1000	e-motif formation at 300 ns in BSC0
DC-2	$(CGGCCC)_2$	no	BSC0, BSC1, OL15	1000	no e-motif transition
SC-3	$(CCCCGG)_2$, slipped	no	BSC0	1000	no e-motif transition
$GCC5_{emotif}$	$(GCC)_5$	yes	BSC0, BSC1, OL15	1000	stable
$CCG5_{emotif}$	$(CCG)_5$	yes	BSC0, BSC1, OL15	1000	mismatches become intra-helical for BSC1 & OL15; e-motif in BSC0 loses H-bonds
$GCC4_{extended}$	C-$(GCC)_4$-G	yes, extended e-motif	BSC0, BSC1, OL15	2000	stable extended e-motif for all three force fields
$CCG4_{extended}$	C-$(GCC)_4$-G	yes, extended e-motif	BSC0, BSC1, OL15	2000	unstable e-motif for BSC1 and OL15; RMSD around e-motif increases for BSC0
DC-1-MUT	5'(CCCGCCCCGGGC)3' 3'(CGGCGCCGCCCC)5'	yes	BSC0, BSC1, OL15	1200	e-motif lost at 250 ns in BSC0, at 170 ns in BSC1, at 35 ns in OL15
$DC\text{-}2_{emotif}$	$(CGGCCC)_2$	yes	BSC0, BSC1, OL15	1200	e-motif lost at 350 ns in BSC0, at 60 ns in BSC1, at 32 ns in OL15

Table 6.2 Steps and pseudo steps exhibited by the homoduplexes with mismatches.

Nomenclature	Description
L=GpC=GC/GC M=CpG=CG/CG N=GG/CC=CC/GG	standard basepair step
L_C=GC/CC=CC/GC M_C=CG/CC=CC/CG W=CC/CC	pseudo GpC step L containing intrahelical C mismatches pseudo CpG step M containing intrahelical C mismatches pseudo step containing two intrahelical C mismatches
L_L=GC//GC M_M=CG//CG	pseudo steps where the two basepairs of the step are simply stacked on top of each other but not covalently linked along the backbone (because Cs have been extruded)
L_{LC}=GC//CC=CC//GC M_{MC}=CG//CC=CC//CG	pseudo steps like L_C and M_C, but the G-C or C-G basepairs are not covalently linked to the C intrahelical mismatches (because one of the two C·C mismatches has been extruded)

Table 6.3 Step changes for different DNA homoduplexes before and after e-motif formation.

Label	Steps without e-motif	Steps after e-motif formation
GCC4	M-L_C-L_C-M-L_C-L_C-M-...	M-L_L-M-L_L-M...
CCG4	L-M_C-M_C-L-M_C-M_C-L-...	L-M_M-L-M_M-L...
DC-1	L_C-N-M-N-L_C-W-L_C-N-M-N-L_C	L_C-N-M-N-L_C-L_{LC}-N-M-N-L_C
DC-2	M_C-N-L-N-M_C-W-M_C-N-L-N-M_C	M_C-N-L-N-M_C-M_{MC}-N-L-N-M_C
DC-1-MUT	L_C-N-M-L-M_C-W-L_C-M-N-N-L_C	L_C-N-M-L-M_C-L_{LC}-M-N-N-L_C

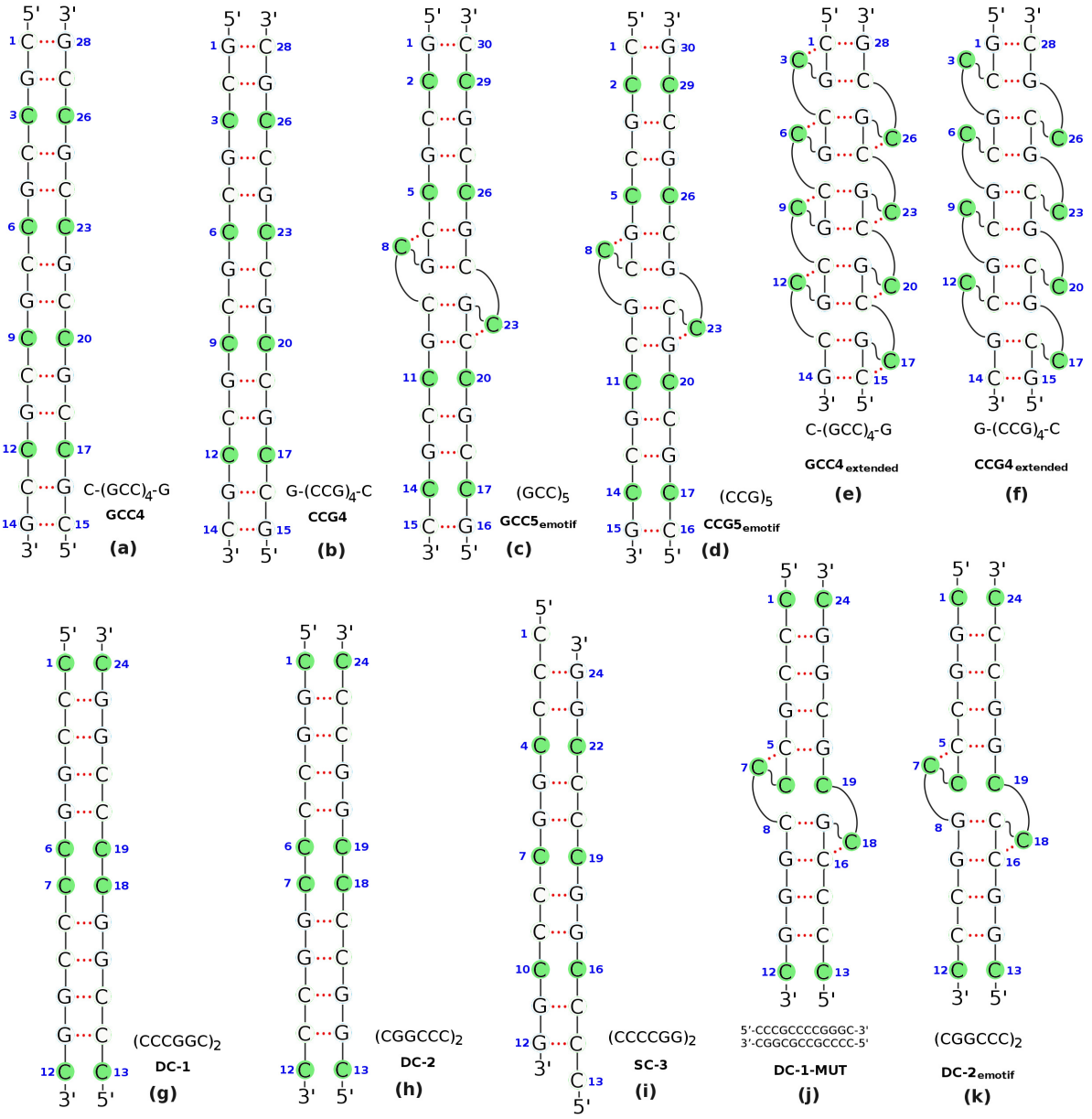


Figure 6.1 Schematics of the initial DNA helical duplexes considered in this study. The C mismatched bases are marked by solid green circles. Nucleotide indexes are labeled by blue numbers. Hydrogen bonds are indicated by dashed red lines. More details about the duplexes and the corresponding simulation results are provided in Table 1.

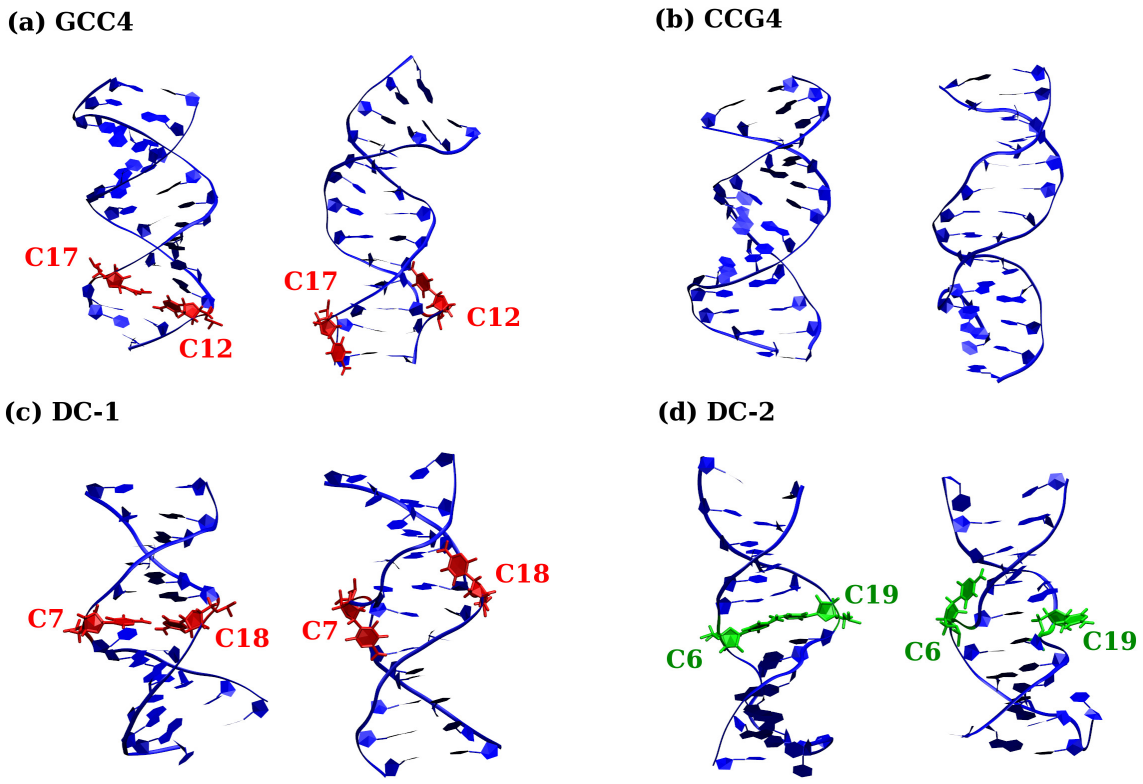


Figure 6.2 Initial (left) and final (right) structures for (a) GCC4, (b) CCG4, (c) DC-1, and (d) DC-2 structures obtained from the molecular dynamics simulations. The bases in the C·C mismatches that form an e-motif are shown in red. Those flipped out of the inner helix but not forming an e-motif are shown in green.

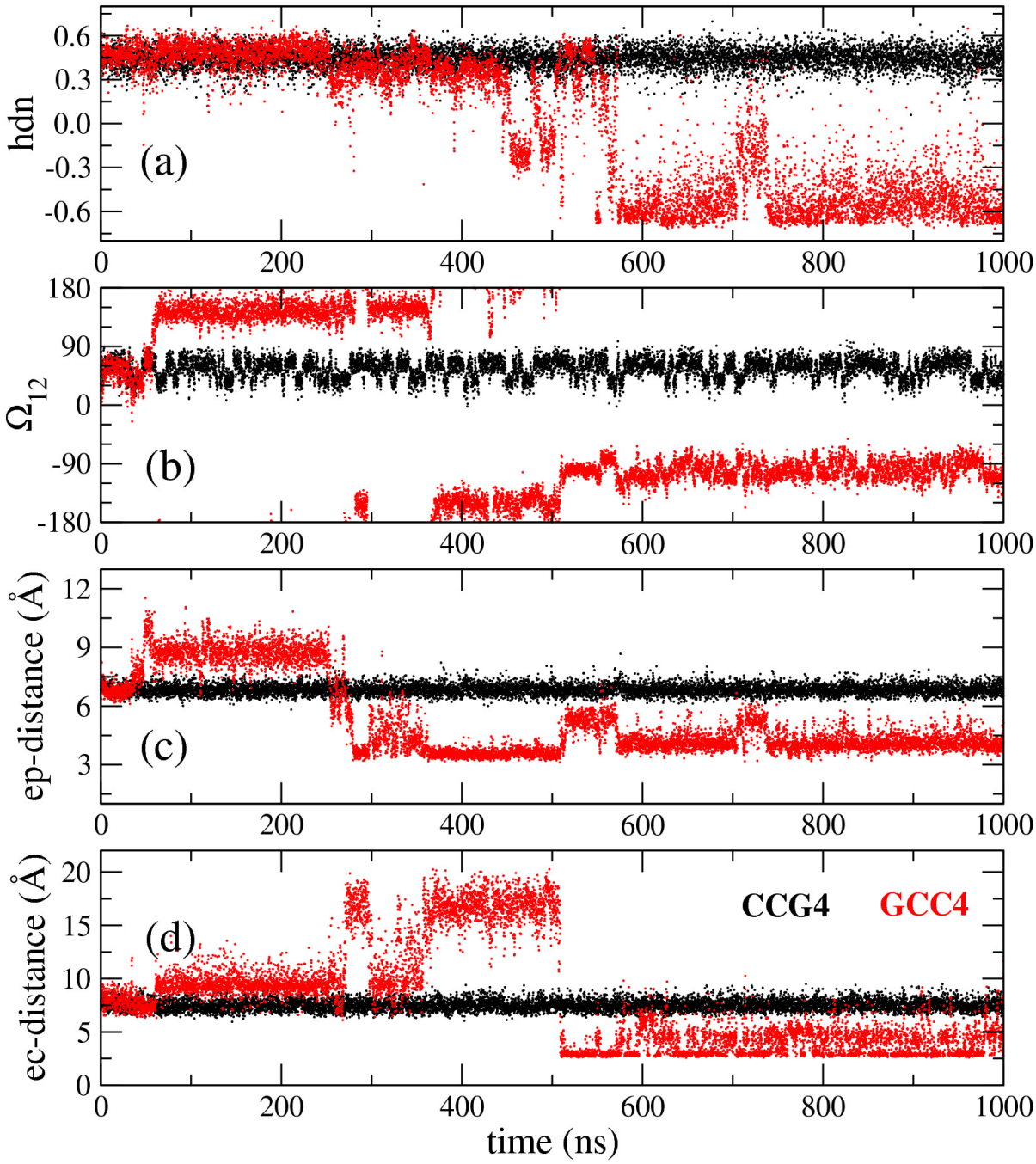


Figure 6.3 Time dependence of quantities characterizing the transition to an e-motif. Results for duplexes CCG4 and GCC4 are shown in black and red, respectively. (a) Partial handedness of the C_{12}-C_{17} mismatch; (b) pseudodihedral angle (Ω_{12}) describing the base unstacking of C_{12} with respect to the helical axis; (c) center-of-mass distance (ep-distance) between basepairs 11-16 and 13-18; (d) "e-motif distance" (ec-distance), defined as the distance between the N4 atom of C_{12} and the O2 atom of C_{10} in GCC4 or the N3 atom of G_{10} in CCG4.

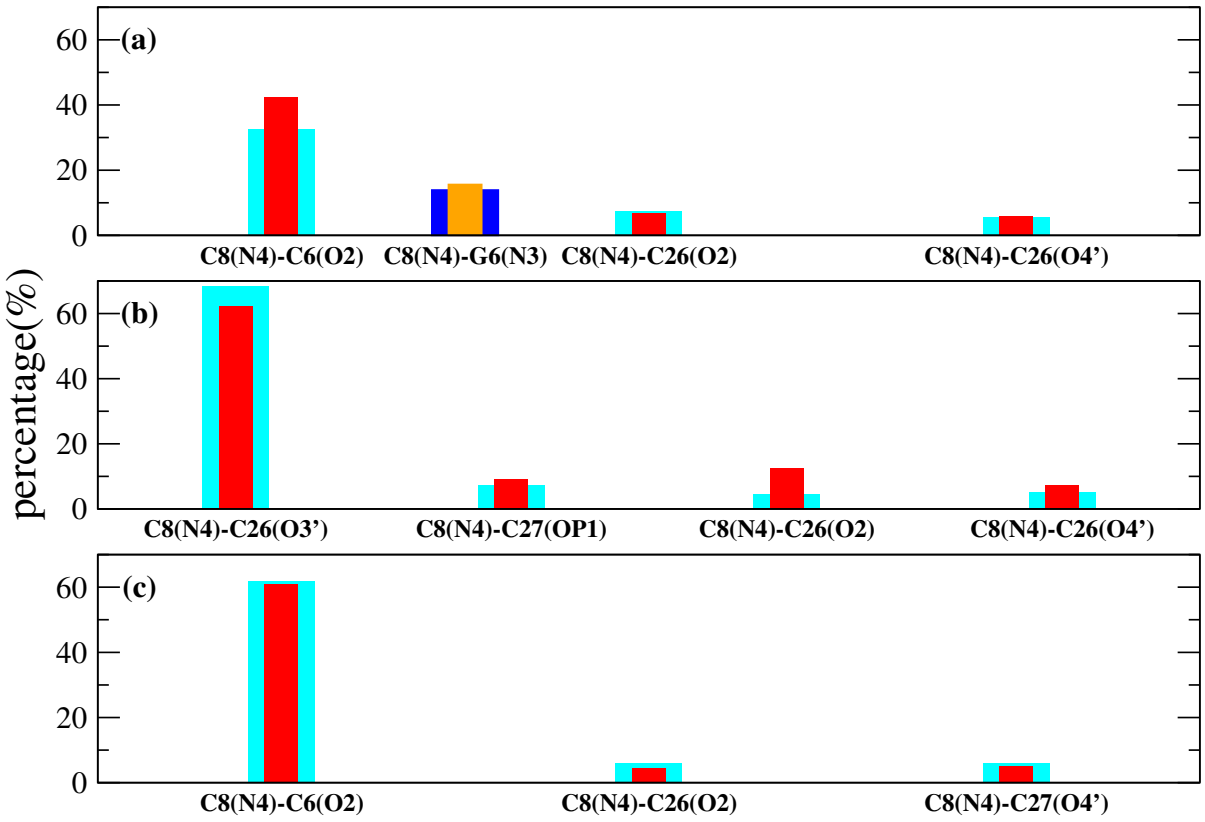


Figure 6.4 Hydrogen bond population for 1 μs simulation of GCC5$_{emotif}$ as obtained form the three force fields: (a) BSC0; (b) BSC1; (c) OL15. The x-axis indicates the hydrogen bond, while the y-axis gives its percentage over the duration of the simulation. Cyan color shows the percentage of the hydrogen bond on one strand and the red color shows the symmetric bond on the other strand. Blue and orange bars show hydrogen bonds for CCG5$_{emotif}$.

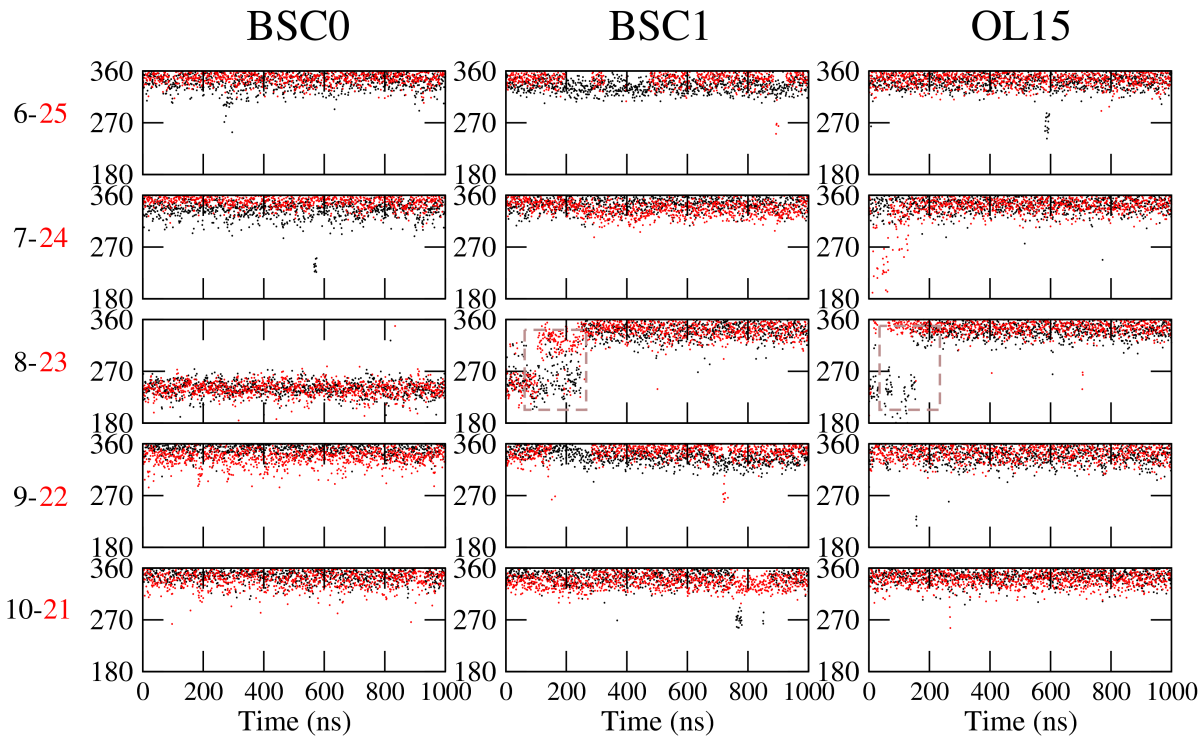


Figure 6.5 Backbone torsion angles $(\alpha + \gamma)$ for trinucleotide repeats. Sum of torsion angles $(\alpha + \gamma)$ for bases 6–10 (black) and 21–25 (red) as a function of time for CCG5$_{emotif}$ as obtained from the different force fields. The rectangle with dashed lines indicates the transition in BSC1 and OL15.

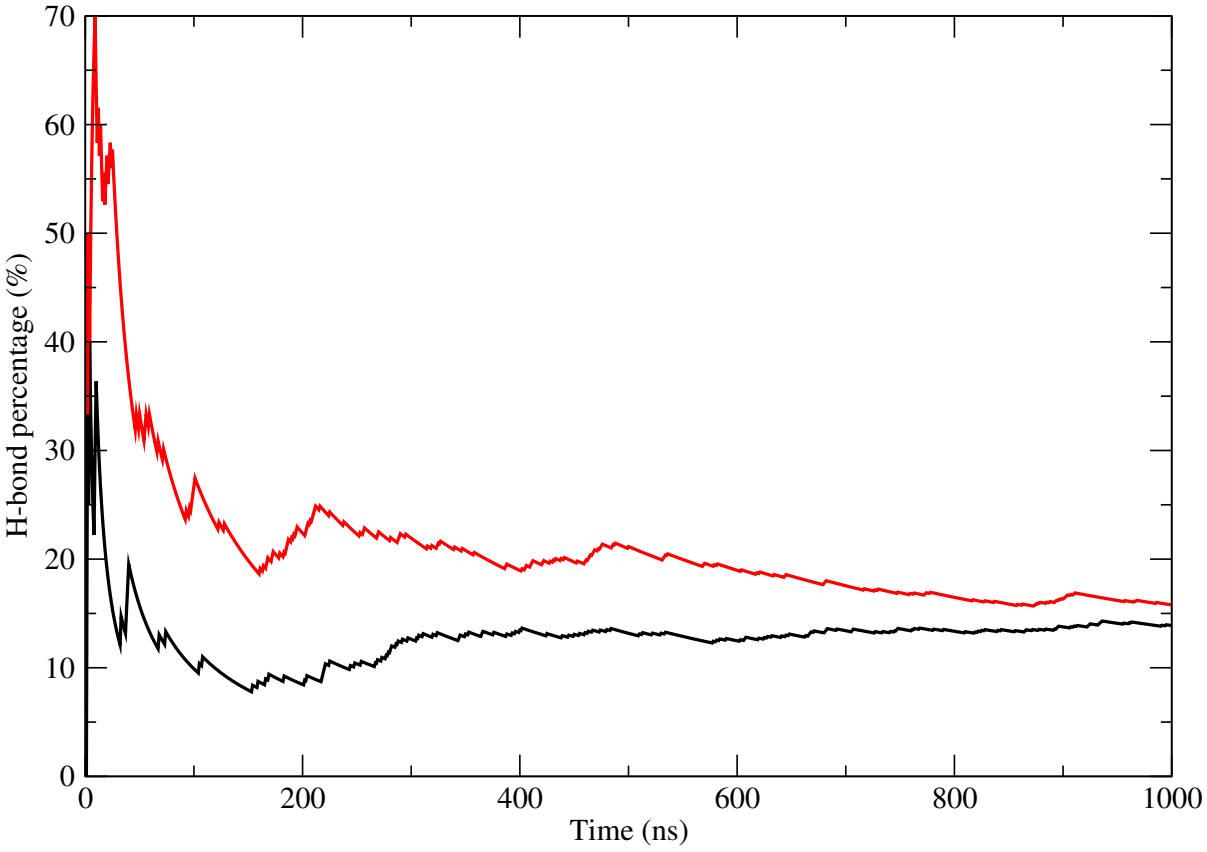


Figure 6.6 Hydrogen bond population versus time for CCG5$_{emotif}$ as obtained from BSC0 simulations. Black: C8(N4)-G6(N3); red: C23(N4)-G21(N3).

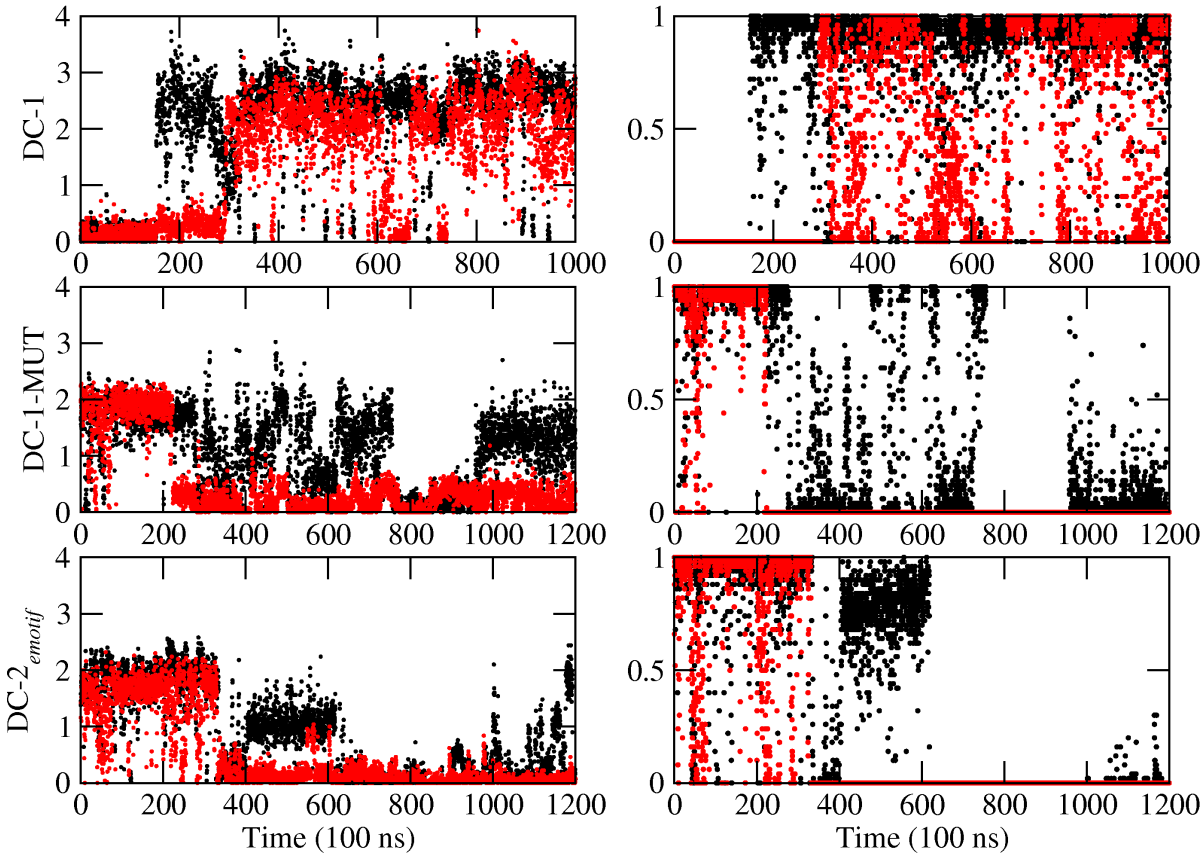


Figure 6.7 For hexanucleotide repeats, number of hydrogen bonds between mismatched bases at position i and nucleotides along the same strand at position $i-2$. Left: Number of total hydrogen bonds between C7 and nucleotide 5 (black); and C18 and nucleotide 16 (red). Right: Number of most important hydrogen bonds, in black: C7(N4)-G5(N3) for DC-1; C7(N4)-C5(O2) for DC-1-MUT and DC-2$_{emotif}$; and between equivalent positions in the other strand (C18 and G16 or C16) in red. Data is averaged every 250 ps.

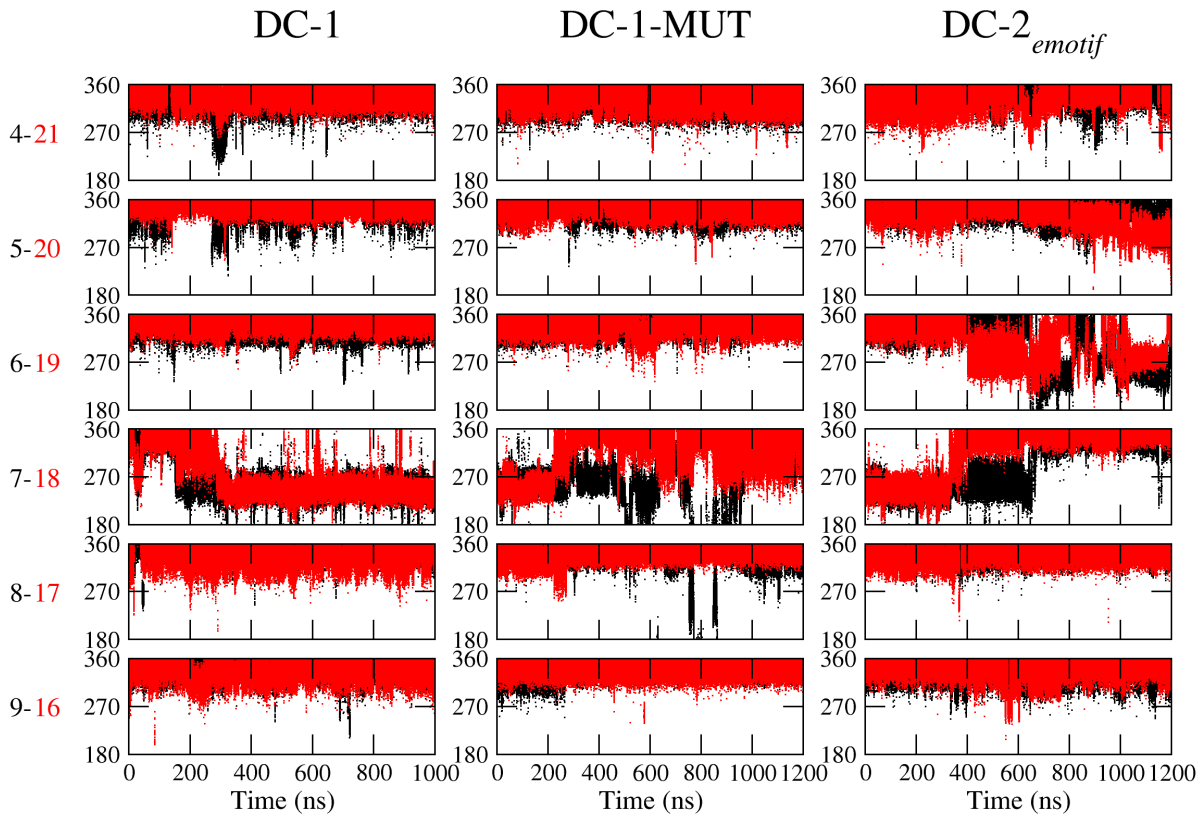


Figure 6.8 Backbone torsion angles ($\alpha+\gamma$) for hexanucleotide repeats. Sum of torsion angles ($\alpha+\gamma$) for bases 4–6 (black) and 16–21 (red) as a function of time for DC-1 (left), DC-1-MUT (middle) and DC-2$_{emotif}$ (right).

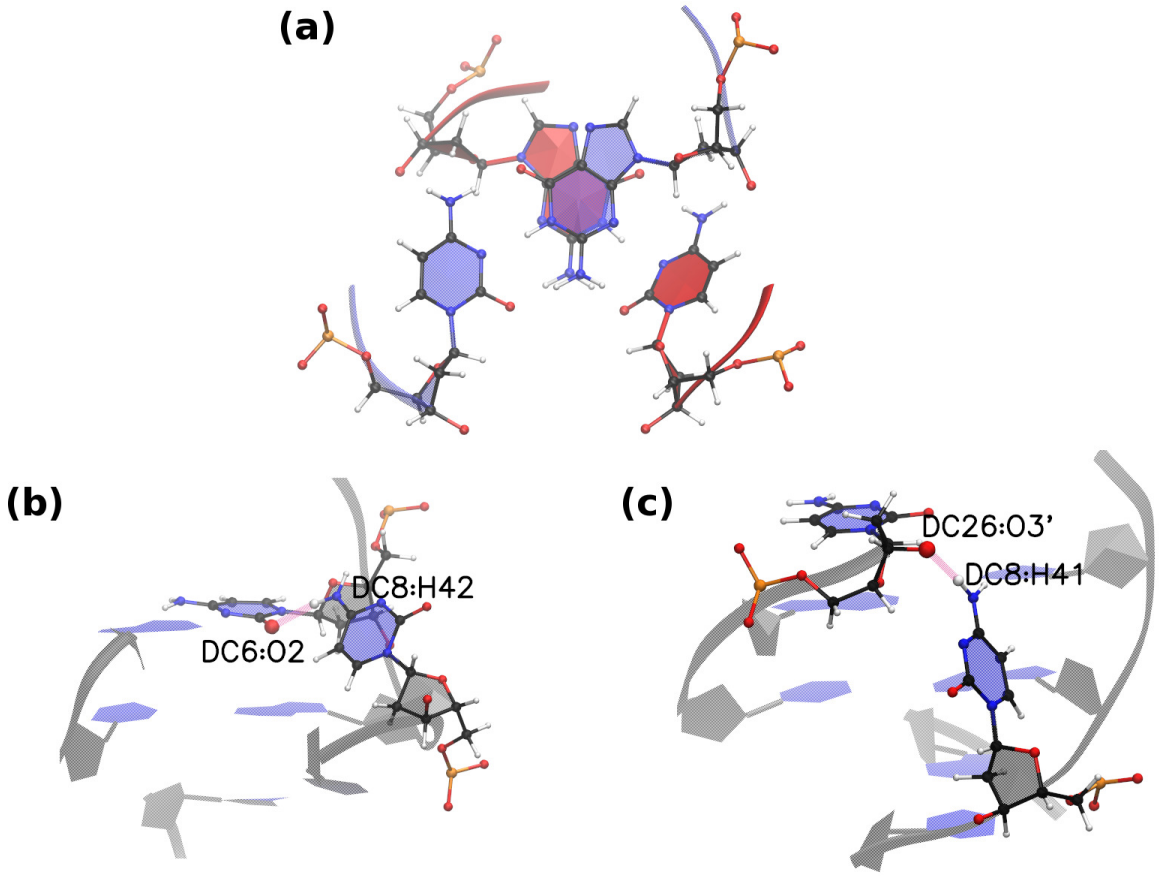


Figure 6.9 Pseudo GpC stacking L_L in GCC trinucleotide repeats. (a) G-G stacking of the hexagon part on the base for the pseudo GpC step L_L; (b) Most populated hydrogen bond O2-N4 for the BSC0 and OL15 force fields; (c) Most populated hydrogen bond O3'-N4 for BSC1 results.

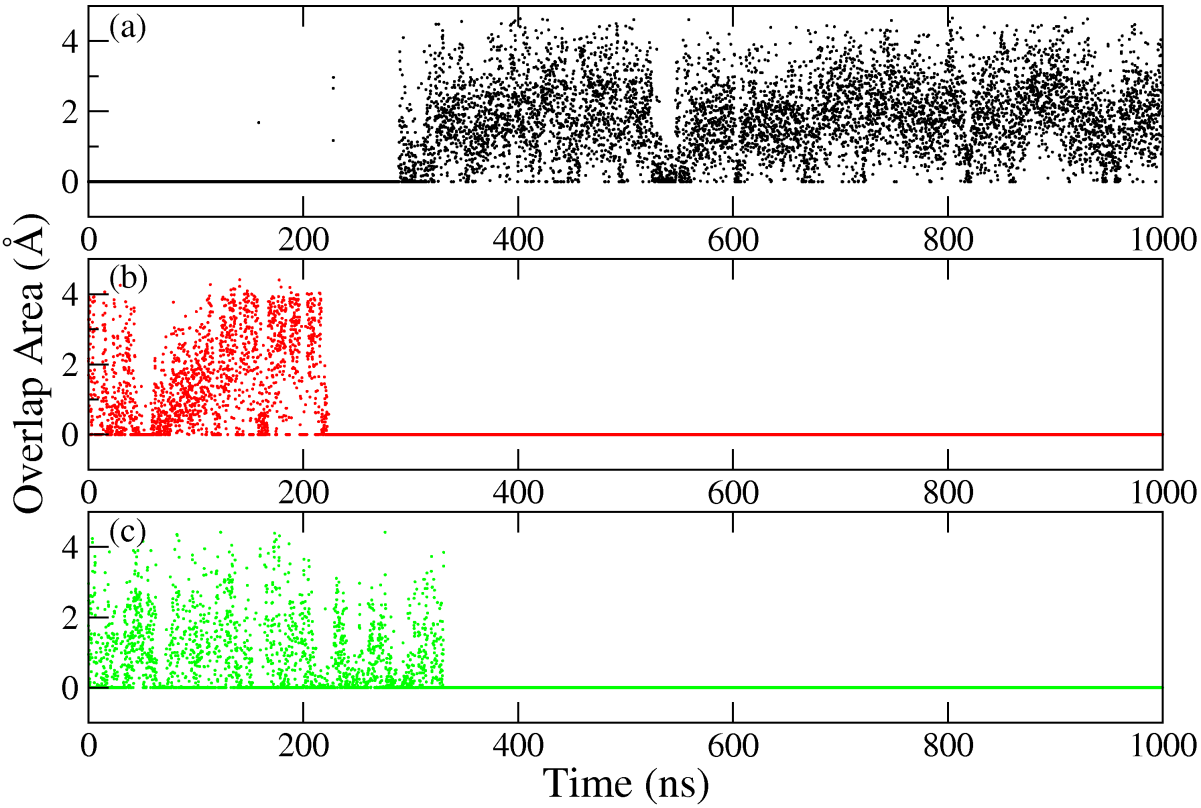


Figure 6.10 Overlap areas of the basepair ring atoms of pseudo steps in hexanucleotides. Specifically, bases 6 and 8; and 17 and 19 in Fig. 1 (g), (j), (k) are considered. Here, we show results for (a) pseudo GpC step L_{LC} in DC-1; (b) pseudo GpC step L_{LC} in DC-1-MUT; (c) pseudo CpG step M_{MC} in DC-2_{emotif}.

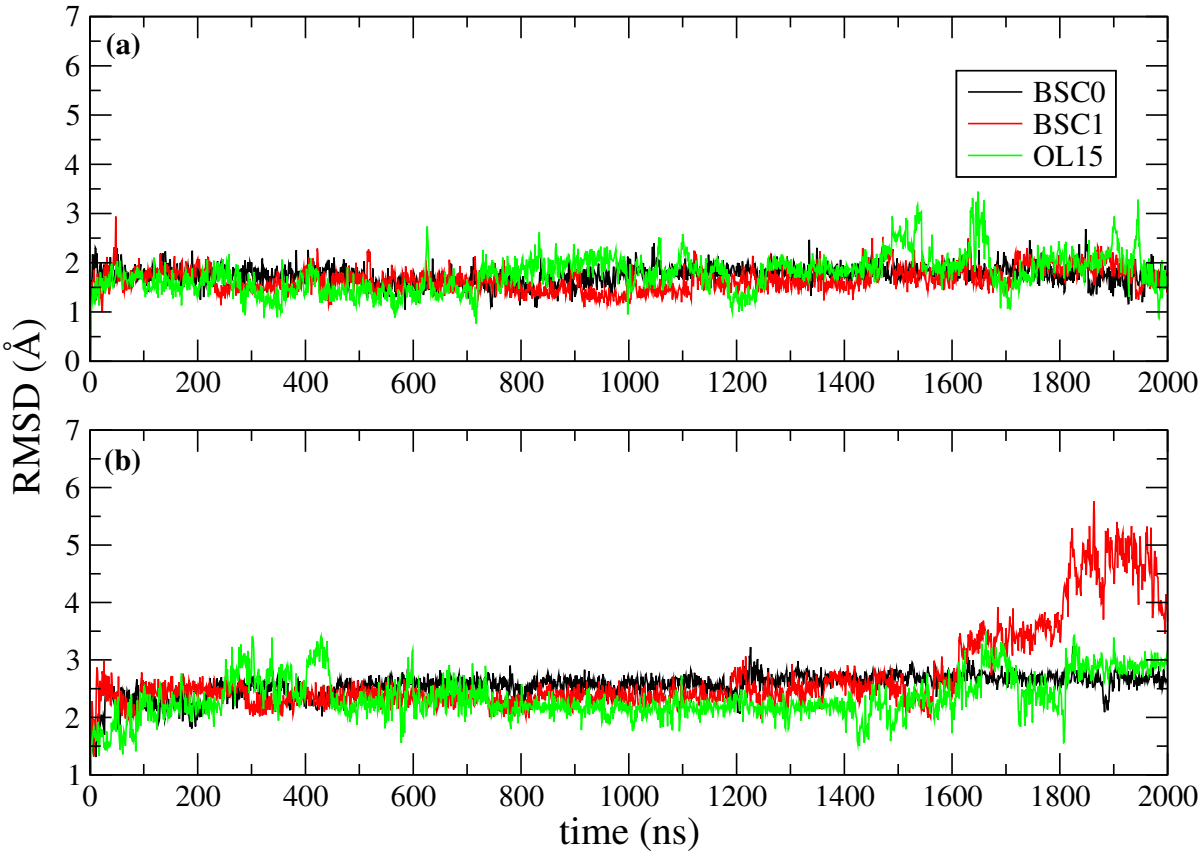


Figure 6.11 RMSD of the extended e-motif in trinucleotide repeats. RMSD of the central section of the extended e-motif (residues 4-12,18-26) with respect to the initial frame in a 2 μs MD simulation. Different colors are used to represent different force field results. Results are shown for: (a) $GCC4_{extended}$ (b) $CGG4_{extended}$.

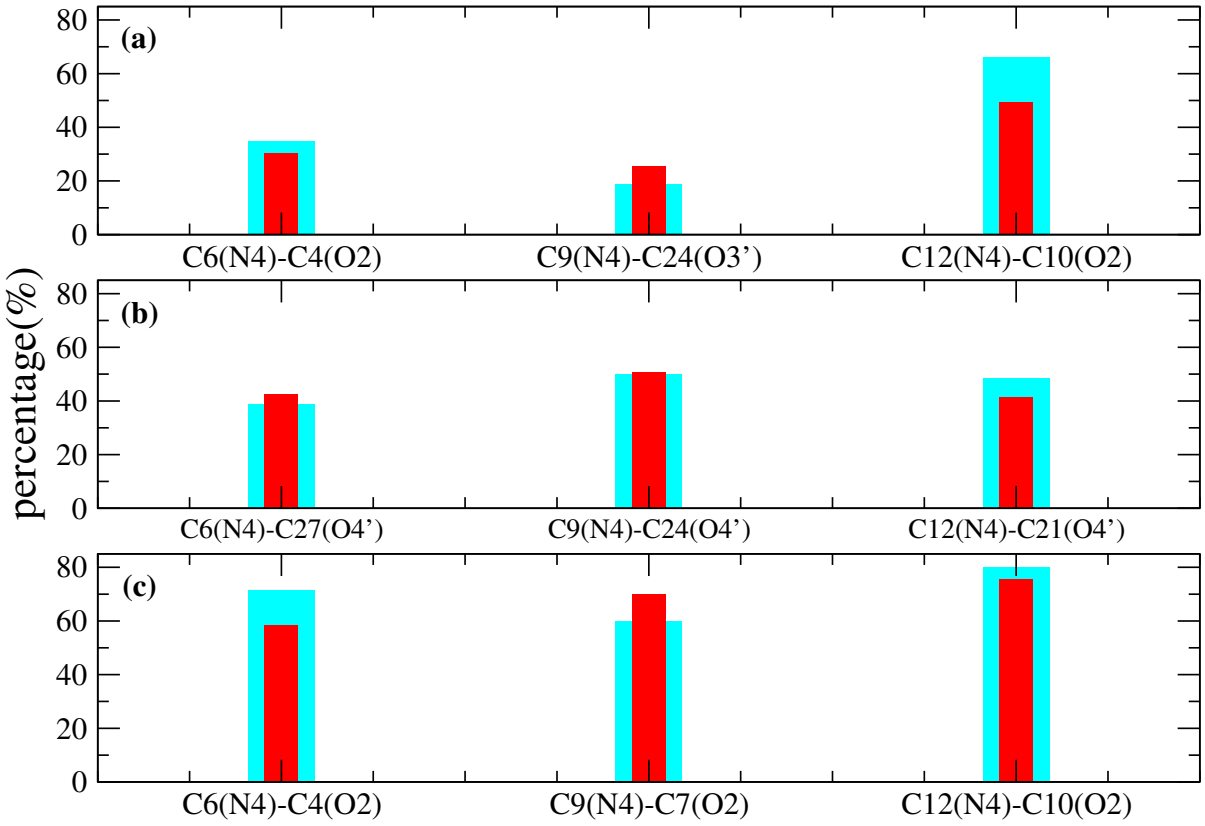


Figure 6.12 Hydrogen bonds in extended e-motif GCC duplexes. Hydrogen bonds with highest percentage in GCC4$_{extended}$ associated with the extruded C6, C9, C12 bases and the symmetric ones on the other strand. Cyan color shows the percentage of the labeled hydrogen bonds and red color shows the symmetric ones on the other strand. Results are given for the different force fields: (a) BSC0; (b) BSC1; (c) OL15.

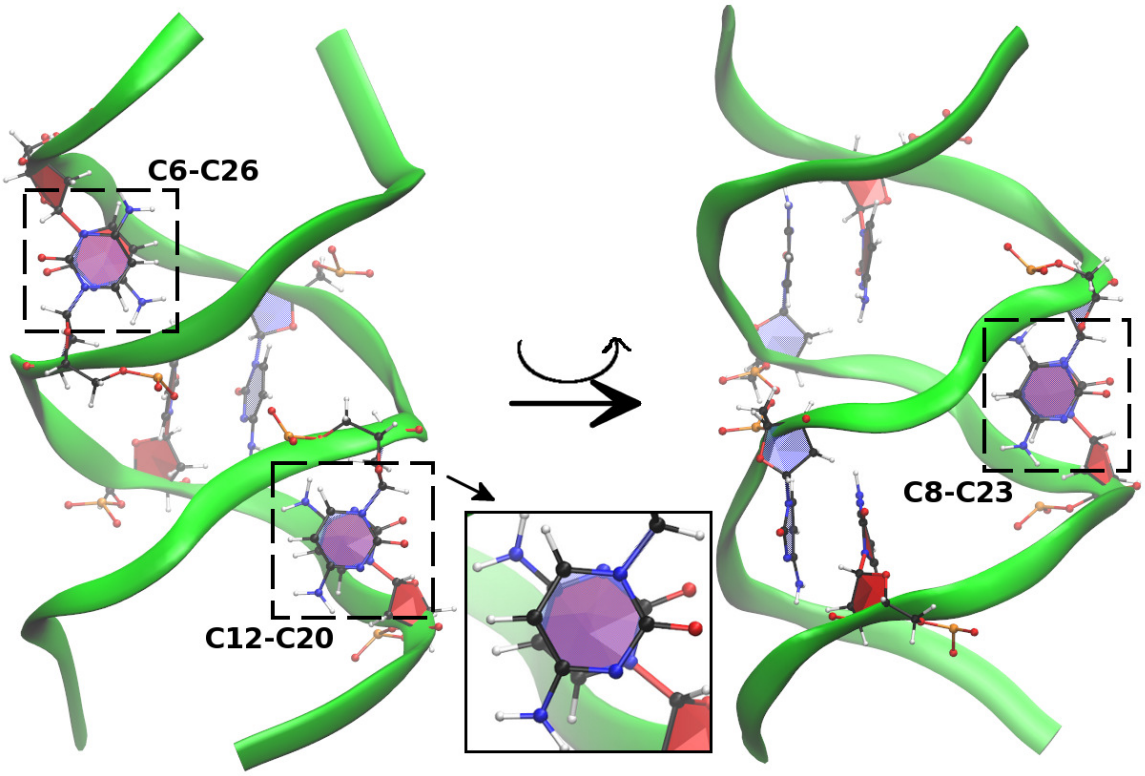


Figure 6.13 CC stacking pattern for the GCC4$_{extended}$ duplex for the BSC1 results. Left figure shows the stacking of C6-C26 and C12-C20, right figure shows the stacking of C8-C23 after a rotation around the central axis. The inset in the middle shows the close view of C12-C20 stacking.

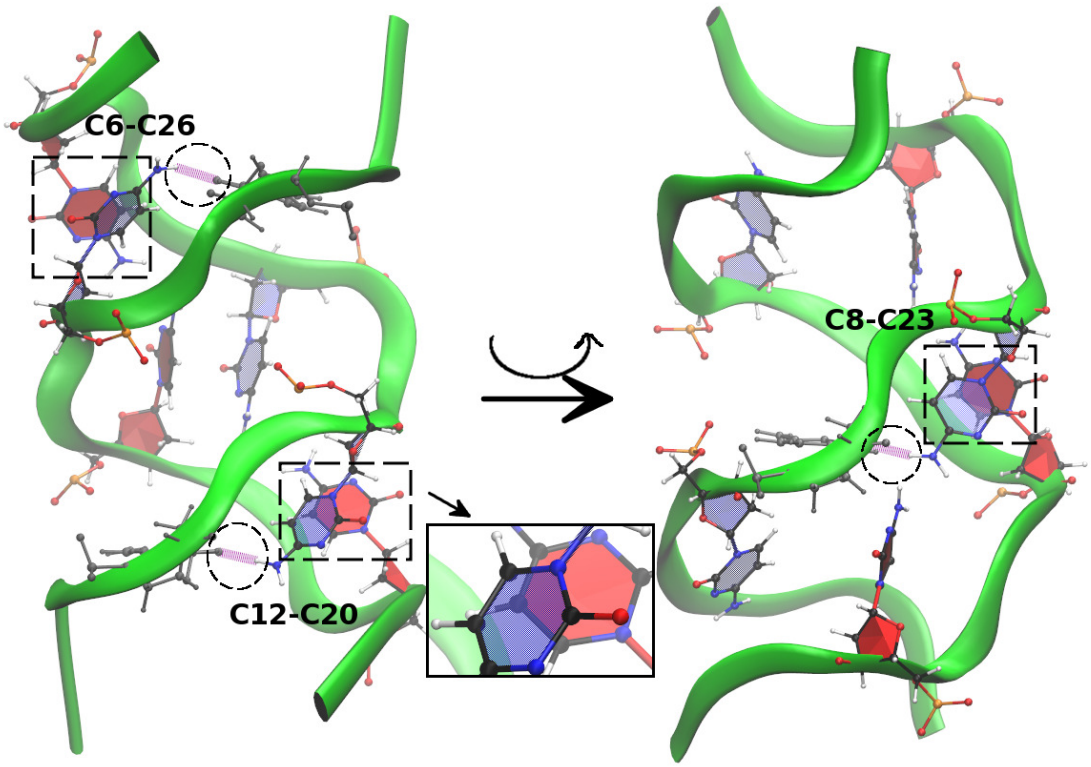


Figure 6.14 CC stacking pattern for the GCC4$_{extended}$ duplex for the OL15 results. Left figure shows the stacking of C6-C26 and C12-C20, right figure shows the stacking of C8-C23 after a rotation around the central axis. The stacking is not as strong as in BSC1 because the extruded C bases have hydrogen bonds with bases along the same strand. Hydrogen bonds are shown in purple inside the circles. The inset in the middle shows the close view of C12-C20 stacking.

Acknowledgements

The work was supported by the National Institute of Health [NIH-R01GM118508]; the National Science Foundation (NSF) [SI2-SEE-1534941]; and the Extreme Science and Engineering Discovery Environment (XSEDE) [TG-MCB160064].

References

[1] CT McMurray. DNA secondary structure: A common and causative factor for expansion in human disease. *Proc. Natl. Acad. Sci. USA*, 96:1823–1825, 1999.

[2] CE Pearson, KN Edamura, and JD Cleary. Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.*, 6:729–742, 2005.

[3] Mirkin, S. Expandable DNA repeats and human disease. *Nature*, 447:932, 2007.

[4] I Oberle, F. Rouseau, D. Heitz, D. Devys, S. Zengerling, and J.L. Mandel. Molecular-basis of the fragile-X syndrome and diagnostic applications. *Am. J. of Human Genet.*, 49:76, 1991.

[5] P Giunti, MG Sweeney, M Spadaro, C Jodice, A Novelletto, P Malaspina, M Frontali, and Harding AE. The trinucleotide repeat expansion on chromosome 6p (sca1) in autosomal dominant cerebellar ataxias. *Brain*, 117:645–649, 1994.

[6] V Campuzano, L Montermini, MD Molto, L Pianese, M Cossee, F Cavalcanti, E Monros, F Rodius, F Duclos, A Monticelli, F Zara, J Canizares, H Koutnikova, SI Bidichandani, C Gellera, A Brice, P Trouillas, G DeMichele, A Filla, R DeFrutos, F Palau, PI Patel, S DiDonato, JL Mandel, S Cocozza, M Koenig, and M Pandolfo. Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, 271:1423–1427, 1996.

[7] Wells, R.D. and Warren, S. *Genetic instabilities and neurological diseases*. Academic Press, San Diego, CA, Elsevier, 1998.

[8] CE Pearson and RR Sinden. Slipped strand DNA (S-DNA and SI-DNA), trinucleotide repeat instability and mismatch repair: A short review. In Sarma, RH and Sarma, MH, editor, *Structure, Motion, Interaction and Expression of Biological Macromolecules, Vol 2*, pages 191–207. US NIH, 1998. 10th Conversation in Biomolecular Stereodynamics Conference, SUNY Albany, JUN 17-21, 1997.

[9] Sergei M. Mirkin. DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. in Struct. Biol.*, 16:351–358, 2006.

[10] Orr, H. and Zoghbi, H. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, 30:575, 2007.

[11] Hutton Moore, Patricia W. Greenwell, Chin-Pin Liu, Norman Arnheim, and Thomas D. Petes. Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc. Natl. Acad. Sci.*, 96(4):1504–1509, 1999.

[12] RD Wells, R Dere, ML Hebert, M Napierala, and LS Son. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucl. Acids Res.*, 33:3785–3798, 2005.

[13] Jane C. Kim and Sergei M. Mirkin. The balancing act of DNA repeat expansions. *Curr. Opin. in Genet. & Devel.*, 23:280–288, 2013.

[14] Vincent Dion and John H. Wilson. Instability and chromatin structure of expanded trinucleotide repeats. *Trends in Genet.*, 25:288–297, 2009.

[15] Cynthia T. McMurray. Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair*, 7:1121–1134, 2008.

[16] Yunfu Lin and John H Wilson. Transcription-induced DNA toxicity at trinucleotide repeats: double bubble is trouble. *Cell Cycle*, 10:611–618, 2011.

[17] Laura P. W. Ranum and Thomas A. Cooper. RNA-mediated neuromuscular disorders. *Ann. Rev. of Neuroscience*, 6:259–277, 2006.

[18] Ling-Bo Li and Nancy M. Bonini. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends in Neurosciences*, 33:292–298, 2010.

[19] P Jin, DC Zarnescu, FP Zhang, CE Pearson, JC Lucchesi, K Moses, and ST Warren. RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in Drosophila. *Neuron*, 39:739–747, 2003.

[20] H Jiang, A Mankodi, MS Swanson, RT Moxley, and CA Thornton. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum. Mol. Genet.*, 13:3079–3088, 2004.

[21] R. Daughters, D. Tuttle, W. Gao, Y. Ikeda, M. Moseley, T. Ebner, M. Swanson, and L. Ranum. RNA Gain-of-Function in Spinocerebellar Ataxia Type 8. *PLoS Genet.*, 5:e1000600, 2009.

[22] W. Krzyzosiak, K. Sobczak, M. Wojciechowska, A. Fiszer, A. Mykowska, and P. Kozlowski. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nuc. Acids Res.*, 40:11–26, 2012.

[23] V Campuzano, L Montermini, Y Lutz, L Cova, C Hindelang, S Jiralerspong, Y Trottier, SJ Kish, B Faucheux, P Trouillas, FJ Authier, A Durr, JL Mandel, A Vescovi, M Pandolfo, and M Koenig. Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum. Mol. Genet.*, 6:1771–1780, 1997.

[24] E. Kim, M. Napierala, and S. Dent. Hyperexpansion of GAA repeats affects post-initiation steps of FXN transcription in Friedreich's ataxia. *Nucl. Acids Res.* , 39:8366–8377, 2011.

[25] D. Kumari, R. Biacsi, and K. Usdin. Repeat expansion in intron 1 of the Frataxin gene reduces transcription initiation in Friedreich ataxia. *Faseb J.*, 25:895, 2011. Experimental Biology Meeting 2011, Washington, DC, APR 09-13, 2011.

[26] T. Punga and M. Buehler. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *Embo Mol. Medicine*, 2:120–129, 2010.

[27] Ying-Hui Fu, Derek P.A. Kuhl, Antonio Pizzuti, Maura Pieretti, James S. Sutcliffe, Stephen Richards, Annemieke J.M.H. Verkert, Jeanette J.A. Holden, Raymond G. Fenwick Jr., Stephen T. Warren, Ben A. Oostra, David L. Nelson, and C.Thomas Caskey. Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. *Cell*, 67(6):1047–1058, 1991.

[28] Nan Zhong, Weina Ju, James Pietrofesa, Daowen Wang, Carl Dobkin, and W. Ted Brown. Fragile X "gray zone" alleles: AGG patterns, expansion risks, and associated haplotypes. *Am J Med Genet.*, 64(2):261–5, 1996.

[29] C. Dombrowski, S. L*é*vesque, M. L. Morel, P. Rouillard, K. Morgan, and F. Rousseau. Premutation and intermediate-size FMR1 alleles in 10 572 males from the general population: loss of an AGG interruption is a late event in the generation of fragile X syndrome alleles. *Hum Mol Genet.*, 11(4):371–378, 2002.

[30] RJ Hagerman, M Leehey, W Heinrichs, F Tassone, R Wilson, J Hills, J Grigsby, B Gage, and PJ Hagerman. Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X. *Neurology*, 57(1):127–30, 2001.

[31] Stephanie L. Sherman. Premature Ovarian Failure among Fragile X Premutation Carriers: Parent-of-Origin Effect? *Am J Hum Genet.*, 67(1):11–3, 2000.

[32] I.A. Glass. X linked mental retardation. *J. Med. Genet.*, 28:361–371, 1991.

[33] Yanghong Gu, Ying Shen, Richard A. Gibbs, and David L. Nelson. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nat. Genet.*, 13(1):109–113, 1996.

[34] Baorong Zhang, Jun Tian, Yaping Yan, Xinzhen Yin, Guohua Zhao, Zhiying Wu, Weihong Gu, Kun Xia, and Beisha Tang. CCG polymorphisms in the huntingtin gene have no effect on the pathogenesis of patients with Huntington's disease in mainland Chinese families. *J. Neurol. Sci.*, 312(1-2):92–96, 2012.

[35] Claudia Braida, Rhoda K.A. Stefanatos, Berit Adam, Navdeep Mahajan, Hubert J.M. Smeets, Florence Niel, Cyril Goizet, Benoit Arveiler, Michel Koenig, Clotilde Lagier-Tourenne, Jean-Louis Mandel, Catharina G. Faber, Christine E.M. de Die-Smulders, Frank Spaans, and Darren G. Monckton. Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum. Mol. Genet.*, 19(1-2):1399–1412, 2010.

[36] Mariely DeJesus-Hernandez, Ian R. Machkenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, NiCole A. Finch, Heather Flynn, Jennifer Adamson, Naomi Kouri, Aleksandra Wojtas, Pheth Sengdy, Ging-Yuek R. Hsiung, Anna Karydas, William W. Seeley, Keith A. Josephs, Giovanni Coppola, Daniel H. Geschwind, Zbigniew K. Wszolek, Howard Feldman, David S. Knopman, Ronald C. Petersen, Bruce L. Miller, and Dennis W. Dickson. Expanded ggggcc hexanucleotide repeat in noncoding region of c9orf72 causes chromosome 9p-linked ftd and als. *Neuron*, 72(2):245–256, Oct. 2011.

[37] Alan E. Renton, Elisa Majounie, Adrian Waite, Javier Simon-Sanchez, Sara Rollinson, J. Raphael Gibbs, Jennifer C. Schymick, Hannu Laaksovirta, John C. van Swieten, Liisa Myllykangas, Hannu Kalimo, Anders Paetau, Yevgeniya Abramzon, Anne M. Remes, Alice Kaganovich, Sanja W. Scholz, Jamie Duckworth, Jinhui Ding, Daniel W. Harmer, Dena G. Hernandez, Janel O. Johnson, Kin Mok, Mina Ryten, Danyah Trabzuni, and Rita J. Guerreiro. A hexanucleotide repeat expansion in c9orf72 is the cause of chromosome 9p21-linked als-ftd. *Neuron*, 72(2):257–268, Oct. 2011.

[38] XL Gao, XN Huang, GK Smith, MX Zheng, and HY Liu. New antiparallel duplex motif of DNA CCG repeats that is stabilized by extrahelical basis symmetrically located in the minor-groove. *J. Am. Chem. Soc.*, 117:8883–8884, 1995.

[39] MX Zheng, XN Huang, GK Smith, XY Yang, and XL Gao. Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study. *J. of Mol. Biol.*, 264:323–336, 1996.

[40] A Yu, MD Barren, RM Romero, M Christy, B Gold, JL Dai, DM Gray, IS Haworth, and M Mitas. At physiological pH, d(CCG)(15) forms a hairpin containing protonated cytosines and a distorted helix. *Biochem*, 36:3687–3699, 1997.

[41] Pornchai Rojsitthisak, Rebecca M. Romero, and Ian S. Haworth. Extrahelical cytosine bases in DNA duplexes containing d[GCC]$_n$·d[GCC]$_n$ repeats: detection by a mechlorethamine crosslinking reaction. *Nucleic Acids Res.*, 29(22):4716–23, 2001.

[42] Y. Zhang, C. Roland, and C. Sagui. Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases. *ACS Chemical Neuroscience*, 8:578–591, 2016.

[43] D. A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T.Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *"AMBER 16"*. University of California, San Francisco, 2016.

[44] I. Ivani, P. Dans, A. Noy, A. Prez, I. Faustino, A. Hopsital, J. Walther, P. Andrio, R. Goni, A. Balaceanu, G. Portella, F. Battistini, J. GelpA, C. Gonzlez, M. Vendruscolo, C. Laughton, S. Harris, D. Case, and M. Orozco. Parmbsc1: A refined force field for DNA simulations. *Nature Meth.*, 13:55–58, 2016.

[45] A. Perez, I. Marchan, 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.*, 92:3817–3829, 2007.

[46] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham III, Rodrigo Galindo-Murillo, and Petr 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.*, 11(12):5723–36, 2015.

[47] W. L. Jorgensen, J. Chandrasekhar, J. Madura, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926 – 935, 1983.

[48] I. S. Joung and T. E. Cheatham. Determination of Alkali and Halide Monovalent Ion Parameters for use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B*, 112:9020–9041, 2008.

[49] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577 – 8593, 1995.

[50] M. Moradi, V. Babin, C. Roland, and C. Sagui. Reaction path ensemble of the b-z-dna transition: a comprehensive atomistic study. *Nucleic Acids Res.*, 41:33–43, 2013.

[51] F. Pan and V. Man and C. Roland and C. Sagui. Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats. *Biophys. J.*, 113:19–36, 2017.

[52] JM Darlow and DRF Leach. Secondary structures in d(CGG) and d(CCG) repeat tracts. *J. Mol. Biol.*, 275:3–16, 1998.

[53] JM Darlow and DRF Leach. Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts in vivo. *J. Mol. Biol.*, 275:17–23, 1998.

[54] D. L. Nelson, H. T. Orr, and S. T. Warren. The unstable repeats three evolving faces of neurological disease. *Neuron*, 77:825–843, 2013.

[55] C. T. McMurray. Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.*, 11:786–799, 2010.

[56] A. R. La Spada and J. P. Taylor. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.*, 11:247–258, 2010.

[57] C. T. McMurray. Mechanisms of DNA expansion. *Chromosoma*, 104:2–13, 1995.

[58] C. T. McMurray. Influence of hairpins on template reannealing at trinucleotide repeat duplexes: a model for slipped DNA. *Biochemistry*, 37:9426–9434, 1998.

[59] Do-Yup Lee and C. T. McMurray. Trinucleotide expansion in disease: why is there a length threshold? *Curr. Opin. Genet. Dev.*, 26:131–140, 2014.

Chapter 7

Conclusions

This thesis presents a study focused on the structure and dynamics of atypical DNA and RNA double helices related with trinucleotide repeats, include CAG, CCG and CGG repeats. Before the study of atypical structures, we have the investigation of ion atmosphere and a new fast laser melting method on regular duplexes. Besides, the study of e-motif structure that induced by the extrahelical C bases in atypical structures is presented as an individual chapter.

In Chapter 2, a comparative study of ion distributions was carried out on four kinds of duplexes. The ions investigated include Na^+, K^+, and Mg^{2+}, with various concentrations of their chloride salts. Our results quantitatively describe the characteristics of the ionic distributions for different structures at varying ionic strengths, tracing these differences to nucleic acid structure and ion type. Several binding pockets with rather long ion residence times are described, both for the monovalent ions and for the hexahydrated $Mg[(H_2O)_6]^{2+}$ ion. The conformations of these binding pockets include direct binding through desolvated ion bridges in the GpC steps in B-DNA and A-RNA; direct binding to backbone oxygens; binding of $Mg[(H_2O)_6]^{2+}$ to distant phosphates, resulting in acute bending of A-RNA; tight "ion traps" in Z-RNA between C-O2 and the C-O2' atoms in GpC steps; and others.

In Chapter 3, we use the classical laser pulse, which is typical a free-electron based laser [1–5], with specific oscillation characteristics to target specific bonds within a given structure and thereby study the unfolding of the structure. Specifically, this method was used to investigate the dissociation of amyloid fibrils into soluble monomers [3–6], and the break-up of a peptide-based nanotube [7]. Comparing healing and, possibly, relative stability of different structures via laser induced melting is clearly not a universal

technique. Key to the application of the technique is the fact that the perturbation must affect both structures in the same way (i.e., the damage to the structures must be the same). Our study is targeted on the melting of B-DNA and Z-DNA with the same sequence. Quantitative interpretations shoe that B-DNA is more stable than Z-DNA. This is consistent with former experimental results [8, 9] and related simulations [10, 11]. In summary, this laser melting process does not give such quantities as relative free energies. Rather, it provides one with a "more-or-less" (stable) comparison. This, however, could be extremely useful for stability rankings in competing candidate structures that cannot be easily predicted with other tools, including non-standard secondary structures with noncanonical base pairs and mismatches. In addition, the method applies to very realistic, fully solvated, fully atomistic systems, that in principle can faithfully reproduce experimental setups.

In Chapter 4, 5 and 6, we present the studies of different kinds of trinucleotide repeats. Chapter 4 shows free energy calculations and MD studies to determine the preferred conformations of the A-A mismatches in $(CAG)_n$ and $(GAC)_n$ trinucleotide repeats (n=1 or 4) and the way in which these mismatch conformations affect the overall structure of RNA and DNA duplexes. The main findings are following: 1) The global minimum (A1) of the various free energy maps corresponds to A-A mismatches stacked inside the core of the helix with anti-anti conformations in the RNA sequences and (high-anti)-(high-anti) conformations in the DNA sequences. In terms of the free energy, the next higher minimum corresponds to anti-syn conformations, while syn-syn conformations are even higher. 2) DNA helices near the global minimum are very dynamic exhibiting large fluctuations. RNA helices still fluctuate, but with considerably lesser amplitude than DNA. 3) Free energy barriers between minima corresponding to different states of the glycosyl torsion angle χ are rather high, which results in low transition rates during regular MD. 4) Rates of MD transitions of the A mismatches between different χ categories are higher for RNA than DNA. 5) Several mechanisms for the transitions anti-syn $\rightarrow$ anti-anti and syn-syn $\rightarrow$ anti-syn have been identified both through the major and minor grooves. 6) DNA-$(CAG)_4$ and DNA-$(GAC)_4$ duplexes in anti-anti conformations experience some degree of unwinding. In Chapter 5, we have presented results for MD simulations and free energy calculations for both CCG and GGC trinucleotide repeats, with either CpG or GpC steps, for both RNA and DNA. The global minimum of the free energy maps associated with C·C mismatches in the four duplexes RNA/DNA CCG/GCC correspond to anti-anti conformations, and

G·G mismatches in the four duplexes RNA/DNA GGC/CGG correspond to anti-syn conformations. For DNA, the force fields BSC0, BSC1 and OL15 give similar free energy maps for the mismatch configurations. The main difference is that the minima are deeper in BSC1 and OL15 providing for a more rigid DNA duplex with respect to that in BSC0. When mismatches are initially placed in non-equilibrium conformations, intra-helical C·C mismatches make the transition towards the global minimum faster than G·G mismatches. DNA duplexes in the GCC reading frame, with CpG steps between the Watson-Crick basepairs, exhibit the e-motif. This e-motif is individually discussed in Chapter 6, together with the homoduplexes of hexanucleotide repeats. In that chapter, a comparative study is carried out based on different force fields and different structures, leading to the following conclusions: 1) The e-motif transition can be observed under BSC0 force field and it can remain stable under BSC0, BSC1 and OL15 force fields. 2) In the e-motif, the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the $i-2$ residue (i.e., towards the 5' direction in each strand). Also, it is partially stabilized by the formation of hydrogen bonds between the C bases (i residue) in a mismatch and the $i-2$ bases 3) Creation of the e-motif is favored by the formation of pseudo GpC steps when the bases in the C·C mismatches are extruded. Consequently, the e-motif is stable in paired-end homoduplexes of (GCC) and (CCCGGC) SSRs, but not in the other reading frames. 4) The extended e-motif is stabilized by highly cooperative interactions, including the stacking of pseudo GpC steps, the hydrogen bonds between the mismatched bases and other nucleotides, and the stacking of the extruded C bases themselves.

Near term future extensions of this thesis fall into two parts: 1) The study of CTG (CUG in RNA) repeats, which are directly related with myotonic dystrophy type 1 (DM1) and spinocerebellar ataxias type 8 (SCA8). Once the expansion of CTG repeats in DNA are transcribed, they form RNA hairpins with the UU mismatches in the stem part. These loops attract muscleblind-like 1 (MBNL1) protein leading to the disease. Similarly, as the completed work, one can obtain the free energy landscape of the UU or TT mismatches using the ABMD method from duplexes with single mismatches. In addition, the regular MD simulation will be used to investigate the structural properties, transition mechanisms, ion atmosphere and other characteristics. Currently, we have finished some work on the RNA duplexes with CUG repeats. The free energy maps show similar results with another simulation using umbrella sampling [12], with the anti-anti as the global minimum. There

are also some observed transitions from anti-syn to anti-anti in regular MD simulation. Furthermore, a more comparative study needs to be carried out, covering the DNA duplexes with CTG repeats. 2) So far, our research has been focused on the double helices with mismatches. However, there is experimental consensus that the most typical DNA and RNA TR secondary structures in the initial stages of expansion are hairpins with stem-loop structure [13–16]. Therefore, it is more essential to study the structure and dynamics of hairpins with TR. In the scheme introduced by Darlow and Leach [17, 18], hairpins were classified according to the alignment of the sides of the hairpins and the presence of an odd or even number of unpaired bases in the loop. For example, the single strand with CAG repeats may form two kinds of hairpins: odd number of repeats leads to triloops with CAG sequence in the loop; even number of repeats leads to tetraloops with AGCA sequence in the loop. So far, there has been experimental researches on different kinds of loops like for RNA tetraloops [19–22], RNA triloop [23] and DNA tetraloop [24], as well as simulations with umbrella sampling to get the free energy maps [25,26]. However, very little research has targeted the hairpins with TR sequences. So studies of hairpins with TR represent an important research direction for the future. Currently, we have finished some preliminary simulations with CAG repeat hairpins. Both the hairpins with odd and even number of repeats appear to be stable and the one with AGCA sequence in the loop is potentially the most stable. Based on this, more work will be carried out on the structural conformations of the loops, specifically as a function of salt concentrations and different force fields.

References

[1] D. Ozawa, H. Yagi, T. Ban, A. Kameda, T. Kawakami, H. Naiki, and Y. Goto. Destruction of amyloid fibrils of a beta-microglobin fragment by laser beam irradiation. *J. Biol. Chem.*, 284:1009, 2009.

[2] H. Yagi, D. Ozawa, K. Sakuri, T. Kawakami, H. Kuyama, O. Nishimura, T. Shimanouchi, R. Kuboi, H. Naiki, and Y. Goto. Laser-induced propagation and destruction of amyloid beta fibrils. *J. Biol. Chem.*, 285:19660, 2010.

[3] T. Kawasaki, J. Fujioka, T. Imai, and K. Tsukiyama. Effect of mid-infrared free-electron laser irradiation on refolding of amyloid-like fibrils of Lysozyme into native form. *The Protein Journal*, 31:710, 2012.

[4] T. Kawasaki, J. Fujioka, T. Imai, K. Torigoe, and K. Tsukiyama. Mid-infrared free-electron laser tuned to the amide I band for converting insoluble amyloid-like protein fibrils into the soluble monomeric form. *Lasers in Medical Science*, 29:1701, 2014.

[5] T. Kawasaki, T. Yaji, T. Imai, T. Ohta, and K. Tsukiyama. Synchrotron-infrared microscopy analysis of amyloid fibrils irradiated by mid-infrared free-electron laser. *Am. J. of Anal. Chem.*, 5:384, 2014.

[6] V. H. Man, Ph. Derreumaux, M. S. Li, C. Roland, C. Sagui, and Ph. Nguyen. Picosecond dissociation of amyloid fibrils with infrared laser: a nonequilibrium simulation study. *J. Chem. Phys.*, submitted:, 2015.

[7] V.H. Man, P.M. Truong, Ph. Derreumaux, M.S. Li, C. Roland, C. Sagui, and Ph. Nguyen. Picosecond melting of peptide nanotube with infrared laser: a nonequilibrium study. *Nano Lett.*, submitted:, 2015.

[8] L.J. Peck and J.C. Wang. Energetics of the B-to-Z transition in DNA. *Proc. Natl. Acad. Sci. (USA)*, 80:6206, 1983.

[9] P.S. Ho. The non-B-DNA structure of d(CA/TG)(N) does not differ from that of Z-DNA. *Proc. Natl. Acad. Sci. (USA)*, 91:9549, 1994.

[10] J. Lee, Y.G. Kim, K.K. Kim, and C. Seok. Transition between B-DNA and Z-DNA: free energy landscape for the B-Z junction propagation. *J. Phys. Chem. B*, 114:9872, 2010.

[11] M. Moradi, V. Babin, C. Roland, and C. Sagui. Reaction path ensemble of the b-z-dna transition: a comprehensive atomistic study. *Nucleic Acids Res.*, 41:33–43, 2013.

[12] I Yildirim, D Chakraborty, M. D. Disney, D. J. Wales, and G. C. Schatz. Computational Investigation of RNA CUG Repeats Responsible for Myotonic Dystrophy 1. *J. Chem. Theory Comput.*, 11:4943–4958, 2015.

[13] W. Krzyzosiak, K. Sobczak, M. Wojciechowska, A. Fiszer, A. Mykowska, and P. Kozlowski. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nuc. Acids Res.*, 40:11–26, 2012.

[14] M Mitas, A Yu, J Dill, and IS Haworth. The trinucleotide repeat sequence D(CGG)(15) forms a heat-stable hairpin containing G(syn).G(anti) base-pairs. *Biochem.*, 34:12803–12811, 1995.

[15] AM Gacy, G Goellner, N Juranic, S Macura, and CT McMurray. Trinucleotide repeats that expand in human-disease form hairpin structures in-vitro . *Cell*, 81:533–540, 1995.

[16] J Petruska, N Arnheim, and MF Goodman. Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nuc. Acids Res.* , 24:1992–1998, 1996.

[17] JM Darlow and DRF Leach. Secondary structures in d(CGG) and d(CCG) repeat tracts. *J. Mol. Biol.*, 275:3–16, 1998.

[18] JM Darlow and DRF Leach. Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts in vivo. *J. Mol. Biol.*, 275:17–23, 1998.

[19] C. R. Woese, S. Winkers, and R. R. Gutell. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci.*, 87:87, 1990.

[20] M Molinaro and Jr Tinoco I. Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. *Nucleic Acids Res.*, 23:3056–63, 1995.

[21] FM Jucker and A Pardi. Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. *Biochemistry*, 34:14416–27, 1995.

[22] Q Zhao, HC Huang, U Nagaswamy, Y Xia, X Gao, and GE Fox. UNAC tetraloops: to what extent do they mimic GNRA tetraloops? *Biopolymers*, 97:617–28, 2012.

[23] V. Lisi and F Major. A comparative analysis of the triloops in all high-resolution RNA structures reveals sequencestructure relationships. *RNA*, 13:1537–1545, 2007.

[24] M Nakano, EM Moody, J Liang, and PC Bevilacqua. Selection for thermodynamically stable DNA tetraloops using temperature gradient gel electrophoresis reveals four motifs: d(cGNNAg), d(cGNABg),d(cCNNGg), and d(gCNNGc). *Biochemistry*, 41:14281–92, 2002.

[25] N-J Deng and P Cieplak. Free Energy Profile of RNA Hairpins: A Molecular Dynamics Simulation Study. *Biophys. J.*, 98:627–636, 2010.

[26] J Miner, A Chen, and A Garcia. Free-energy landscape of a hyperstable RNA tetraloop. *Proc. Natl. Acad. Sci.*, 113:6665–6670, 2016.

APPENDICES

Appendix A

Ion distributions around left- and right-handed DNA and RNA duplexes: a comparative study - Supporting information

Ion distributions around left- and right-handed DNA and RNA duplexes: a Molecular Dynamics study [Supplementary Information]

Feng Pan, Christopher Roland, Celeste Sagui*

Center for High Performance Simulations (CHiPS) and Department of Physics, North Carolina State University, Raleigh, NC 27695-8202.

∗ E-mail: sagui@ncsu.edu

1 Materials and Methods

Large-scale MD simulations were used to explore the ion distribution around DNA and RNA sequences $(CG)_6$ in an explicit solvent environment. The simulations were carried out using the PMEMD module of the AMBER 12 [1] software package with the ff12SB force field with parameters ff99BSC0 [2] for DNA and ff99BSC0+χ_{OL3} [3, 4] for RNA. The TIP3P model [5] was used for the water molecules. The ion parameters are relatively standard and offered in the AMBER force field employed. The monovalent ion parameters for Ewald and TIP3P waters are from Ref. [6], while the the Mg^{2+} parameters were taken from from Aqvist's work [7].

Table 1. Force field parameters for different ions

	radius(Å)	mass(u)	epsilon(kcal/mol)
Na^+	1.369	22.99	0.0874393
K^+	1.705	39.10	0.1936829
Mg^{2+}	0.7926	24.305	0.8947
Cl^-	2.513	35.45	0.0355910

The long-range Coulomb interaction was evaluated by means of the Particle-Mesh Ewald (PME) method [8] with a 9 Å cutoff and an Ewald coefficient of 0.30768. Similarly, the van der Waals interaction were calculated by means of a 9 Å atom-based nonbonded list, with a continuous correction applied to the long-range part of the interaction. The production runs were generated using the leap-frog algorithm with a 1 fs timestep with Langevin dynamics with a collision frequency of 1 ps^{-1}. Data was saved every picosecond of the simulation. The SHAKE algorithm was applied to all bonds involving hydrogen atoms. The nucleic acid structures generated by means of these constant temperature ($T = 300K$) and pressure (1 atm) runs were analyzed by means of the 3DNA (v2.1) package [9] and the PTRAJ package

in AMBERTOOLS [1].

The initial coordinates of the nucleic acid structures were generated as follows. The usual right-handed B-DNA and A-RNA structures were generated using the NAB package in AMBERTOOLS (ver.12) in Arnott's conformation [10, 11], while the left-handed Z-DNA structure was available from our previous investigations [12]. For the Z-RNA structure, we first generated the $(CG)_3$ structure of Z-DNA, and then modified it to generate the corresponding RNA structure. The resulting structure was then equilibrated for 10 *ns* in an explicit solvent environment with a neutralizing number of Na^+ ions. The average structure was then analyzed using 3DNA and compared to the NMR-based Z-RNA structure found in the PDB data bank (PDB id: 1T4X) [13]. Since the local base-pair and step parameters of this equilibrated structure were found to be quite close to the experimental structure, we simply replicated the $(CG)_3$ structure to generate the initial $(CG)_6$ Z-RNA duplex. The duplexes were first equilibrated in a smaller water box (with neutralizing ions) and then placed in a larger cubic box of $\sim$ 82 Å side, filled with a suitable number of water molecules. Such a large box is necessary, since cylindral distribution functions need to be calculated to a distance of at least 30 Å. To achieve electroneutrality, 22 water molecules were replaced by Na^+ cations. For simulations involving high salt concentrations, the number of water molecules replaced by salt anions and cations was dictated by the system size and density. In order to minimize any correlation effects on the initial distribution of the ions, these were placed at random at a distance of 10 Å away from each other and the duplex structure.

Five simulations were carried out for each duplex with the following ions: (i) 22 neutralizing Na^+, no excess salt; (ii) 22 neutralizing Na^+, and 0.4 M NaCl; (iii) 22 neutralizing K^+, and 0.4 M KCl; (iv) 11 neutralizing Mg^{2+} and 0.2 M $MgCl_2$; and (v) 22 neutralizing Na^+, and 4.0 M NaCl. Here, the "salt" molarity is used to indicate *excess* salt (over the neutralizing ions). Thus, we will refer to the case (i) above as "zero salt", although it has 0.06M Na^+, due to the neutralizing ions.

The energy of each final system of oligonucleotide, waters and ions was then minimized with the nucleic acid and ions fixed, followed by further minimization where atomic motion was allowed. Subsequently, the temperature was gradually raised at constant volume from zero to 300 K over a period of 50 ps, with the DNA/RNA structure and ions restrained with a harmonic constant of 50 kcal/mol. The resulting structure was then further equilibrated at 300 K for 100 ps. The restraints on the duplex atoms and ions were then gradually reduced by decreasing the restraining harmonic constant in 5 steps during equal intervals of time. The final configuration was then used for a 1 ns constant pressure simulation

which was completely unrestrained, thereby adjusting the box size so as to achieve a system density of approximately 1 g/ml. This was then followed by 120 ns of constant pressure and temperature production runs. Equilibrium data was collected from only the last 100 ns of these runs, since a minimum of 20 ns is required in order to stabilize the ion distribution.

Our analysis was focused on calculating the distribution of the mobile ions around the nucleic acid structures. To that end, we calculated the diffusion coefficients for the ions, the cylindrical and radial distribution functions, and carried out an analysis of the efficacy of different sites to localize the ions. The diffusion coefficient (D) for the ions was obtained by conventional means:

$$D = \left\langle |\vec{r}(t) - \vec{r}(0)|^2 \right\rangle / (6\Delta t), \tag{1}$$

i.e., D is obtained by calculating the slope of the mean-square displacement as a function of time. We calculated this quantity over a 10 ns time interval, taking data every 10 ps, and averaging the result over all the ions. Turning to the radial distribution function (RDF), this is defined in terms of the distance r of an atom B from a central atom A:

$$g_{AB}(r) = \frac{\rho_B(r \mid r_A = 0)}{\rho_B}, \tag{2}$$

where ρ_B is the bulk density of B and $\rho_B(r \mid r_A = 0)$ is the conditional distance-dependent density of atoms of type B a distance r from central site A. In practice the RDF is calculated as:

$$g_{AB}(r) = \left(\frac{V}{N_B}\right) \frac{N_B(r, \Delta)}{V_B(r, \Delta)} = \left(\frac{V}{N_B}\right) \frac{N_B(r, \Delta)}{4\pi r^2 \Delta} \tag{3}$$

where $N_B(r, \Delta)$ is the average number of atoms of type B located between distances $r - \frac{1}{2}\Delta$ and $r + \frac{1}{2}\Delta$ from central atom A, and $V_B(r, \Delta)$ is the volume of the spherical slice between $r - \frac{1}{2}\Delta$ and $r + \frac{1}{2}\Delta$. RDFs are particularly useful for investigating the binding properties between specific atoms such as negatively charged nitrogens and oxygens on the duplexes and positively charged ions. In these RDFs the location of the peaks reveals the typical binding distances. In the results presented here, the RDFs between two atom types were computed as an average over all the equivalent atom pairs along the duplex. An ion can be attracted to more than one atom in the duplex, and thus it can contribute to the RDFs of different atom types in the duplex. A particularly useful variant of RDFs involves the cylindrical distribution

functions, which track the ion concentration measured radially outwards from the central axis of the duplex. The calculation of these functions is the same as for the RDFs, except that r is measured on the plane perpendicular to the duplex axis and $V_B(r, \Delta)$ now represents the volume of a cylindrical slice between $r-\Delta/2$ and $r+\Delta/2$ from the chosen central axis. In our analysis, the global z-axis was calculated by the package 3DNA, as in SCHNAaP [14]. The vectors used to define the axis are a combination of C1' and G-N9/C-N4 atoms along the same strand, as developed by Rosenberg *et al.* [15]. Calculating the cylindrical distribution function not only gives insight into the properties of the ion distribution, but also provides for a good check for the convergence of the simulation.

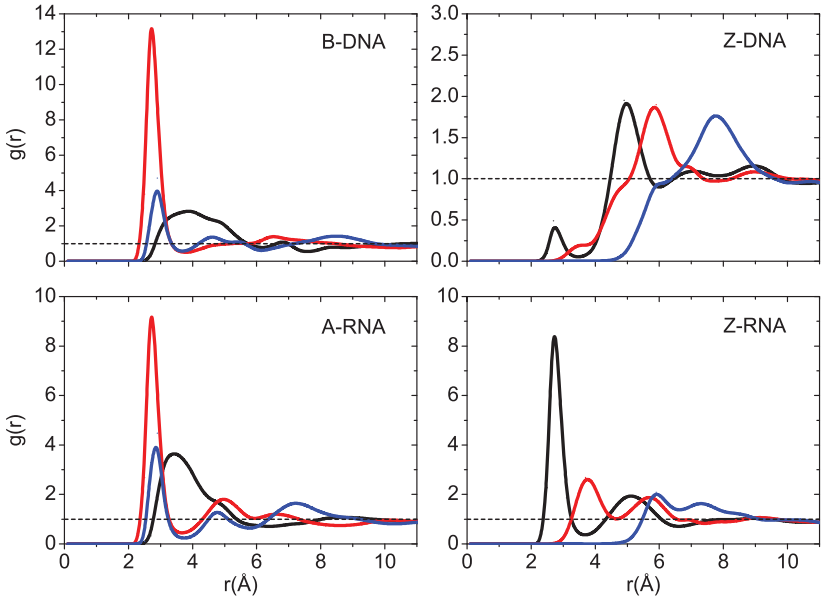
2 Results

Diffusion Constants. Table 2 summarizes the calculated diffusion coefficients for cations in the presence of the different duplexes. Generally speaking, these results are in good agreement with values obtained in B-DNA studies [16]. First, we consider the results for the Na^+ ion. At low ion concentration, the diffusion coefficients for the right-handed structures are higher than for the corresponding left-handed structures. For comparison, the calculated diffusion coefficient for 22 free Na^+ ions in TIP3P waters is $1.56 \times 10^{-9} m^2/s$, which is higher than the experimental value of $1.33 \times 10^{-9} m^2/s$ for ions in pure water [17]. The reduced values for Z-DNA/Z-RNA are indicative of a reduced mobility, due to more binding to the nucleic acids. At high salt concentrations (0.4 M), the diffusion coefficients increase for all structures, and their values are comparable to that of the free ion. Here, the number of mobile ions overwhelms the number of localized ions, giving rise to increased values of the diffusion coefficients. Table 2 also presents the diffusion coefficients for K^+ at 0.4 M KCl and for Mg^{2+} at 0.2 M $MgCl_2$. As a percentage of their bulk values (computed in TIP3P waters) the Mg^{2+} ion has the smallest value, indicating that it is in general more localized.

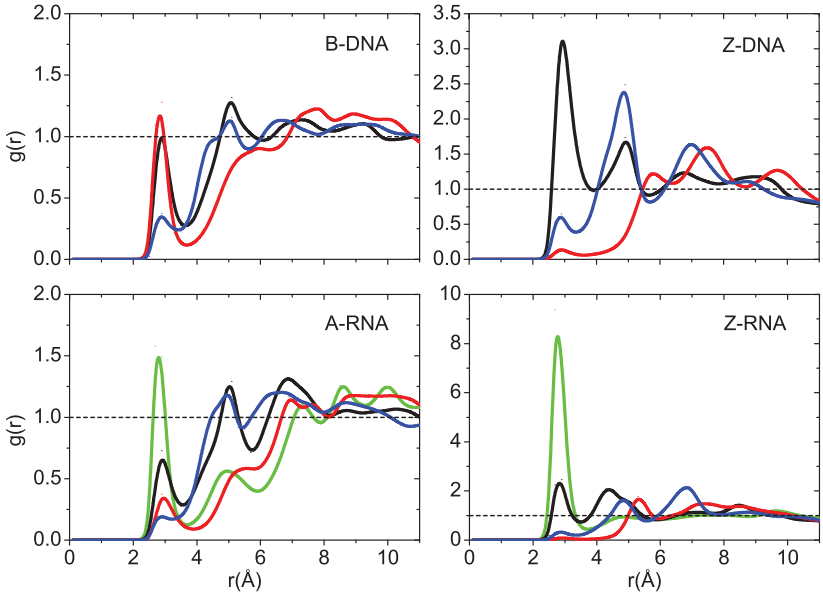
Table 2. Diffusion coefficient of cations (in units of $10^{-9}m^2/s$) for the four different nucleic acid structures. Coefficients are given for a duplex system without excess salt, and for the high excess salt cases. For comparison, the calculated diffusion coefficient in TIP3P waters is 1.56±0.14 for Na^+ ions, 2.59±0.14 for K^+ ions and 1.07±0.11 for Mg^{2+} ions.

Structure	Na^+ (no salt)	Na^+ (0.4M)	K^+ (0.4M)	Mg^{2+} (0.2M)
B-DNA	1.40±0.15	1.48±0.08	2.28±0.17	0.90±0.10
Z-DNA	0.98±0.13	1.44±0.03	2.16±0.11	0.89±0.09
A-RNA	1.28±0.19	1.45±0.10	2.26±0.13	0.92±0.10
Z-RNA	1.04±0.14	1.48±0.10	2.38±0.12	0.90±0.09

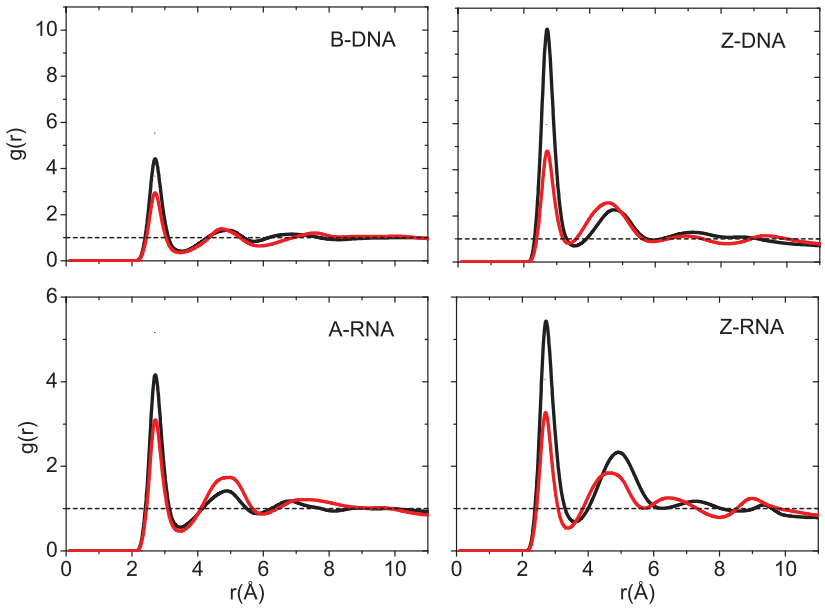
Comparison with the solution of the linear Poisson-Boltzmann equation. Over the years, a number of continuum-based theories have been proposed to model the concentration profile of ions away from charged biomolecules. Two particularly prominent examples include the so-called Manning condensation theory [18], and theories based on various versions of the Poisson-Boltzmann equation. At short distances near the duplexes, these theories are expected to fail because the atomic details matter [19]. However at larger distances, continuum theories can work. To probe this point, we considered a simple model consisting of two concentric cylinders (inner cylinder radius r=a; outer cylinder, r=b), with a constant charge density embedded on the inner cylinder and the electric field vanishing on the outer cylinder. The potential generated by such a model may be solved analytically using the linearized Poisson-Boltzmann equation [20]. The application of this model to our systems gives excellent results at larger distances. For example, SI Fig.11 shows results for 0.4 M Na^+ (radii a=12.0 Å and b=29.5 Å) and shows that this simple continuum model well represents the cylindrical ion distribution for distances larger than about 12.0 Å (RNA) and 14.0 Å (DNA).



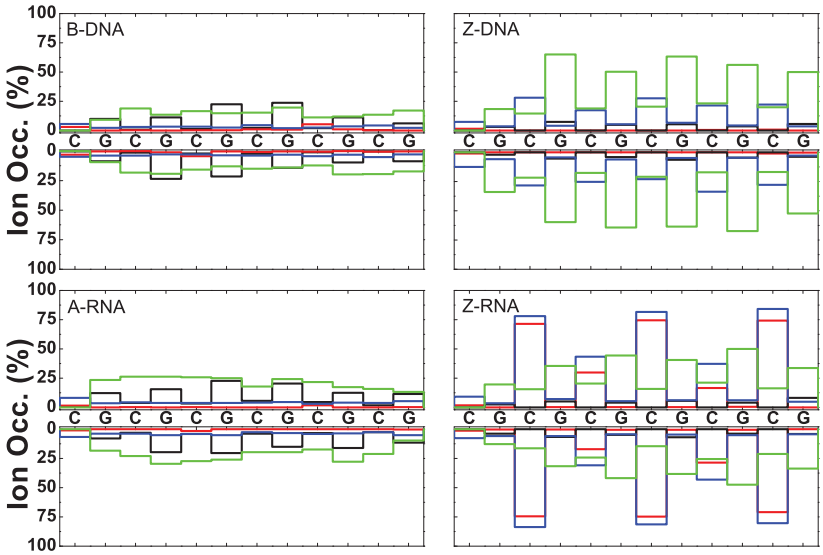
Supplementary Figure 1. RDFs for K^+ with respect to major and minor groove atoms at 0.4 M salt concentration. For B-DNA and A-RNA, the colors indicate RDFs with respect to atoms in the major groove: O6 (red), N7 (blue), N4 (black). For Z-DNA and Z-RNA, the colors indicate RDFs with respect to atoms in the minor groove: O2 (black), N2 (red) and N3 (blue).



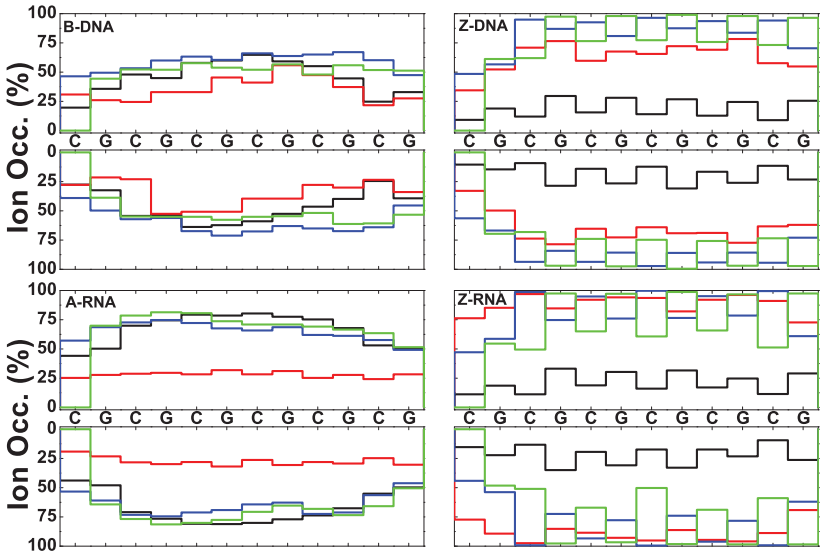
Supplementary Figure 2. RDFs for K^+ with respect to O' backbone oxygens at 0.4 M salt concentration. Colors indicate: O2' (green), O3' (black), O4' (red), and O5' (blue).



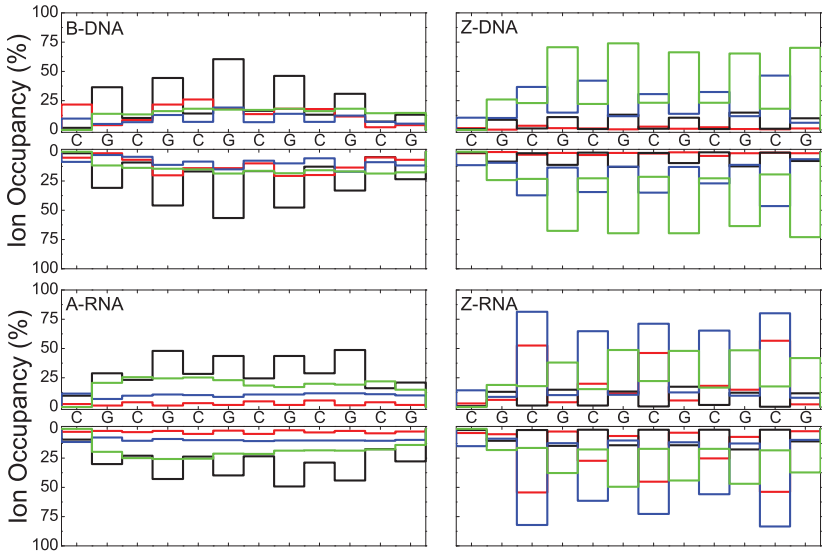
Supplementary Figure 3. RDFs for K^+ with respect to phosphate oxygens at 0.4 M salt concentration. Colors indicate: OP1 (black), OP2 (red). The maximum for the four duplexes occurs approximately at 2.8 Å.



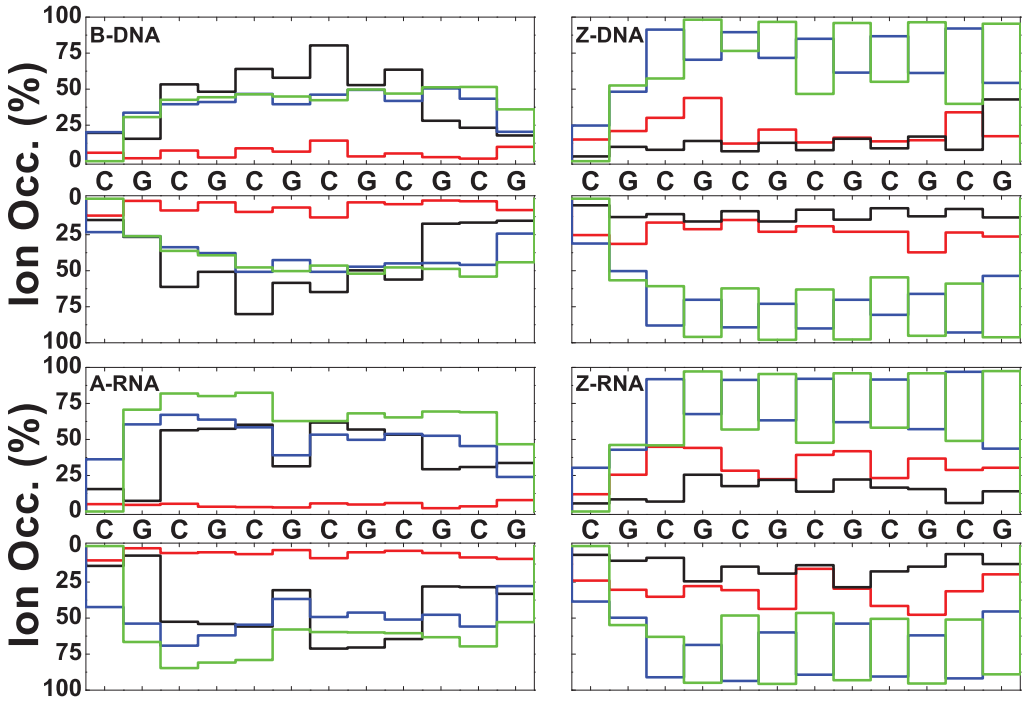
Supplementary Figure 4. Na^+ ion occupancies within 3 Å as function of sequence for 0.4M salt concentration. Colors represent: major groove (black), minor groove (red), O' oxygen atoms on backbone (blue), and phosphate oxygens (green).



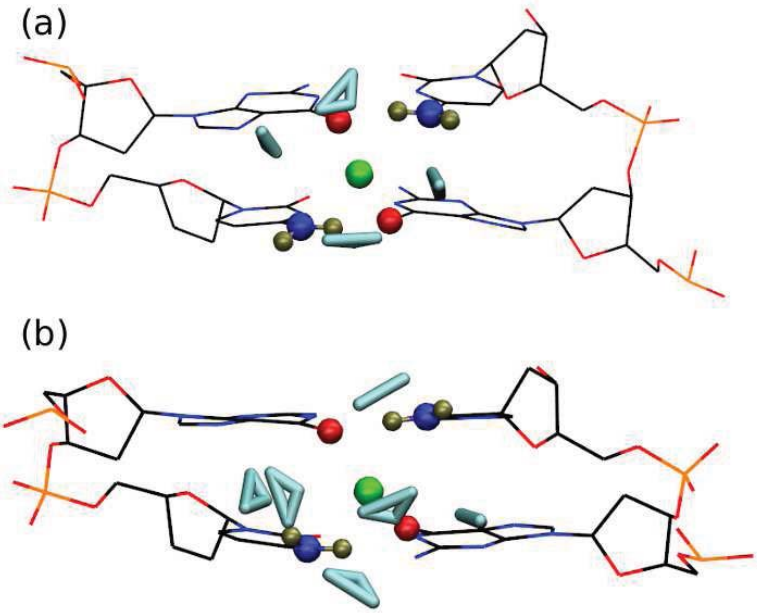
Supplementary Figure 5. Na^+ ion occupancies within 6 Å as function of sequence for 0.4M salt concentration. Colors represent: major groove (black), minor groove (red), O' oxygen atoms on backbone (blue), and phosphate oxygens (green).



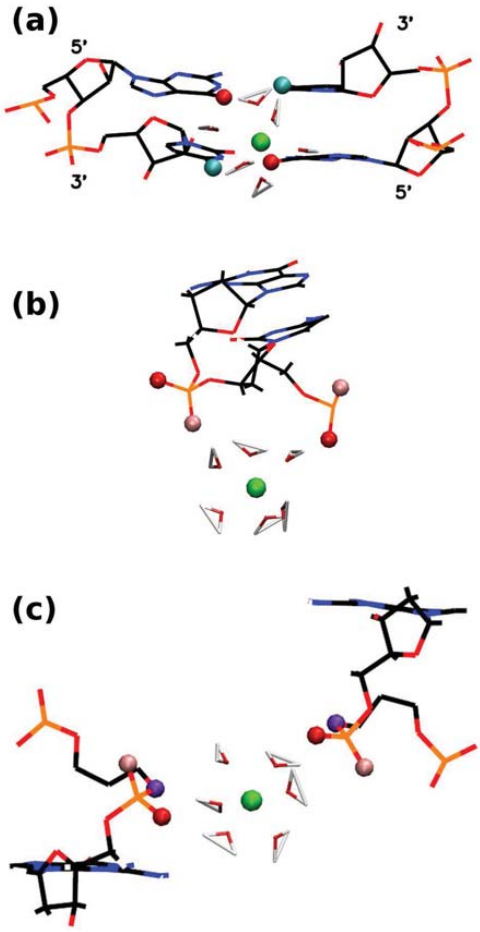
Supplementary Figure 6. K^+ ion occupancies within 3.5 Å as function of sequence for 0.4M salt concentration. Colors represent: major groove (black), minor groove (red), O' oxygen atoms on backbone (blue), and phosphate oxygens (green).



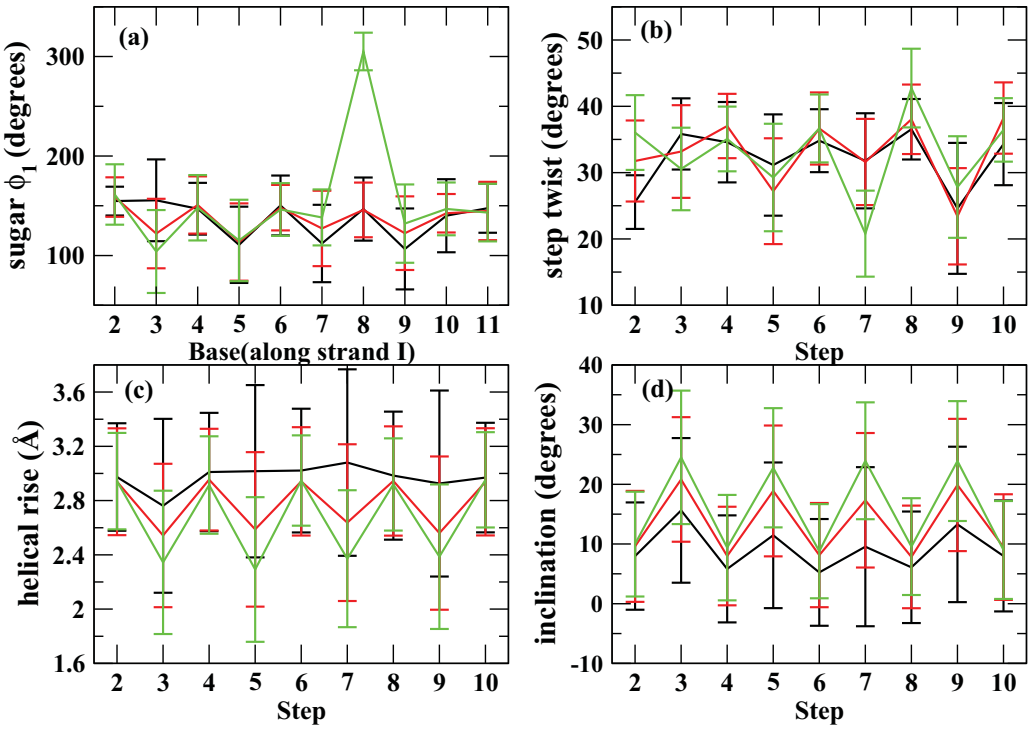
Supplementary Figure 7. $Mg^{2}+$ ion occupancies within 6 Å as function of sequence for 0.2 M $MgCl_2$ salt concentration. Colors represent: major groove (black), minor groove (red), O' oxygen atoms on backbone (blue), and phosphate oxygens (green).



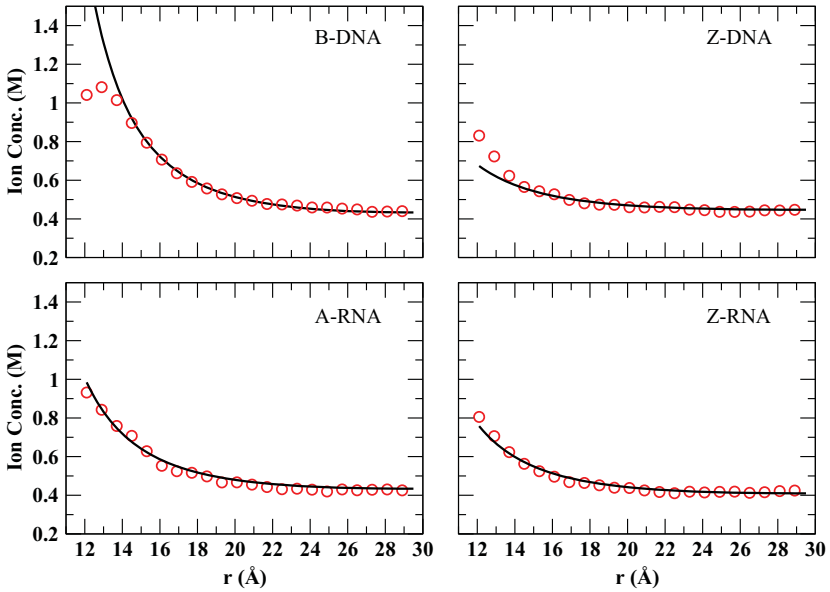
Supplementary Figure 8. Atomic details of direct binding of monoatomic ions in a "G-O6 – C-N4" configuration. These snapshots correspond to configurations similar to those in the main text, Fig. 14 a,b. Color representations: O6 (red), N4 (blue), N4-H (tan), water molecules (cyan) and ion (green). (a) Na^+ and (b) K^+, both in B-DNA at 0.4 M salt concentration. The nearby water molecules are within 3.5Å for Na^+ and 4.0Å for K^+.



Supplementary Figure 9. Atomistic details of binding sites of hexahydrated Mg^{2+} (green) for different nucleic acid structures. (a) Binding to G-O6 (red) in A-RNA (C-N4 atoms are in cyan); (b) binding to phosphate oxygens OP1 (red) and OP2 (pink) in Z-DNA; (c) binding to G-OP1 (red), G-OP2 (pink) and C-O3'(violet) in a CpG step in Z-DNA. The configurations shown are just a snapshot at a given time obtained from the MD simulations.



Supplementary Figure 10. Variation of some nucleic acid parameters for different NaCl excess salt concentrations. Zero excess salt (black), 0.4 M (red), 4 M (green). The top panels are for B-DNA and the bottom panels for A-RNA. (a) sugar phase angle ϕ_1 of B-DNA; (b) step twist of B-DNA; (c) helical rise of A-RNA; (d) helical inclination of A-RNA.



Supplementary Figure 11. Fitting results of the linear Poisson-Boltzmann theory to the cylindrical ion distribution (CDF) for Na^+ at 0.4 M. The red circles represent the simulation results, while the black curve represents the solution of the continuum theory based on the linear Poisson-Boltzamann model [20]. As may be expected, good fits are obtained away from the central axis.

References

[1] Case, D. A., Darden, T. A., Cheatham III, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B. P., Hayik, S., Roitberg, A., Seabra, G., Swails, J., , Kolossvai, A. W. G. I., Wong, K. F., Paesani, F., Vanicek, J., Wolf, R. M., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M. J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. (2012) "AMBER 12", University of California, San Francisco.

[2] Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.,* **92**, 3817–3829.

[3] Banas, P., Hollas, D., Zgarbva, M., Jurecka, P., Orozco, M., Cheatham, T. E., Sponer, J., and Otyepka, M. (2010) Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.,* **6**, 3836–3849.

[4] Zgarbova, M., Otyepka, M., Sponer, J., Mladek, A., Banas, P., Cheatham, T. E., and Jurecka, P. (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.,* **7**, 2886–2902.

[5] Jorgensen, W. L., Chandrasekhar, J., Madura, J., and Klein, M. L. (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.,* **79**, 926 – 935.

[6] Joung, I. S. and Cheatham III, 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.

[7] Aqvist, J. (1990) Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.,* **94**, 8021 – 8024.

[8] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.,* **103**, 8577 – 8593.

[9] Lu, X. J. and Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.,* **31**, 5108 – 5121.

[10] Arnott, S. and Hukins, D. W. L. (1972) Optimised parameters for A-DNA and B-DNA. *Biochem. Bioph. Res. Co.,* **47**, 1504–1509.

[11] Arnott, S., Hukins, D., Dover, S., Fuller, W., and Hodgson, A. (1973) Structures of Synthetic Polynucleotides in the A-RNA and A'-RNA Conformations. X-ray Diffraction Analyses of the Molecule Conformations of (Polyadenylic acid) and (Polyi-nosinic acid).(Polycytidylic acid).. *J. Mol. Biol.,* **81**, 109–122.

[12] Moradi, M., Babin, V., Roland, C., and Sagui, C. (2012) Reaction path ensemble of the B-Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Res.,* **41**, 33–43.

[13] Popenda, M., Milecki, J., and Adamiak, R. W. (2004) High salt solution structure of a left-handed RNA double helix. *Nucleic Acids Res.,* **32**, 4044.

[14] Lu, X. J., Hassan, M. A. E., and Hunter, C. A. (1997) Structure and Conformation of Helical Nucleic Acids: Analysis Program (SCHNAaP). *J. Mol. Biol.,* **273**, 668–680.

[15] Rosenberg, J. M., Seeman, N. C., Day, R. O., and Rich, A. (1976) RNA Double Helices Generated From Crystal-Structures of Double Helical Dinucleoside Phosphates. *Biochem. Bioph. Res. Co.,* **69**, 979–987.

[16] Várnai, P. and Zakrzewska, K. (2004) DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.,* **320**, 4269–4280.

[17] Atkins, P. W. (1998) Physical Chemistry, Oxford University Press, Oxford, .

[18] Manning, G. S. (1978) Molecular Theory of Polyelectrolyte Solutions With Applications to Electrostatic Properties of Polynucleotides. *Q. Rev. Biophys.,* **11**, 179–246.

[19] Chen, A. A., Marucho, M., Baker, N. A., and Pappu, R. V. (2009) Simulations of RNA Interactions with Monovalent Ions. *Method. Enzymol.,* **469**, 411–432.

[20] Kirmizialtin, S. and Elber, R. (2010) Computational Exploration of Mobile Ion Distributions Around RNA Duplex. *J. Phys. Chem. B,* **114**, 8207–8220.

Appendix B

Comparative melting and healing of B-DNA and Z-DNA by an infrared laser pulse - Supporting information

Supplementary Material: Comparative melting and healing of B-DNA and Z-DNA via Fast Melting by an Infrared Laser Pulse

Viet Hoang Man,[†,‡] Feng Pan,[†,‡] Celeste Sagui,[*,†,‡] and Christopher Roland[*,†,‡]

Department of Physics, North Carolina State University, Raleigh, NC 27695-8202, USA, and Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC 27695-8202, USA

E-mail: sagui@ncsu.edu; cmroland@ncsu.edu

[*]To whom correspondence should be addressed
[†]Department of Physics
[‡]CHiPS

Table S1: Bond types index. XX letter indicates heavy atoms.

Bond types index	standrad length (Å)	atom 1	atom 2	location
1	1.480	P	O1P-O2P	backbond
2	1.610	P	O5’	backbond
3	1.526	C5’	C4’	backbond
4	1.526	C4’	C3’	backbond
5	1.526	C3’	C2’	backbond
6	1.526	C1’	C2’	backbond
7	1.410	C3’	O3’	backbond
8	1.410	O4’	C1’	backbond
9	1.410	C4’	O4’	backbond
10	1.410	O5’	C5’	backbond
11	1.610	O3’	P	backbond
12	1.475	C1’	N1	Join C
13	1.475	C1’	N9	Join G
14	1.383	N1	C2	C base
15	1.229	C2	O2	C base
16	1.358	N3	C2	C base
17	1.339	C4	N3	C base
18	1.340	C4	N4	C base
19	1.433	C5	C4	C base
20	1.350	C6	C5	C base
21	1.365	N1	C6	C base
22	1.381	N1	C2	G base
23	1.340	C2	N2	G base

Continued on next page

Table S1 – *Continued from previous page*

Bond types index	standard length (Å)	atom 1	atom 2	location
24	1.339	C2	N3	G base
25	1.354	N3	C4	G base
26	1.370	C5	C4	G base
27	1.419	C5	C6	G base
28	1.229	C6	O6	G base
29	1.388	C6	N1	G base
30	1.374	N9	C4	G base
31	1.391	N7	C5	G base
32	1.304	C8	N7	G base
33	1.371	N9	C8	G base
34	0.96-1.09	H	XX	all

Table S2: native Hbonds index.

native Hbonds index	acceptor group		donor group	
	atom	base index	atom	base index
1	O2	1	N2(H21/H22)	24
2	N3	1	N1(H1)	24
3	O6	24	N4(H41/H42)	1
4	O2	23	N2(H21/H22)	2
5	N3	23	N1(H1)	2
6	O6	2	N4(H41/H42)	23
7	O2	3	N2(H21/H22)	22
8	N3	3	N1(H1)	22
9	O6	22	N4(H41/H42)	3
10	O2	21	N2(H21/H22)	4
11	N3	21	N1(H1)	4
12	O6	4	N4(H41/H42)	21
13	O2	5	N2(H21/H22)	20
14	N3	5	N1(H1)	20
15	O6	20	N4(H41/H42)	5
16	O2	19	N2(H21/H22)	6
17	N3	19	N1(H1)	6
18	O6	6	N4(H41/H42)	19
19	O2	7	N2(H21/H22)	18
20	N3	7	N1(H1)	18
21	O6	18	N4(H41/H42)	7
22	O2	17	N2(H21/H22)	8

Continued on next page

Table S2 – *Continued from previous page*

native Hbonds index	acceptor group		donor group	
	atom	base index	atom	base index
23	N3	17	N1(H1)	8
24	O6	8	N4(H41/H42)	17
25	O2	9	N2(H21/H22)	16
26	N3	9	N1(H1)	16
27	O6	16	N4(H41/H42)	9
28	O2	15	N2(H21/H22)	10
29	N3	15	N1(H1)	10
30	O6	10	N4(H41/H42)	15
31	O2	11	N2(H21/H22)	14
32	N3	11	N1(H1)	14
32	O6	14	N4(H41/H42)	11
34	O2	13	N2(H21/H22)	12
35	N3	13	N1(H1)	12
36	O6	12	N4(H41/H42)	13

Equilibrium quantities

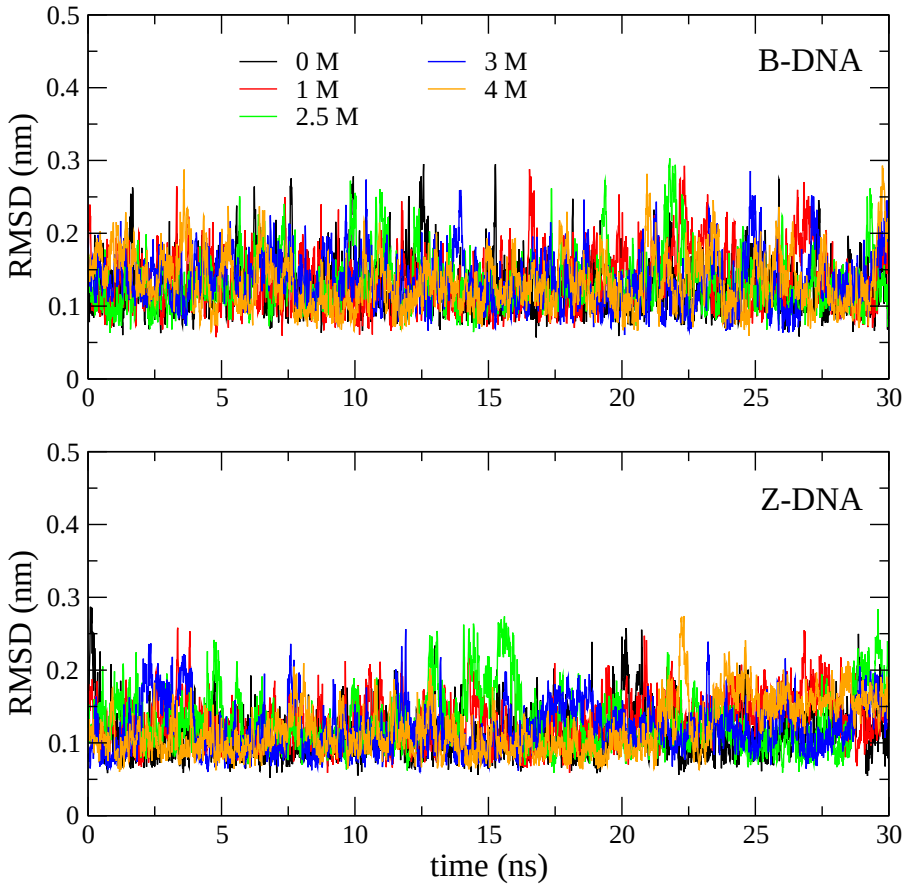


Fig. S1: **Time dependence of RMSD of B-DNA (top) and Z-DNA (bottom).** The duplexes are in equilibrium with no applied field. The labels denote NaCl salt concentration at 0, 1, 2.5, 3 and 4 M, respectively. RMSDs are taken with respect to average structures.

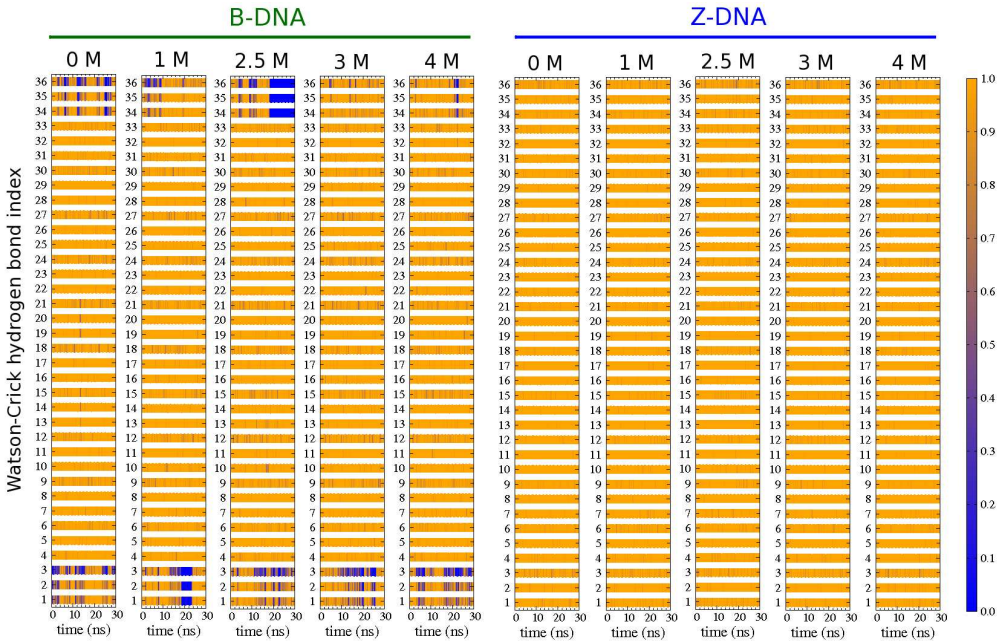


Fig. S2: **Time dependence of the native Watson-Crick hydrogen bond probability.** The labels denote NaCl salt concentration at 0, 1, 2.5, 3 and 4 M, respectively. White gaps separate the different H-bonds. H-bonds are numbered along the residue indices (see Table S2). Terminal H-bonds therefore are 1, 2, 3 corresponding to C_1–G_{24}, and 34, 35, 36 corresponding to C_{13}–G_{12}.

Non-equilibrium quantities

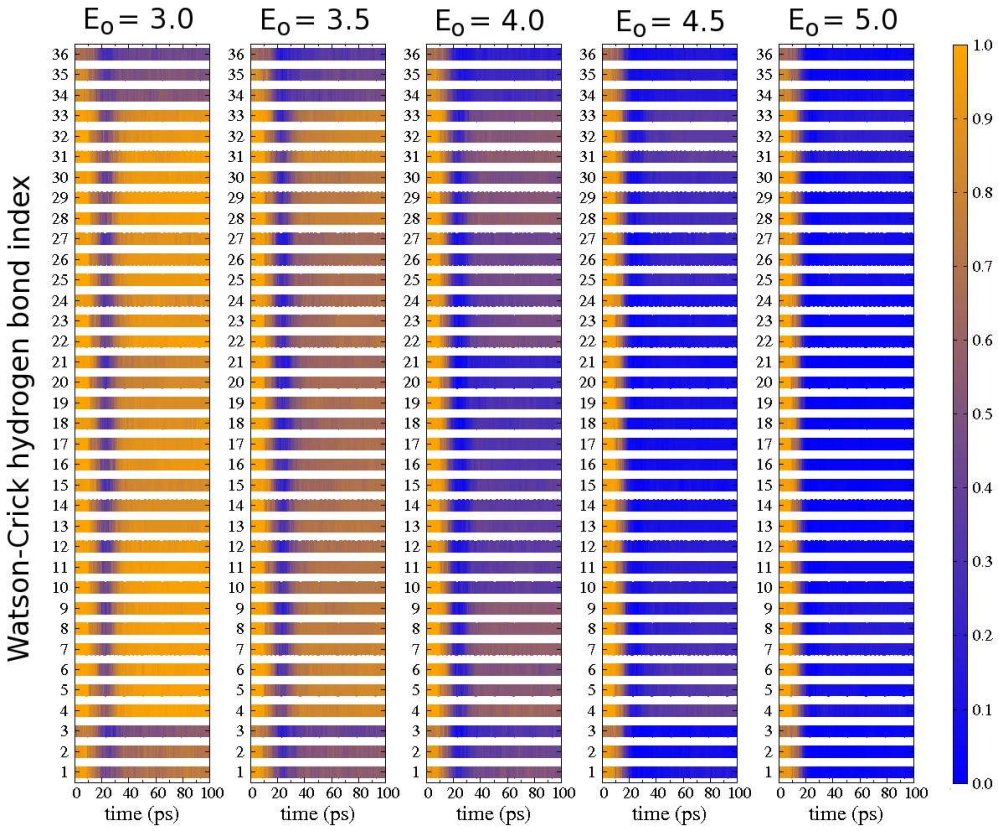


Fig. S3: **Time dependence of the native Watson-Crick hydrogen bond probability in B-DNA at zero salt excess.** White gaps separate the different H-bonds. The unit of E_0 is V/nm. The probability is computed over 50 trajectories. The laser pulse parameters, t_0 = 20 ps and σ = 6 ps.

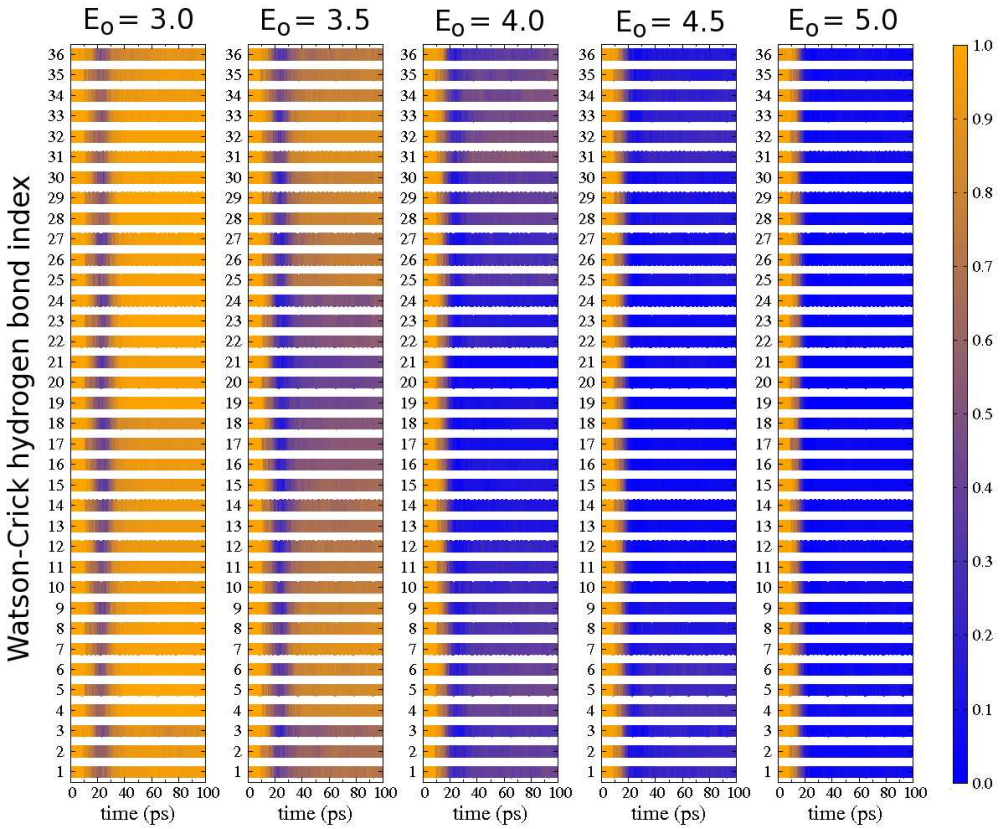


Fig. S4: **Time dependence of the native Watson-Crick hydrogen bond probability in Z-DNA at zero salt excess.** White gaps separate the different H-bonds. The unit of E_0 is V/nm. The probability is computed over 50 trajectories. The laser pulse parameters, t_0 = 20 ps and σ = 6 ps.

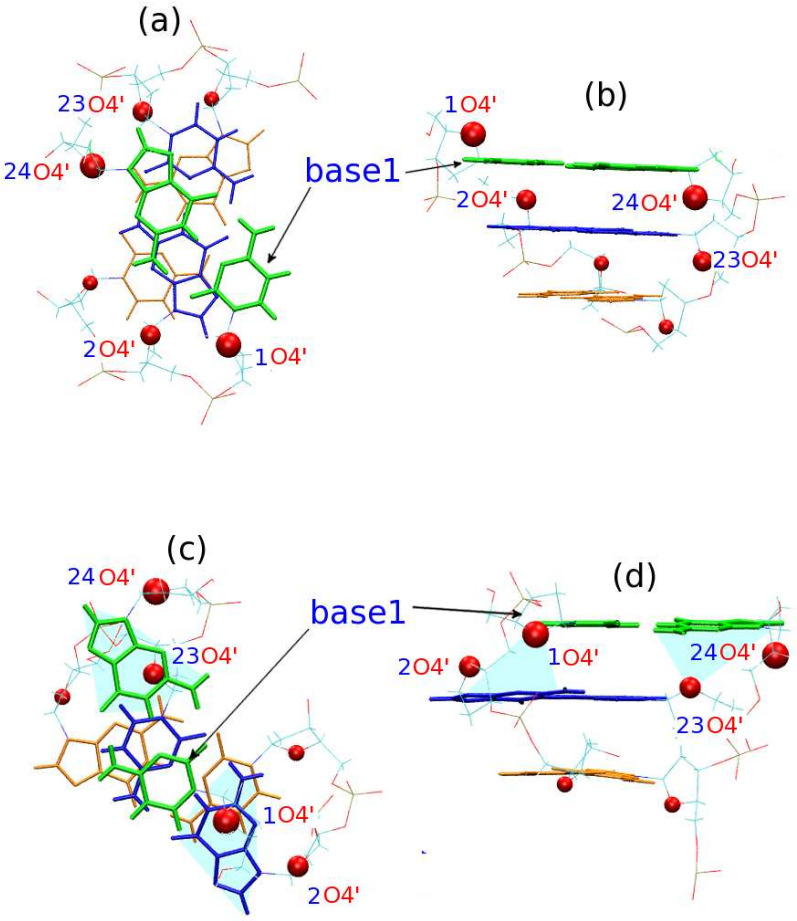


Fig. S5: **Terminal base-pairs in B-DNA and Z-DNA strands.** B-DNA is shown in (a,b) while Z-DNA is shown in (c,d). Parallel view to the helical axes are shown in (a) and (c), and perpendicular views in (b) and (d). The 1-24, 2-23 and 3-22 base-pairs are shown in green, blue and orange sticks, respectively. The red balls indicate the O4' atoms labeled by iO4', in which i is the base index. The light of 1O4' to base2 and 23O4' to base24 are marked in light cyan color. The difference between the terminal pairs of B-DNA and Z-DNA is that the projection of the 1O4' (23O4') atom on the base2 (base24) ring plane is inside the base2 (base24) ring area in Z-DNA, which does not exist in B-DNA. This difference contributes to the larger fraying of B-DNA with respect to Z-DNA.

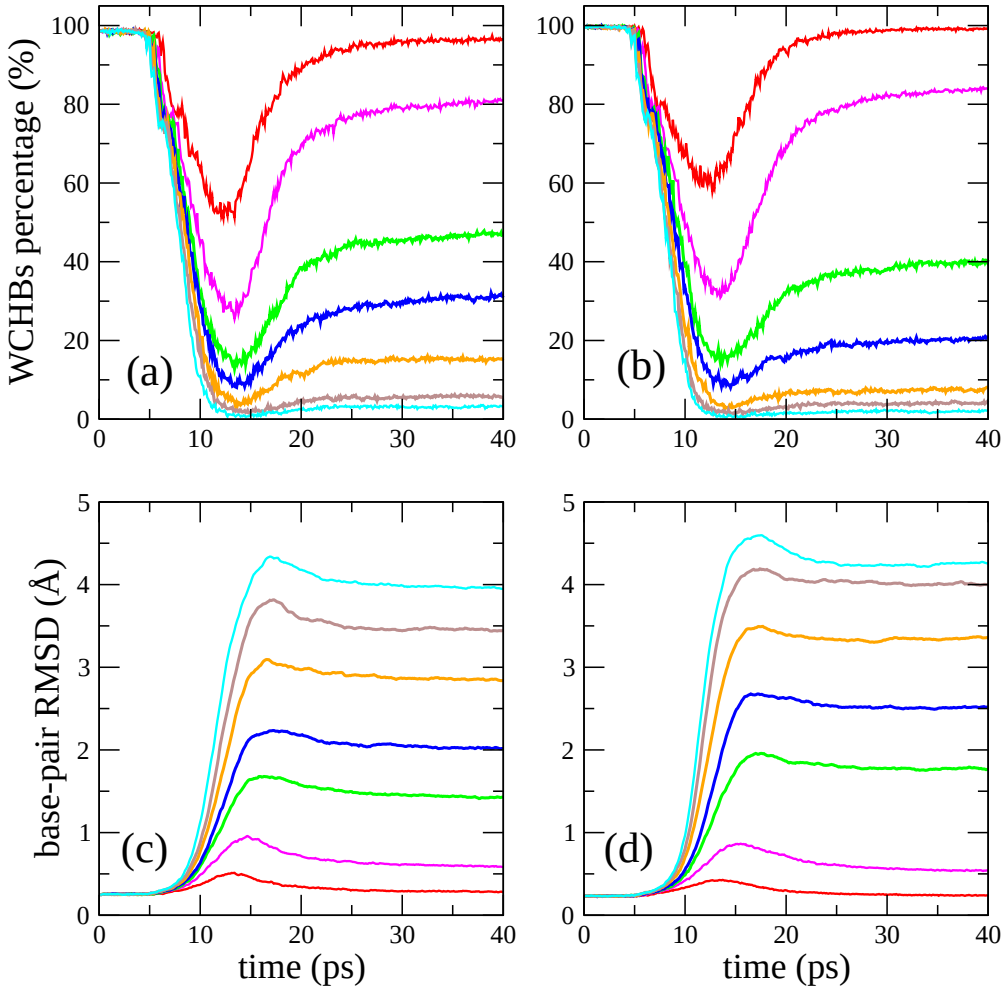


Fig. S6: **Time evolution of the Watson-Crick H-bond (WCHB) percentage and base-pair RMSD of B-DNA and Z-DNA under the application of a laser pulse at zero excess salt.** B-DNA is shown in (a) and (b), and Z-DNA in (c) and (d). Here, different colors represent different values of the electric field amplitude E_o: 3.0 (red); 4.0 (magenta); 4.6 (green); 5.0 (blue); 5.5 (orange); 6.0 (brown); and 6.5 (cyan), respectively. The base-pair RMSD is taken with respect to the initial structure and represents an average over the ten middle base pairs of each DNA sequence. Data is averaged over 50 trajectories and E_0 is given in V/nm. The laser pulse parameters, t_0 = 10 ps and σ = 3 ps.

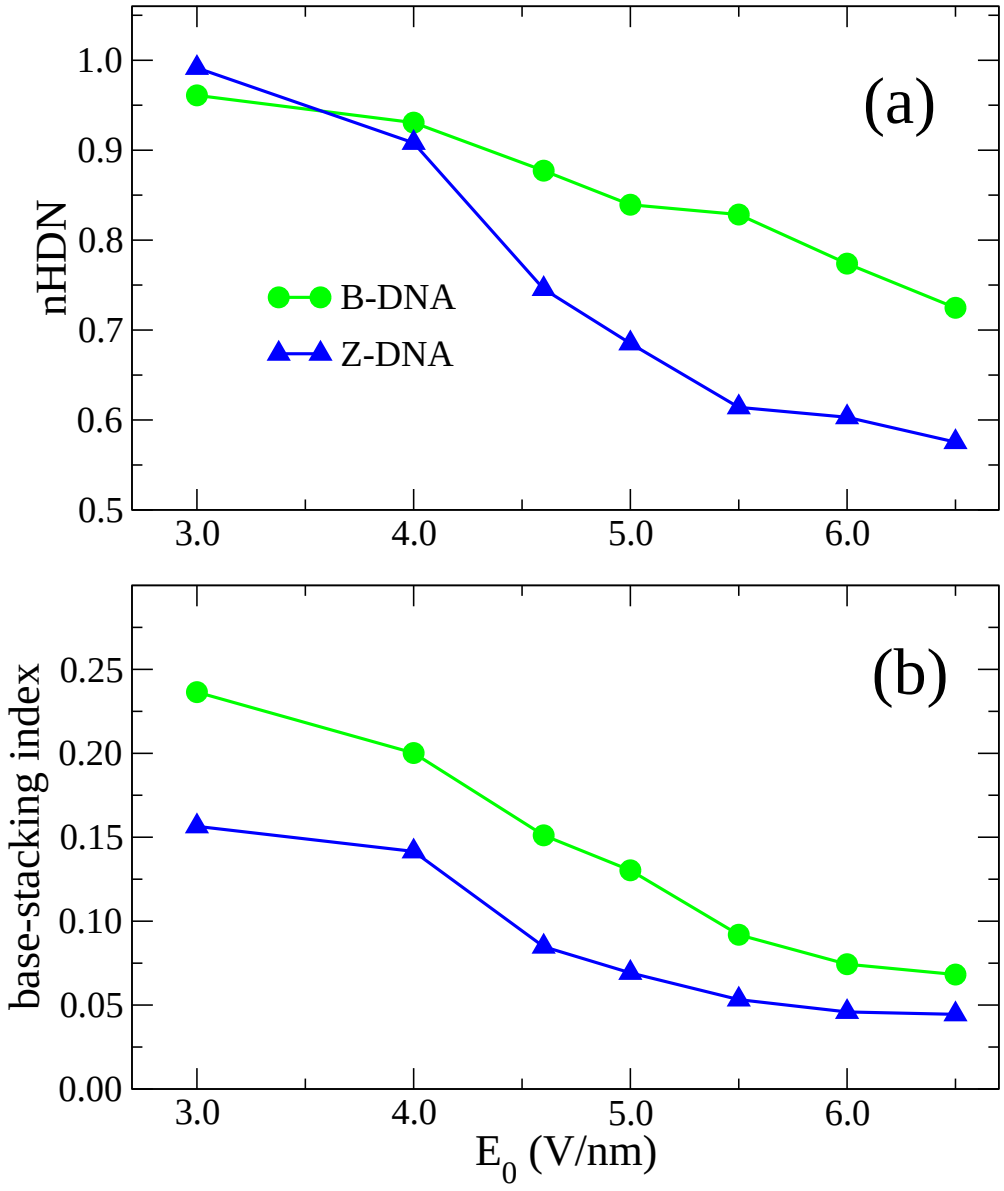


Fig. S7: **Dependence of the normalized handedness and base-stacking index on the magnitude of the electric field E_0.** Data points are averaged over the last 5 ps of the 40 ps simulations with zero excess salt. The laser pulse parameters, t_0 = 10 ps and σ = 3 ps.

Appendix C

Structure and Dynamics of DNA and RNA Double Helices of CAG and GAC Trinucleotide Repeats - Supporting information

Structure and dynamics of DNA and RNA double helices obtained from the CAG and GAC trinucleotide repeats [Supporting Material]

Feng Pan, Viet Hoang Man, Christopher Roland, Celeste Sagui*

Department of Physics, North Carolina State University, Raleigh, NC 27695-8202, USA

* E-mail: sagui@ncsu.edu

1 Definitions of twist and handedness

We use the 3DNA software package(1, 2) to calculate the twist angle of the duplexes. Since non-Watson-Crick base pairs are our main object, we choose the "simple" step parameters, which are "intuitive" for non-Watson-Crick base pairs and were introduced into 3DNA as of v2.3-2016jan01. The regular z-axis defined in 3DNA is used here, which is the average of two base normals, taking into consideration the M-N vs M+N base-pair classification. The "simple" inter-base-pair step parameter calculation uses consecutive C1'-C1' vectors. Since the A-A mismatches may have a anti-syn or syn-syn conformation, the z-axis turns out opposite to the normal one. Thus, one must add 180 degrees to the twist angles involving A(anti)-A(syn) and A(syn)-A(syn) mismatches.

Handedness is a natural choice for investigating left- (H $<$ 0) and right-handed (H $>$ 0)) helical structures, based on a former investigation of the B-Z DNA transition(3). For the double helix, the position of the phosphorus (P) atoms of the backbone phosphate groups was found to be a good choice for the definition of handedness. In brief, the definition of handedness for a portion of DNA/RNA between the base pairs n and m makes use of a sequence of P atoms: P^1_n,P^2_n,P^1_{n+1},P^2_{n+1},...,P^1_m,P^2_m, where the upper index indicates the strand number (1 or 2, labeled arbitrarily) and the lower index indicates the base-pair number labeled in the 5 $\rightarrow$ 3 direction of strand 1. Note that this definition of handedness is independent of the labeling of the strands. Supplementary Fig. S1 (right) shows the P atoms involved in the definition of handedness of a DNA segment between base pairs n and m; the red and purple balls in this figure are the first and last elements in the sequence. The position of these P atoms then defines the handedness via

$$H(p_1p_2p_3...p_n) = \frac{\overrightarrow{AB}}{|\overrightarrow{AB}|} \times \frac{\overrightarrow{CD}}{|\overrightarrow{CD}|} \cdot \frac{\overrightarrow{EF}}{|\overrightarrow{EF}|}, \tag{1}$$

in which each p_i is a point in the sequence discussed above, and

$$H(ABCD) = \sum_{i=1}^{n-3} H(p_ip_{i+1}p_{i+2}p_{i+3}). \tag{2}$$

In this last equation, the points A, B, C, D define the vectors $\overrightarrow{AB}$ and $\overrightarrow{CD}$ and the midpoints of these vectors, called E and F , in turn form the vector $\overrightarrow{EF}$. Supplementary Fig. S1 illustrates this definition for the first term of the sum in the relation (Eq. (1)). The cross product of the unit vectors of $\overrightarrow{AB}$ and

$\overrightarrow{CD}$ defines the (purple) vector whose dot product with the unit vector of $\overrightarrow{EF}$ forms the first term of the sum in the definition of handedness.

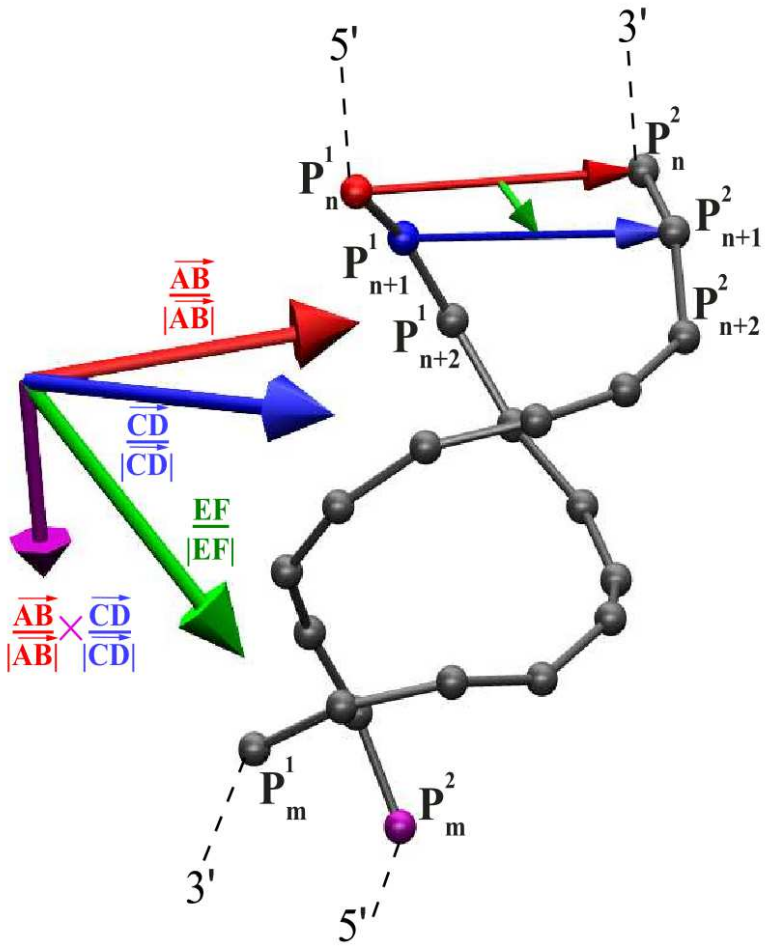


Figure S1: **Definition of handedness**

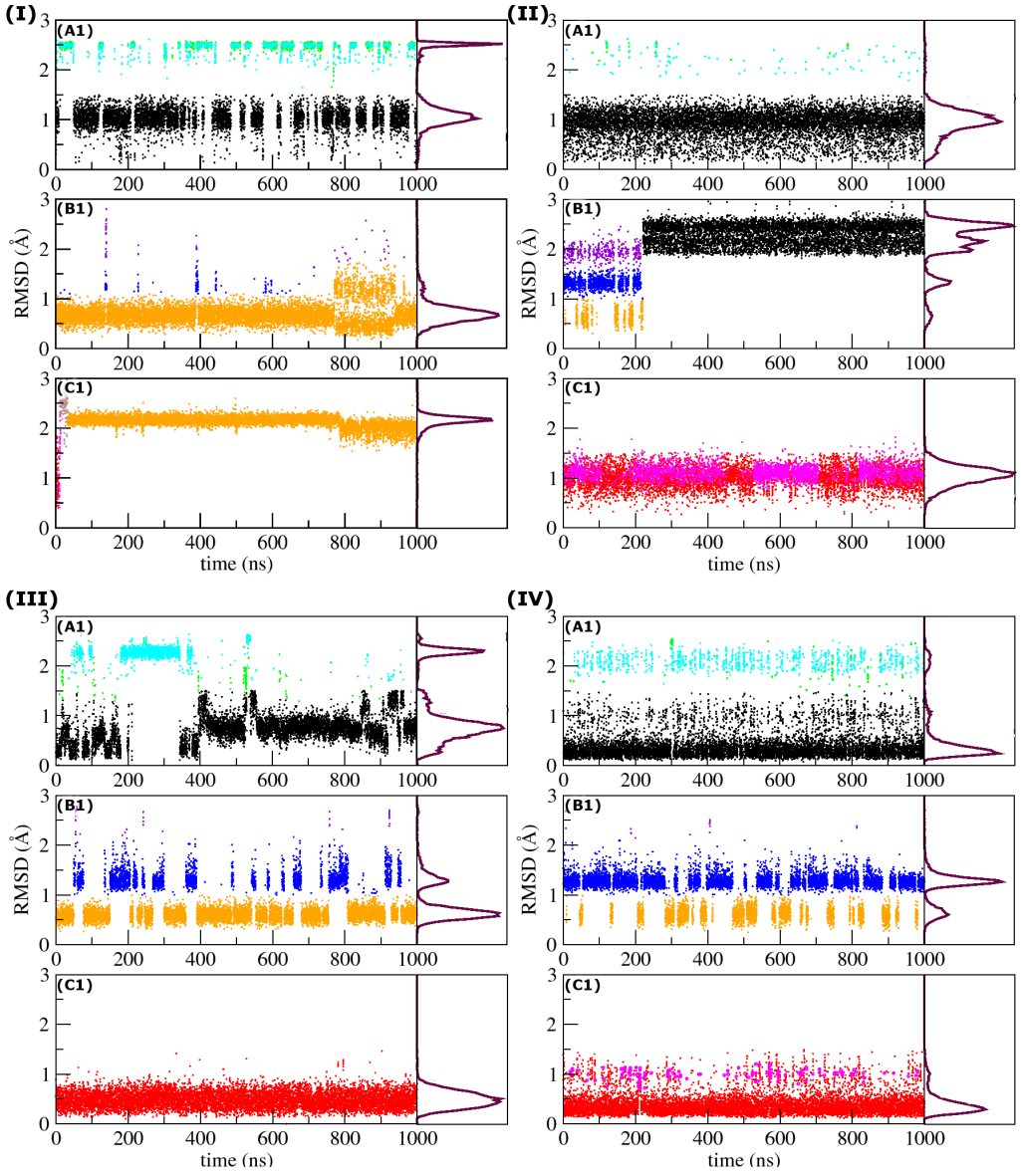


Figure S2: **RMSD for the single internal mismatch A_5-A_{14} during 1 μs simulations.** (I) RNA-CAG (II) RNA-GAC (III) DNA-CAG (IV) DNA-GAC Conformations are color coded to agree with the mismatch conformations in Fig. 2. Initial conformations for each of the three panels in a column are as follows. Top: anti-anti (conformation A1 in Fig. 2); Middle: anti-syn (B1); Bottom: syn-syn (C1).

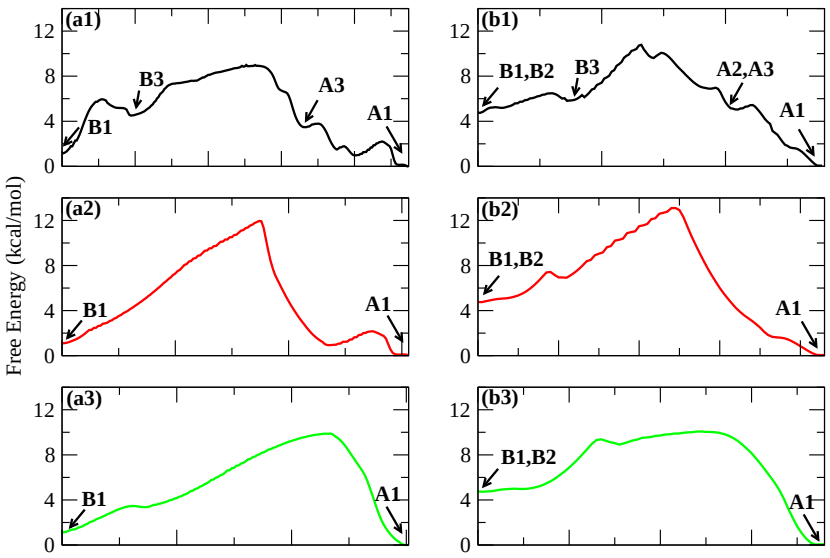


Figure S3: **Possible paths for the B1 → A1 transition on the (Ω,χ_{14}) free energy maps.** (a) RNA-CAG; (b) RNA-GAC. Different colors indicate different paths on the (Ω,χ_{14}) maps.

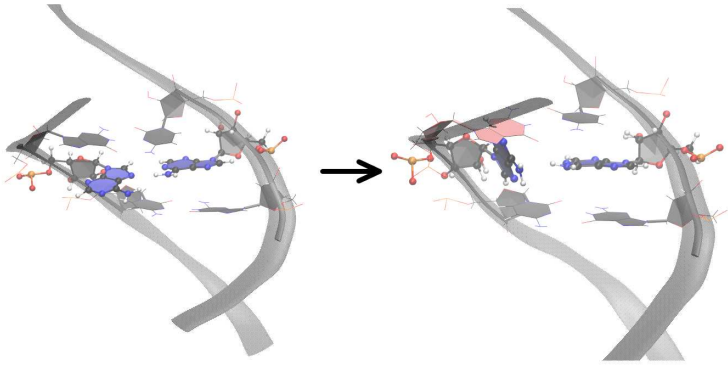


Figure S4: **syn→ anti rotation in a clockwise direction.** This rotation results in a clash betwen A14 and the pink G13.

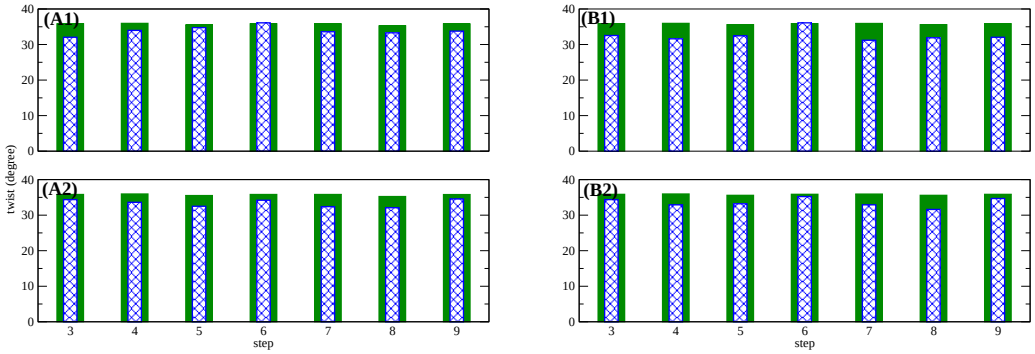


Figure S5: **Simple twist of the middle eight basepairs in DNA with initial anti-syn and syn-syn mismatch conformations.** (A1) anti-syn $(CAG)_4$; (A2) syn-syn $(CAG)_4$; (B1) anti-syn $(GAC)_4$; (B2) syn-syn $(GAC)_4$. Green bars show the initial values. Blue bars show the average value taken from the final 200ns.

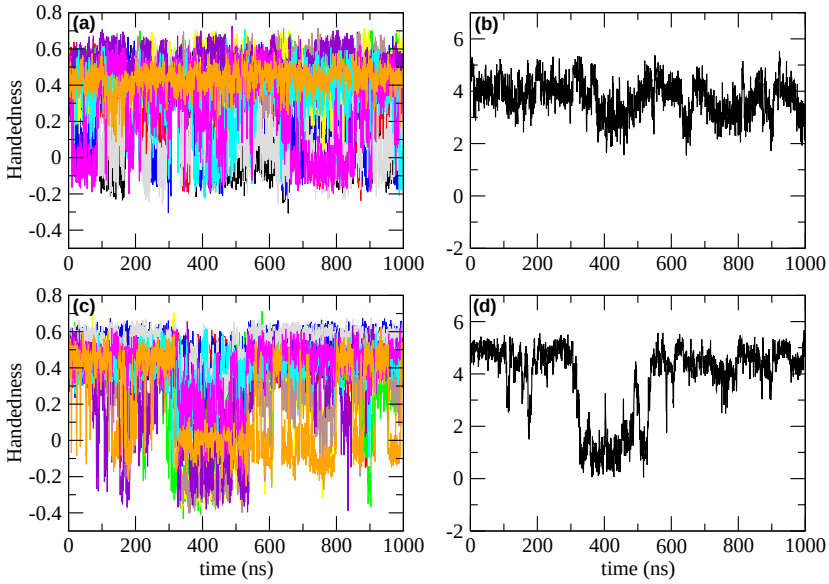


Figure S6: **Handedness of the middle six basepairs in DNA with initial anti-anti mismatch conformations.** Top: $(CAG)_4$; Bottom: $(GAC)_4$. The left column shows local handedness, with different colors representing different turns. The right column shows the total handedness.

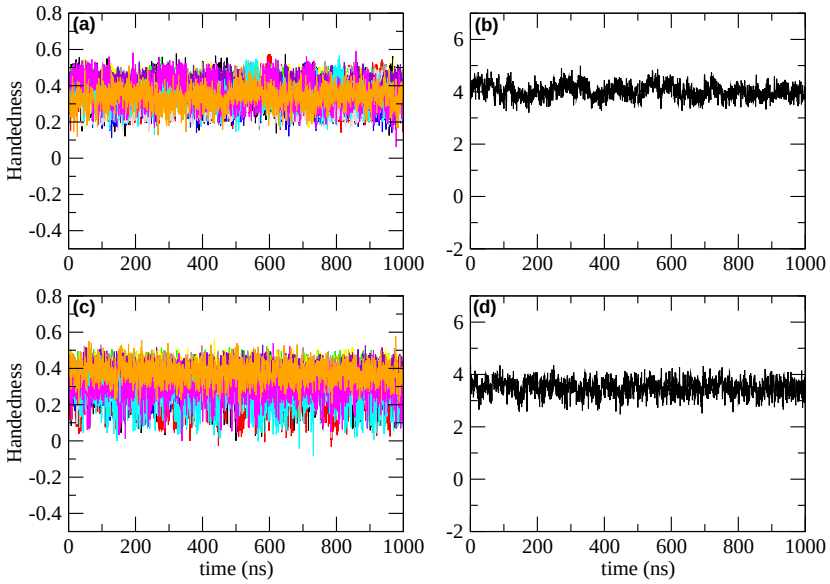


Figure S7: **Handedness of the middle six basepairs in RNA with initial anti-anti mismatch conformations.** Top: $(CAG)_4$; Bottom: $(GAC)_4$. The left column shows local handedness, with different colors representing different turns. The right column shows the total handedness.

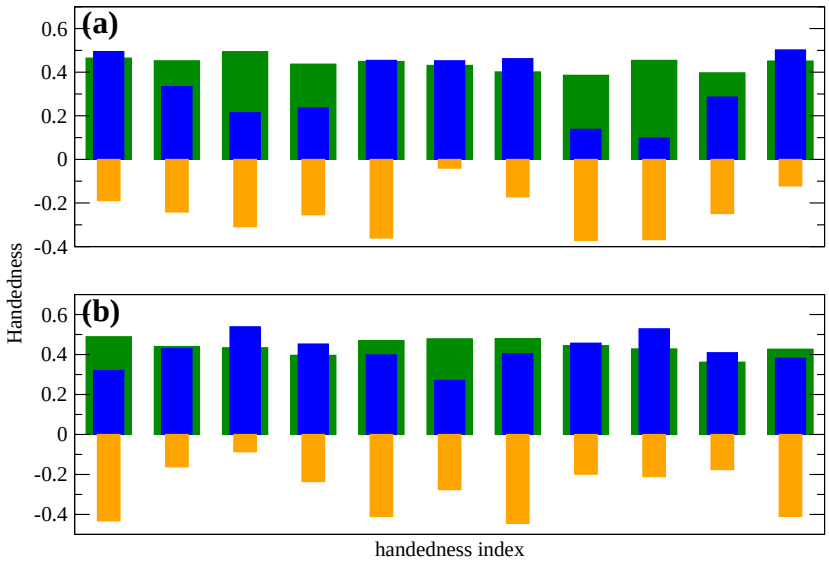


Figure S8: **Average local handedness of the middle six basepairs in DNA with initial anti-anti mismatch conformations.** (a) $(CAG)_4$; (b) $(GAC)_4$. Green bars show initial values taken from the first 10ns. Blue bars show average values taken from the final 200ns. Orange bars show the minimum value throughout the run. The index is specified by the P atoms of different residues, and acts as a sliding window through successive four residues. Thus, the first index is defined by $P_3P_{16}P_4P_{15}$ using Eq. (1) where the lower number represents residue index, the second is defined by $P_{16}P_4P_{15}P_5$, etc.

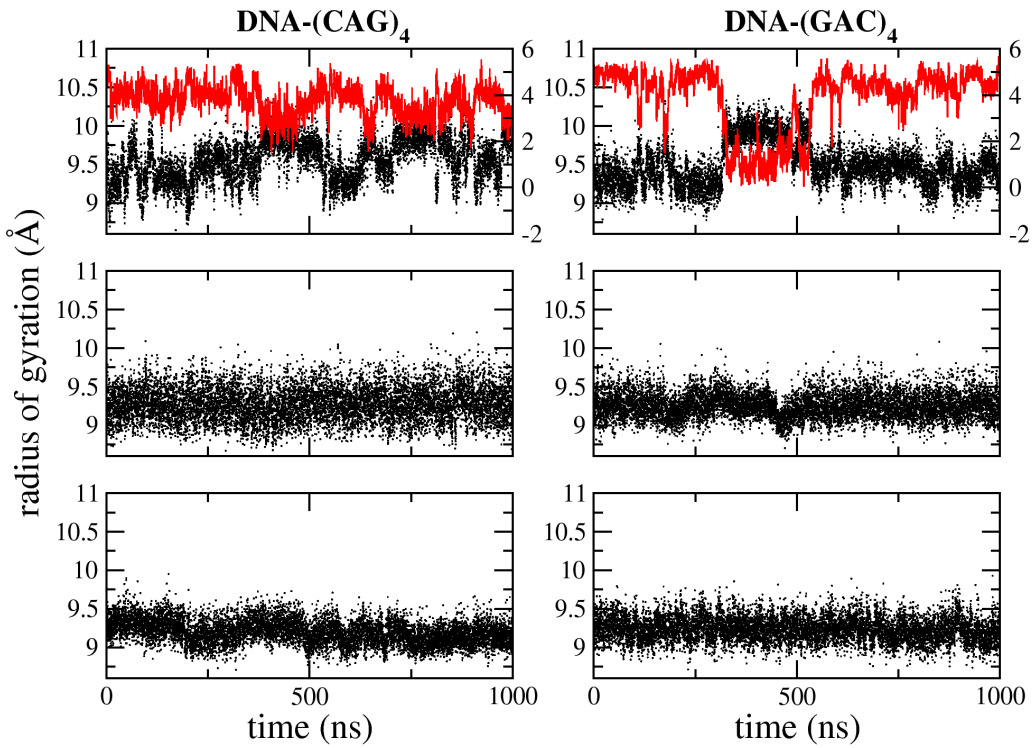


Figure S9: **Radius of gyration during 1 μs simulations** Considered here are the residues 4-9 on one strand and the complementary residues 16-21 on the other. Left column: DNA-$(CAG)_4$; Right column: DNA-$(GAC)_4$. Top: anti-anti; Middle: anti-syn; Bottom: syn-syn. The red lines show the total handedness as comparison, with its scale shown on the right side.

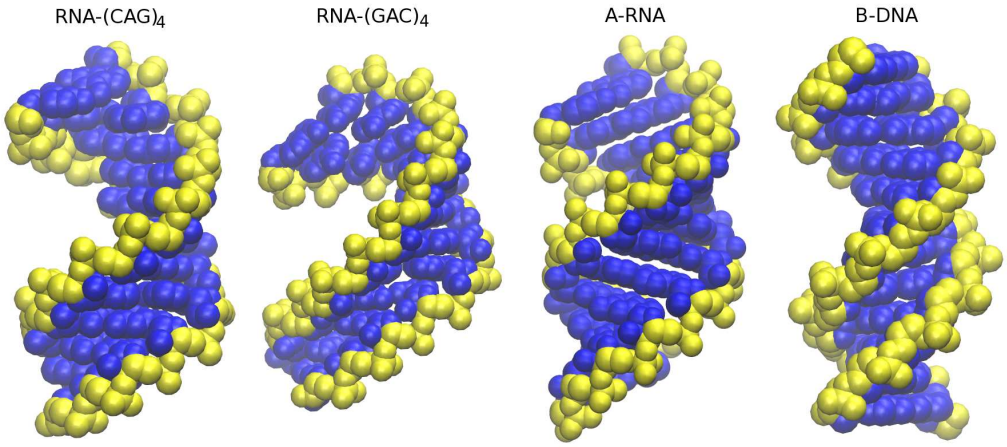


Figure S10: **Comparison between RNA-$(CAG)_4$, RNA-$(GAC)_4$ and standard B-DNA, A-RNA helices in ball model.** The RNA-$(CAG)_4$ and RNA-$(GAC)_4$ structures are determined by choosing the lowest combined RMSD value of the middle two AA mismatches, with respect to the anti-anti minimum A1.

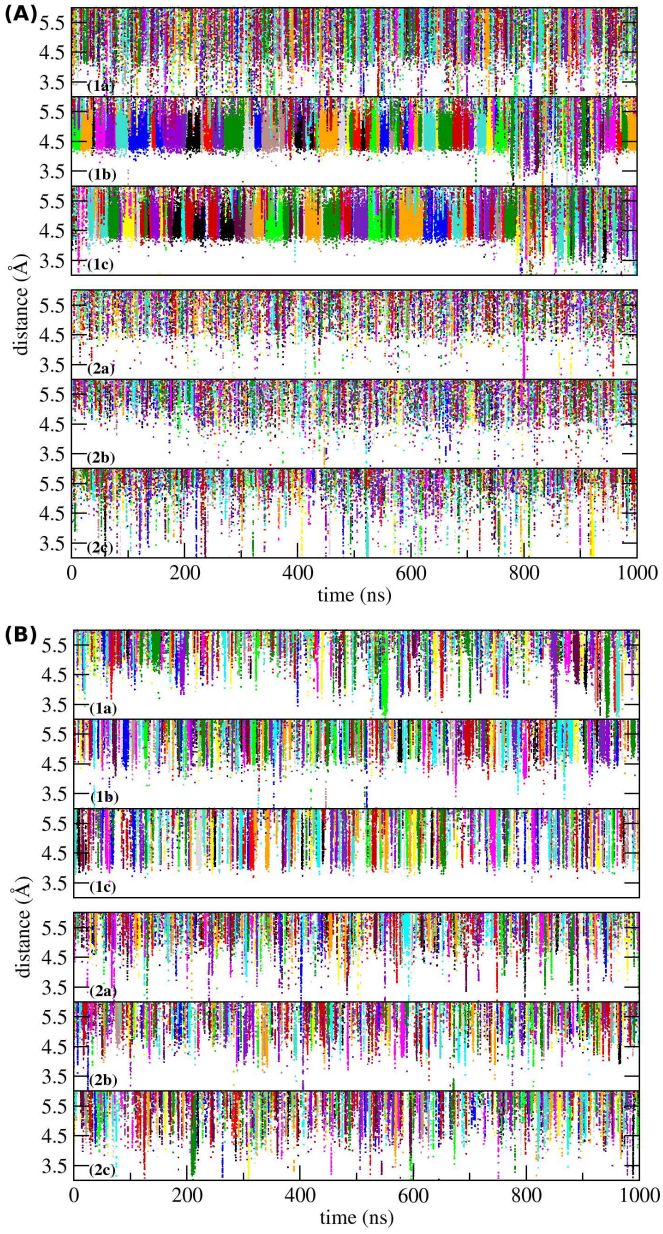


Figure S11: **Distance between Na^+ ions and the center of mass of the A-A mismatches.** The single mismatch duplexes are: (A) RNA-CAG (top) and RNA-GAC (bottom); (B) DNA-CAG (top) and DNA-GAC (bottom). For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

	RNA-$(CAG)_4$(anti-anti)	RNA-$(CAG)_4$(anti-syn)	RNA-$(GAC)_4$(anti-anti)	RNA-$(GAC)_4$(anti-syn)	A-RNA	B-DNA
major groove width (Å) (direct PP distance)	23.1±1.8	16.9±1.2	24.8±1.9	18.8±3.1	15.2	17.3
minor groove width (Å) (direct PP distance)	16.6±0.6	18.1±0.5	16.3±0.6	19.0±1.0	18.8	11.5
inclination (degree)	5.6±4.5	14.4±3.7	4.6±5.0	16.3±4.6	19.0	-5.5

Table S1: **Major/minor groove width and basepair inclination for RNA-$(CAG)_4$, RNA-$(GAC)_4$ and standard B-DNA, A-RNA.** The results of RNA-$(CAG)_4$ and RNA-$(GAC)_4$ are taken from the middle five base pairs and averaged through 50ns.

Supporting References

[1] Lu, X. J., and W. K. Olson, 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 31:5108 – 5121.

[2] Lu, X., and W. K. Olson, 2008. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* 3:1213–1227.

[3] Moradi, M., V. Babin, C. Roland, and C. Sagui, 2012. Reaction path ensemble of the B-Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Res.* 41:33–43.

Appendix D

Structure and dynamics of DNA and RNA double helices obtained from the CCG and GGC trinucleotide repeats - Supporting information

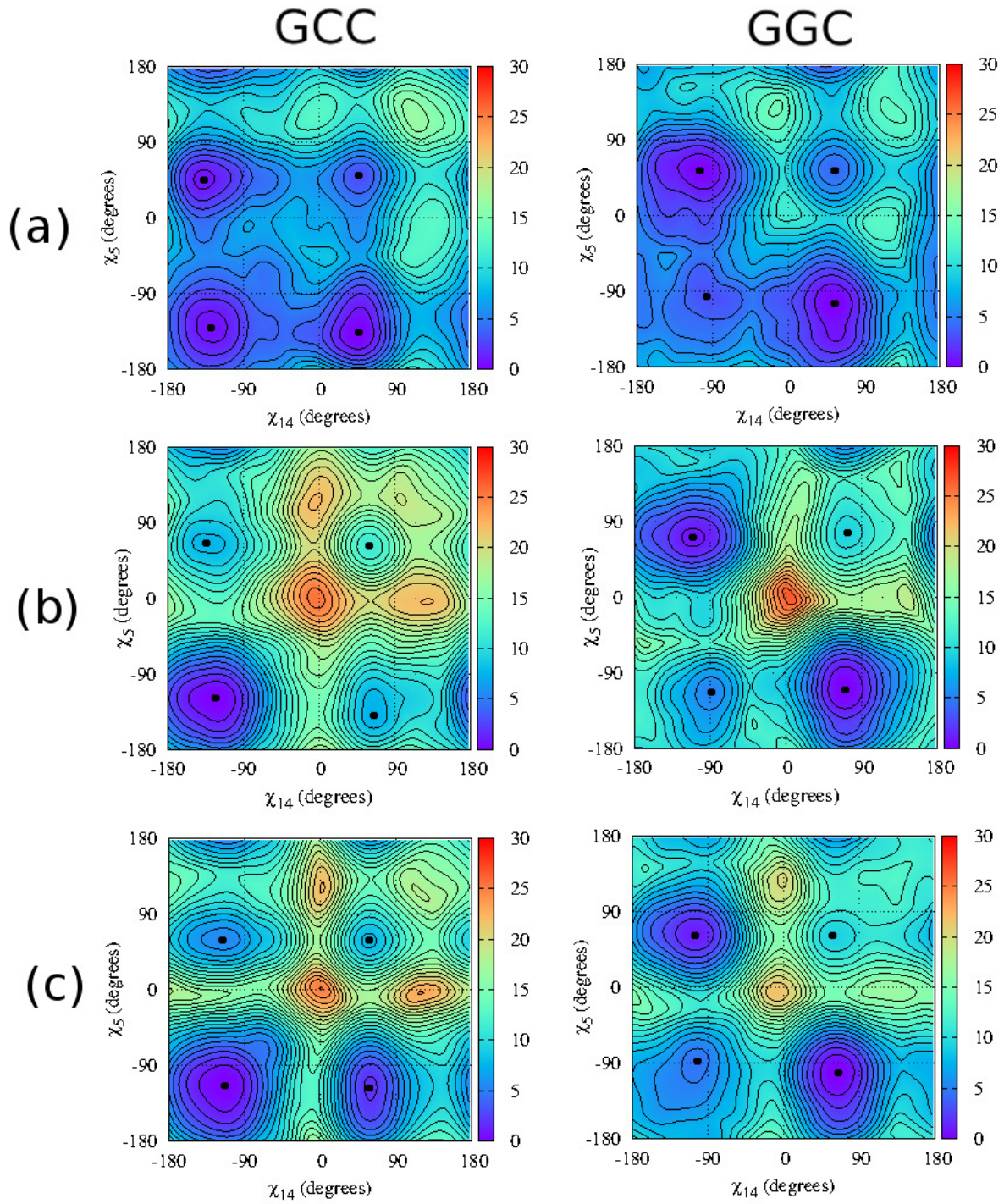


Figure D.0 (χ_5, χ_{14}) free energy maps for single mismatches in DNA-GCC (left panels), DNA-GGC (right panels) for BSC0(a), BSC1(b) and OL15(c) force fields.

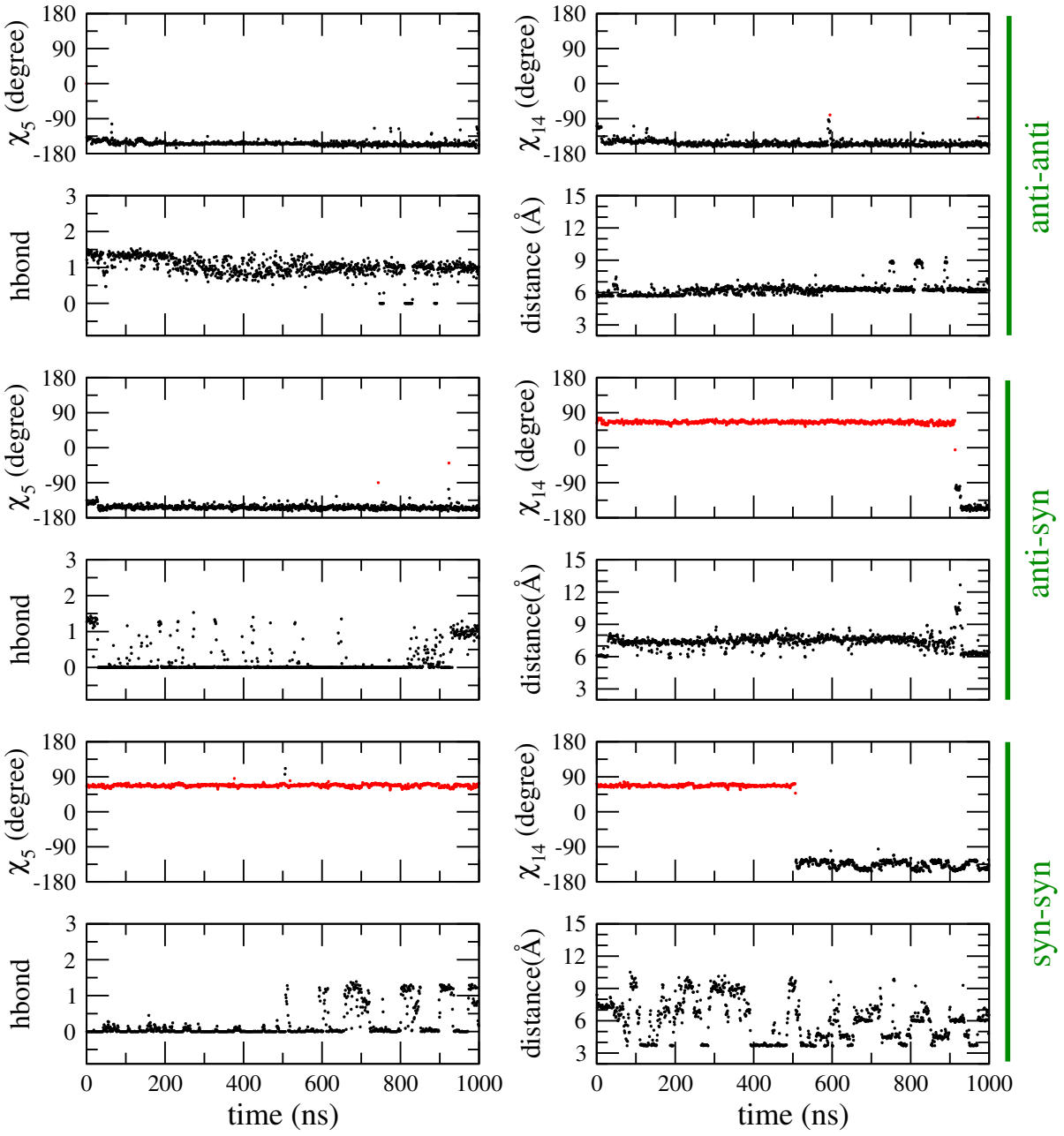


Figure D.1 Characterization of the single C·C mismatch in RNA-CCG1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

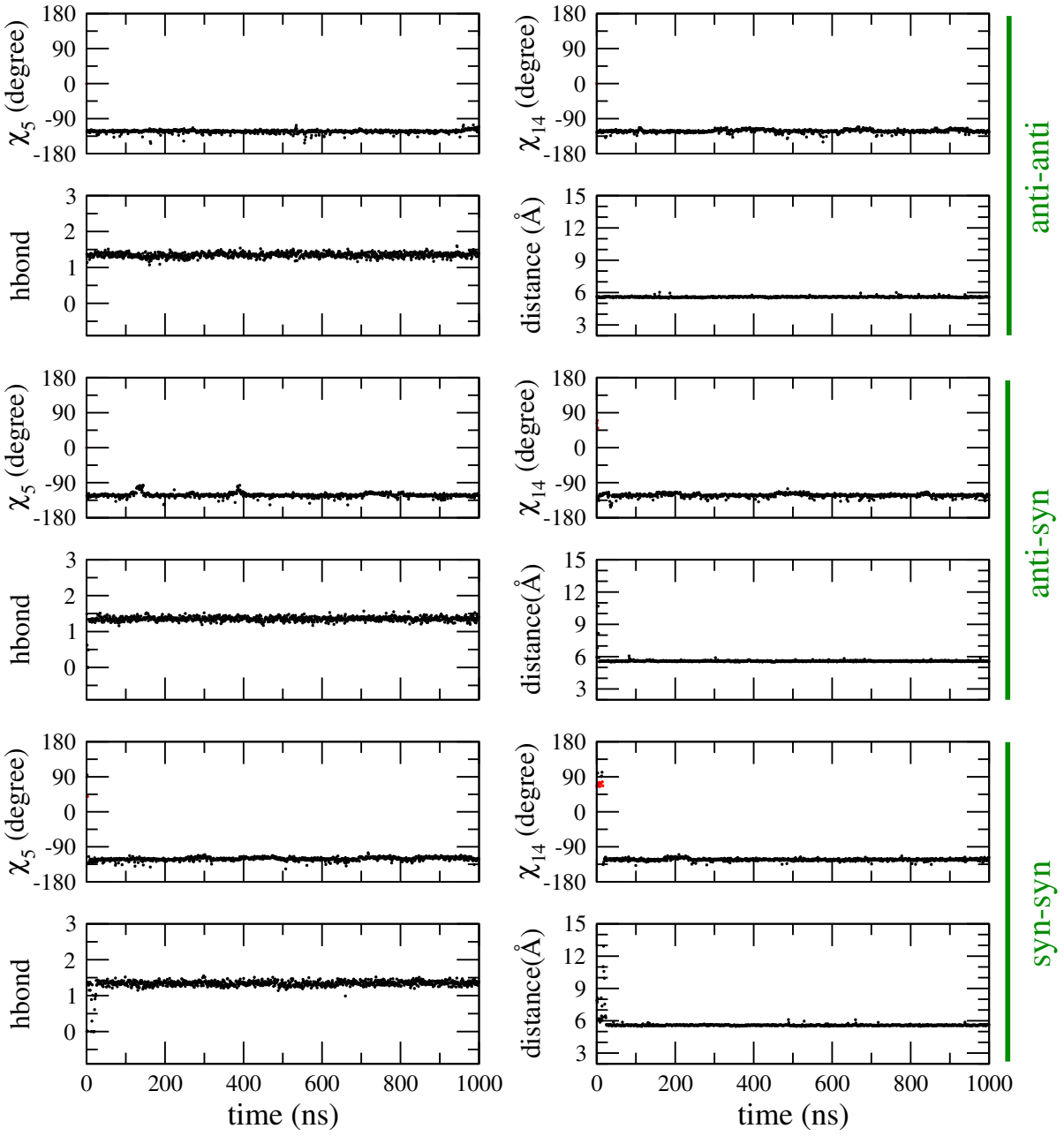


Figure D.2 Characterization of the single C·C mismatch in DNA-CCG1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

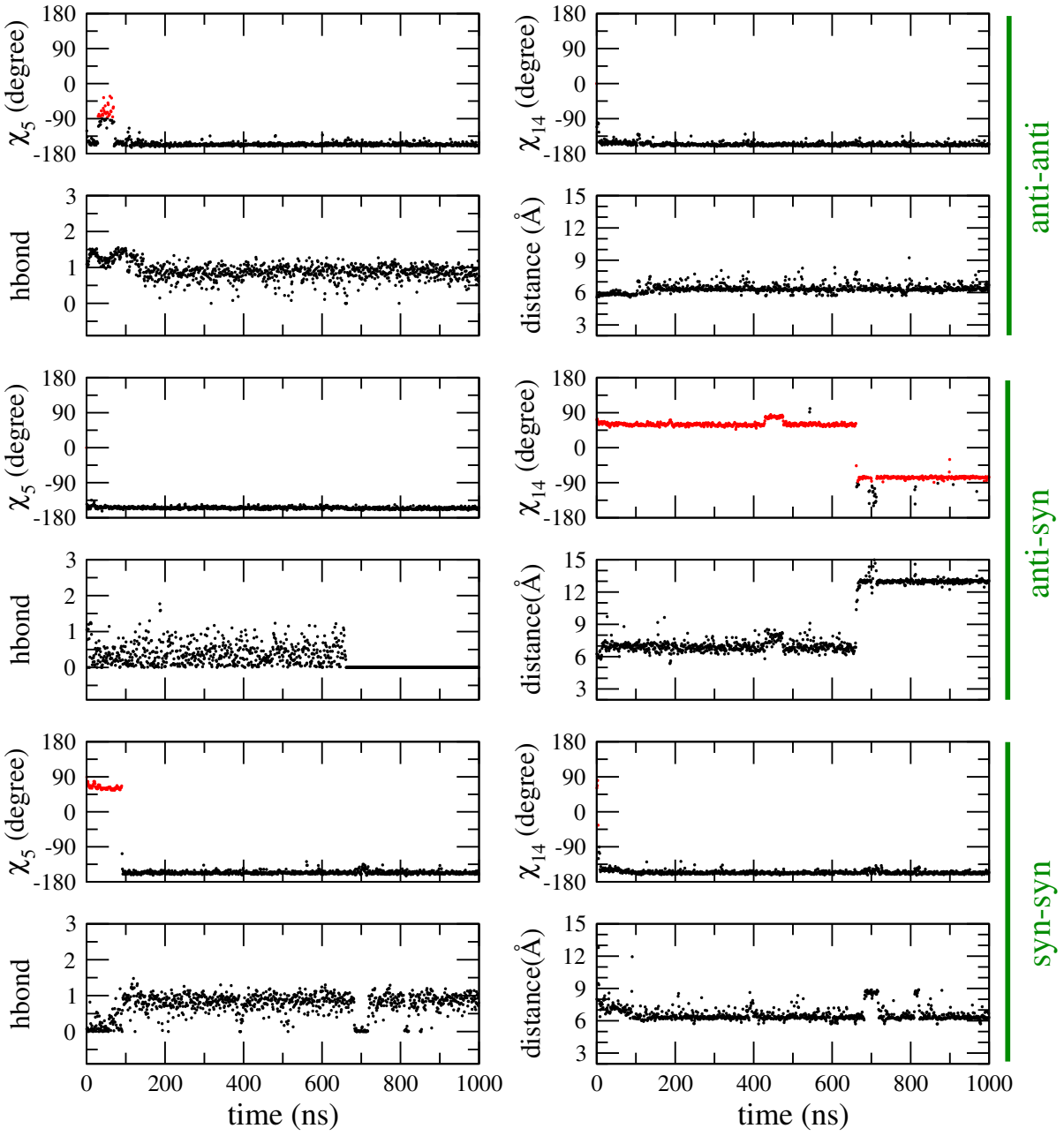


Figure D.3 Characterization of the single C·C mismatch in RNA-GCC1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

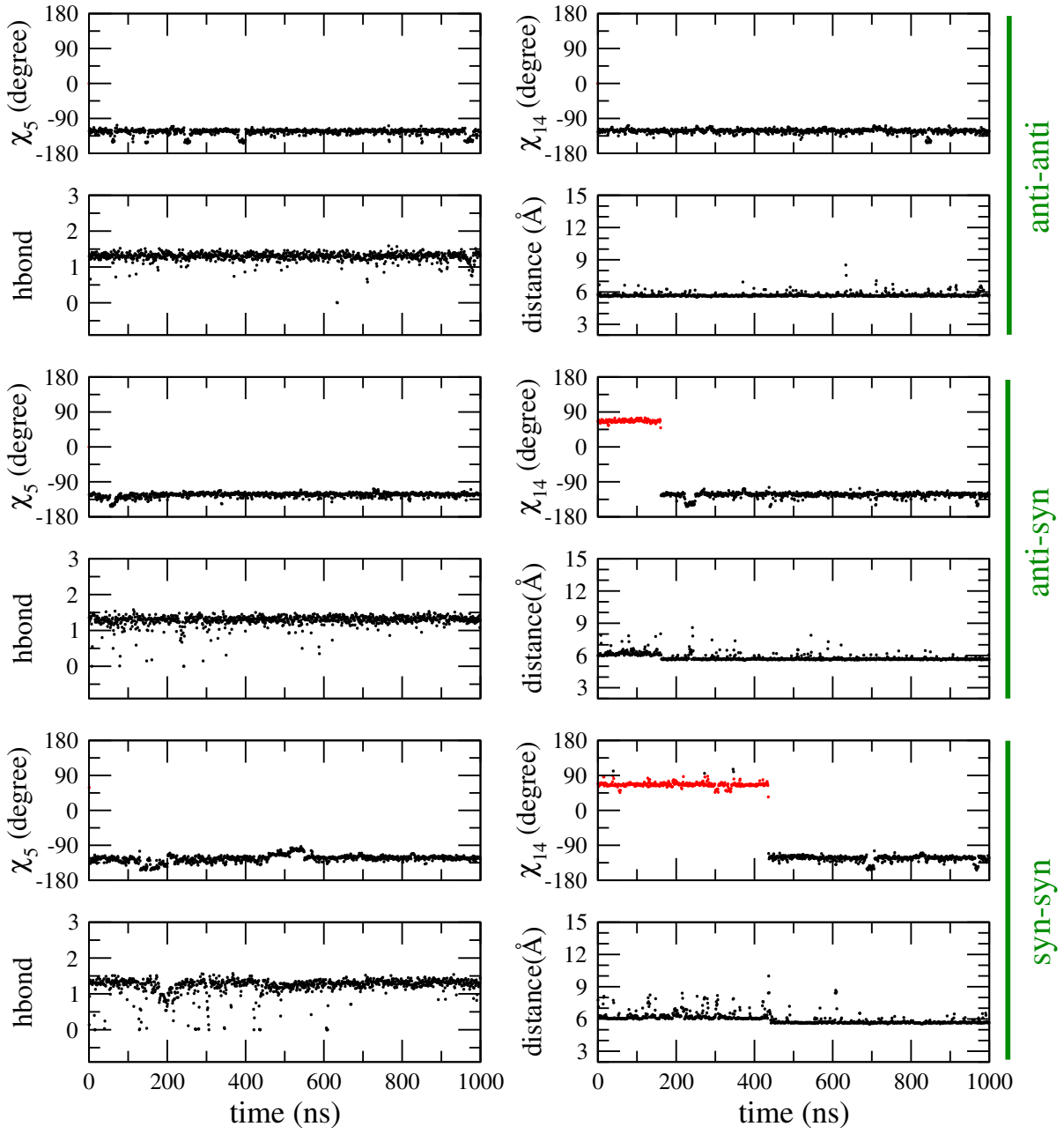


Figure D.4 Characterization of the single C·C mismatch in DNA-GCC1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

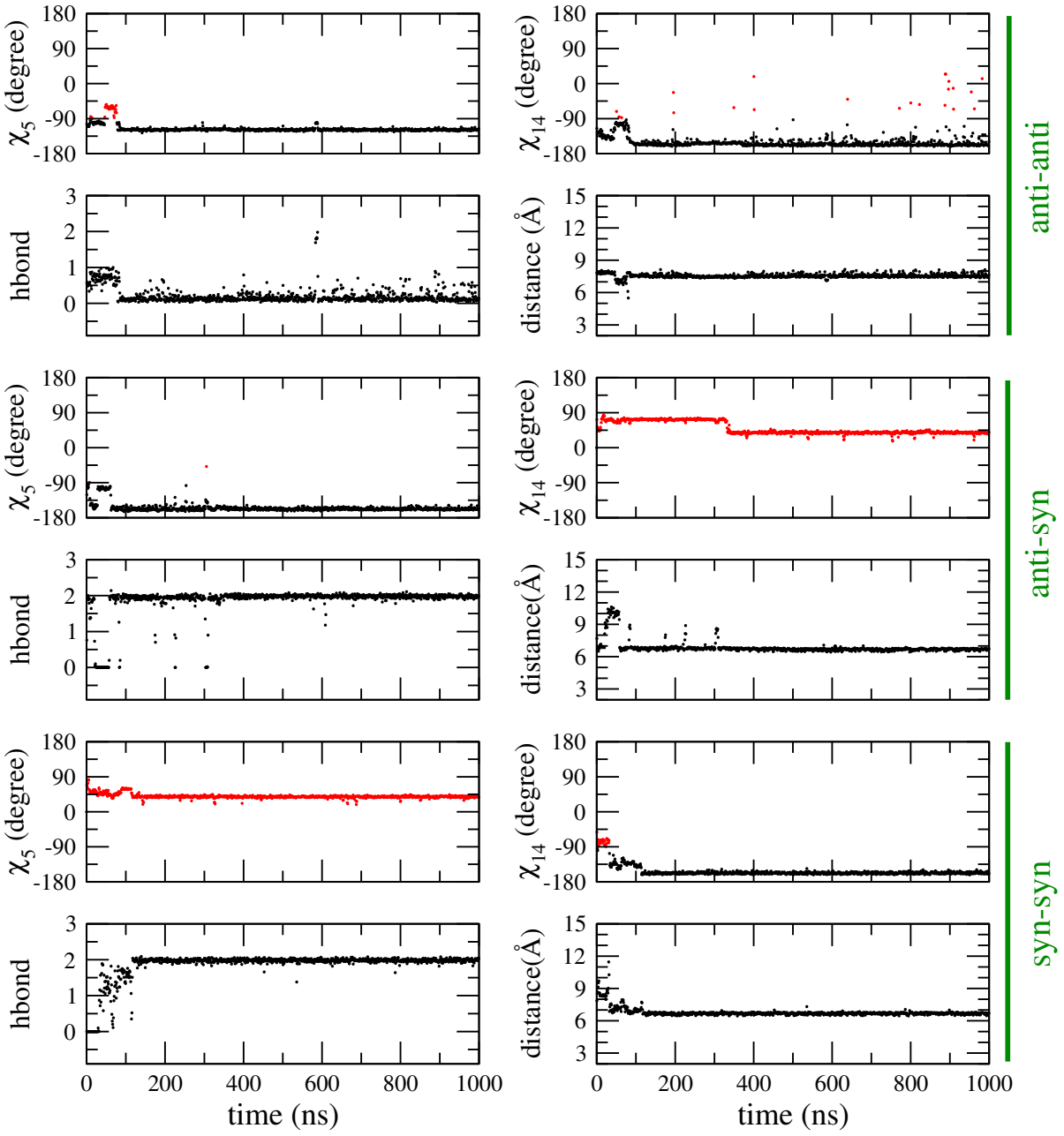


Figure D.5 Characterization of the single G·G mismatch in RNA-CGG1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

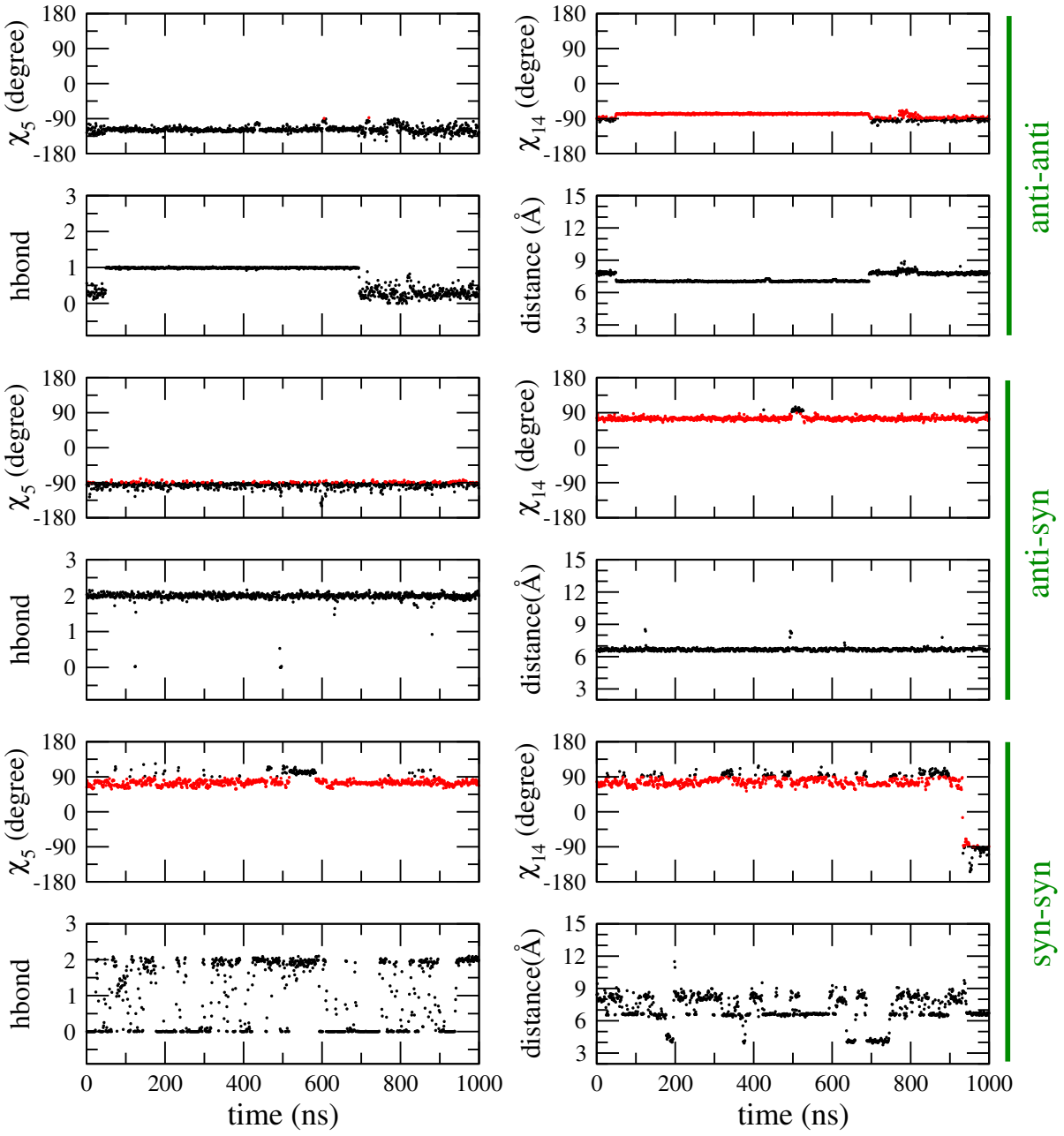


Figure D.6 Characterization of the single G·G mismatch in DNA-CGG1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

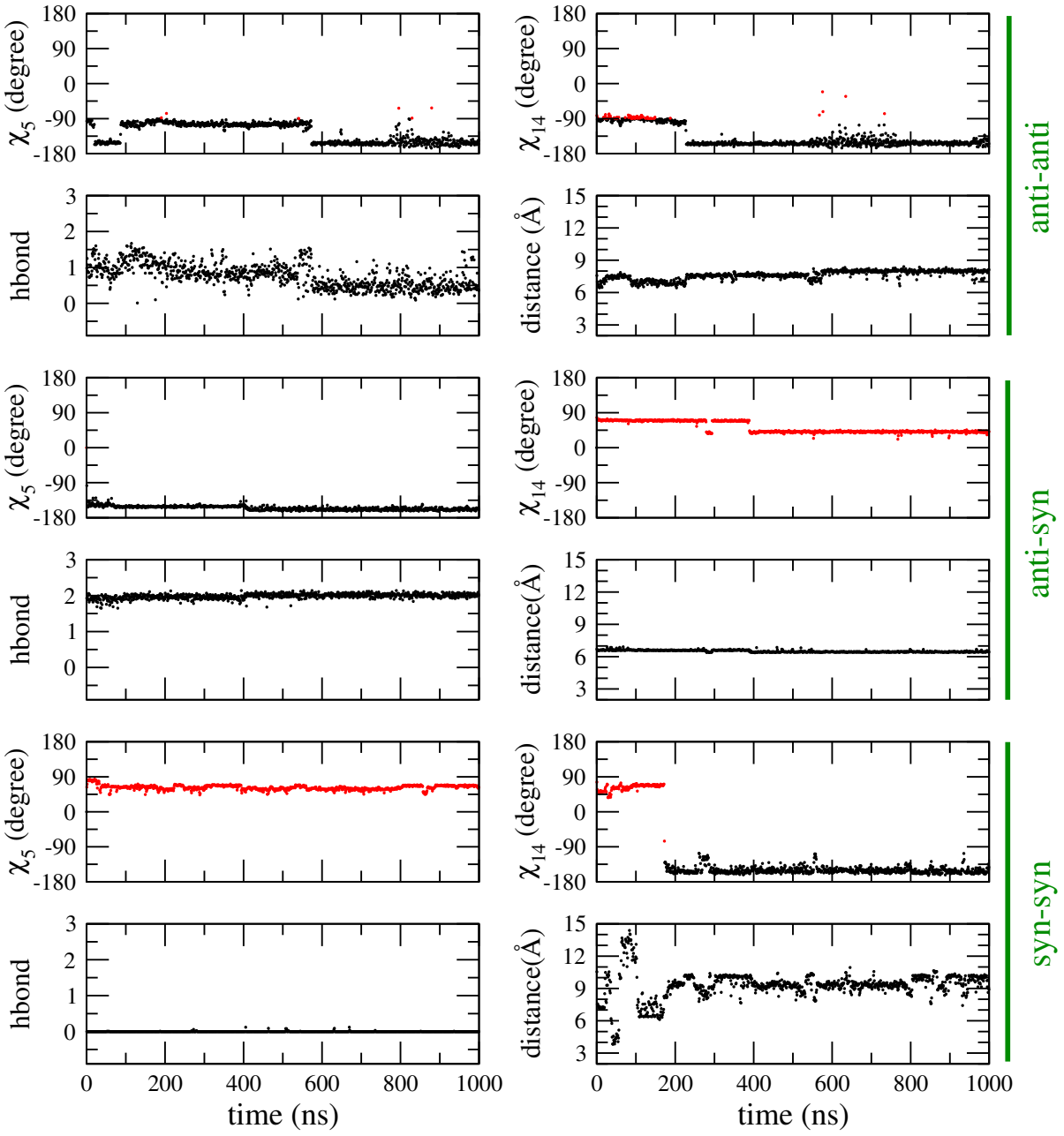


Figure D.7 Characterization of the single G·G mismatch in RNA-GGC1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

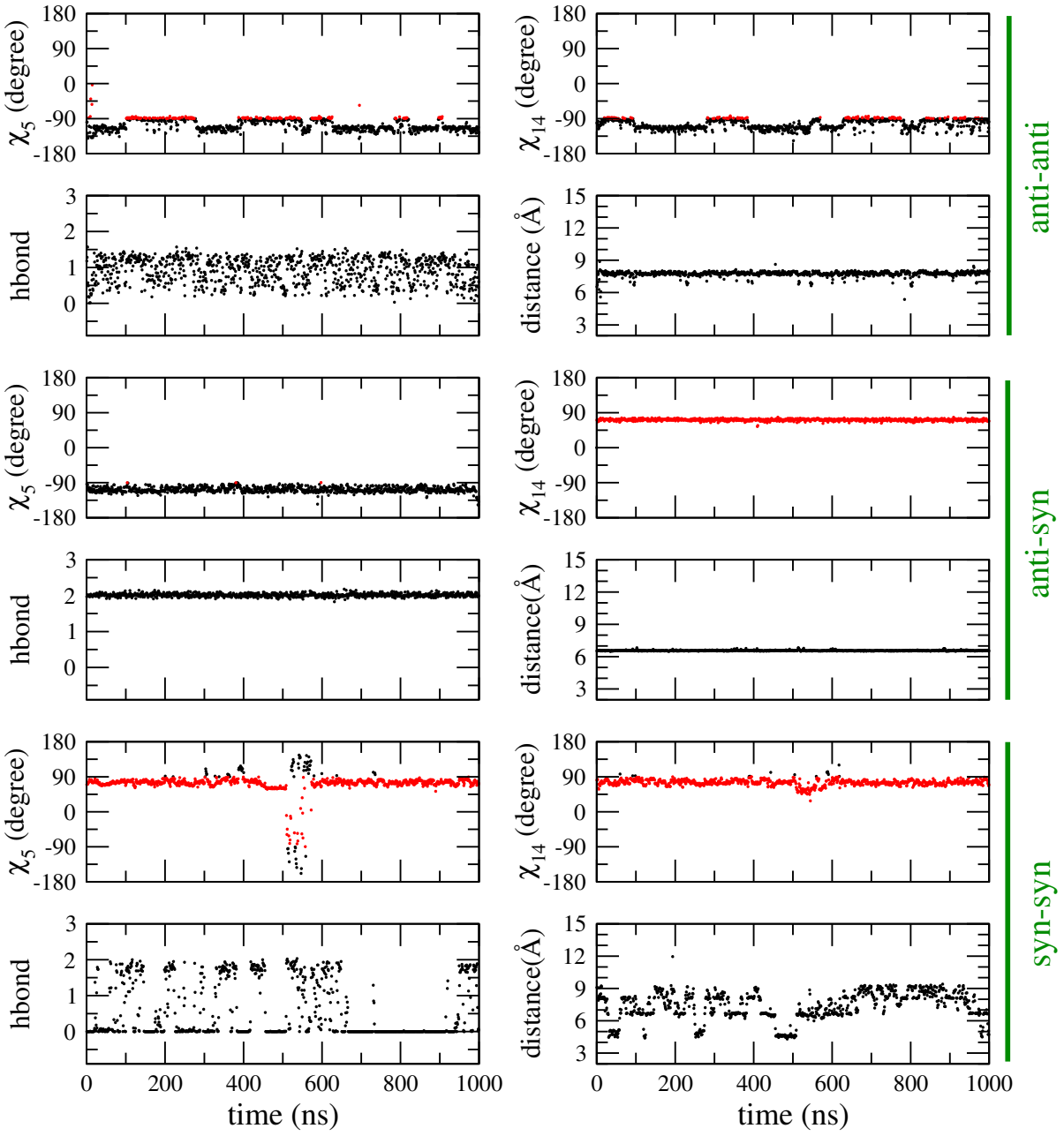


Figure D.8 Characterization of the single G·G mismatch in DNA-GGC1. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

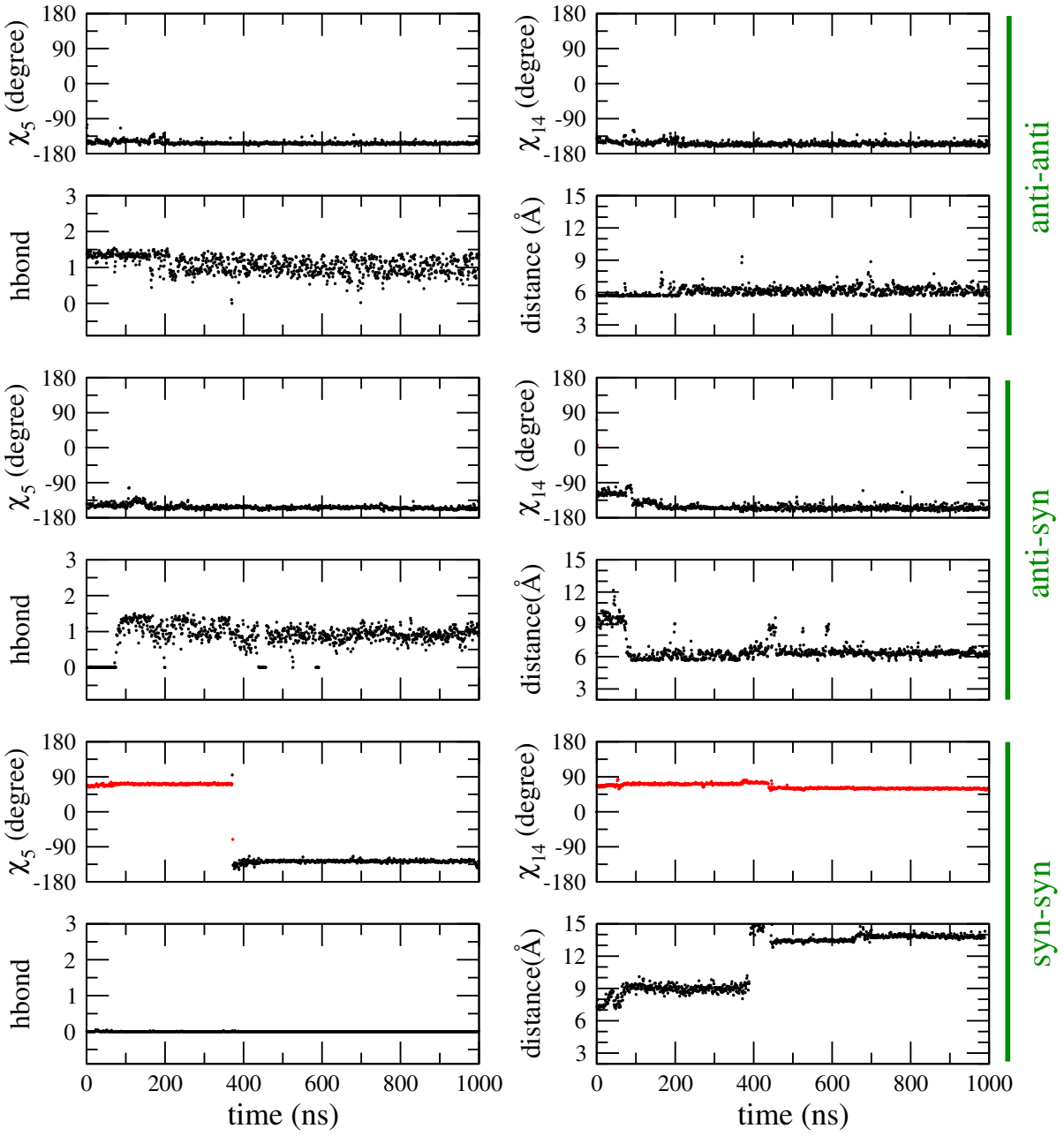


Figure D.9 Characterization of the internal C·C mismatch in RNA-CCG3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

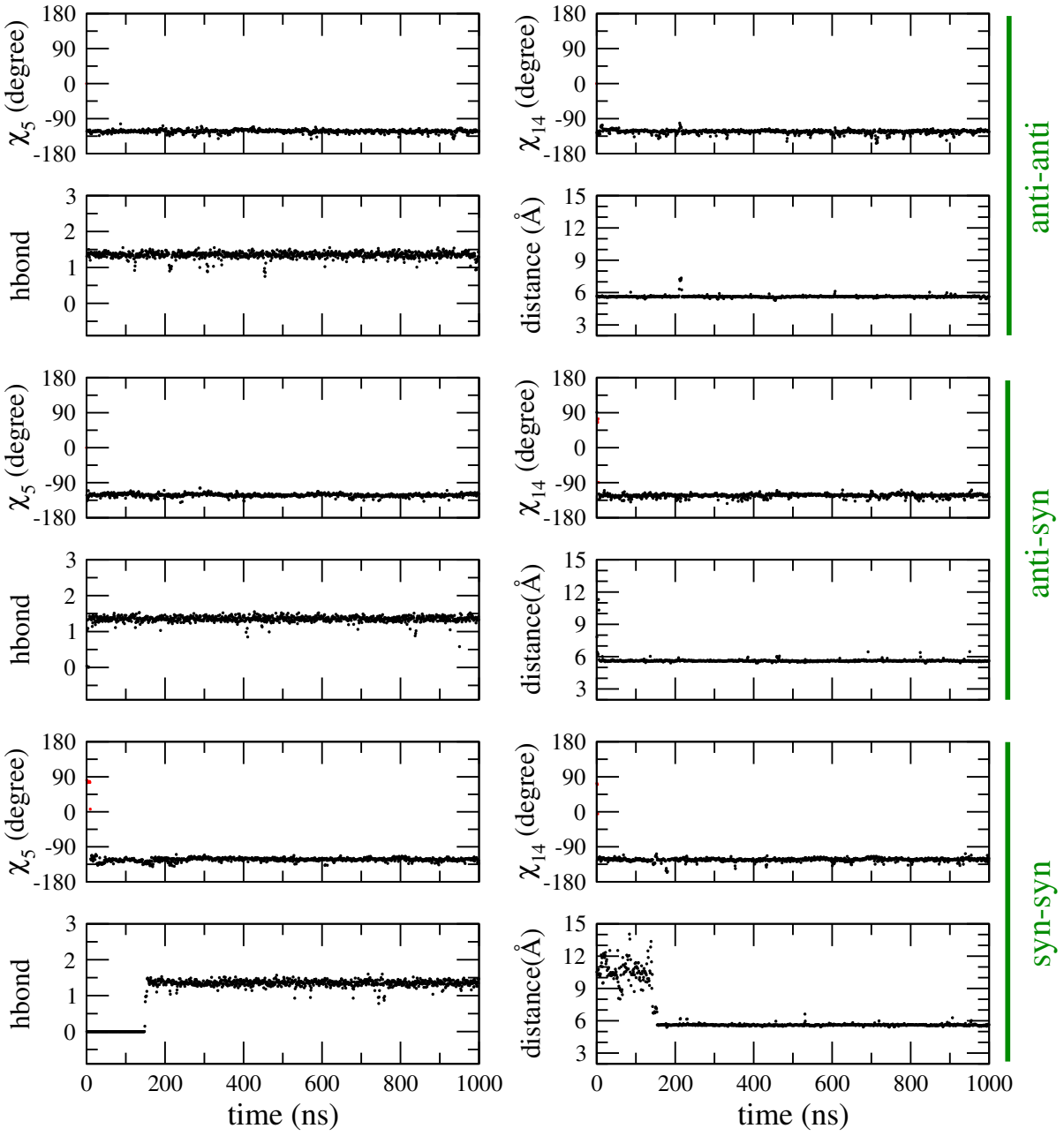


Figure D.10 Characterization of the internal C·C mismatch in DNA-CCG3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

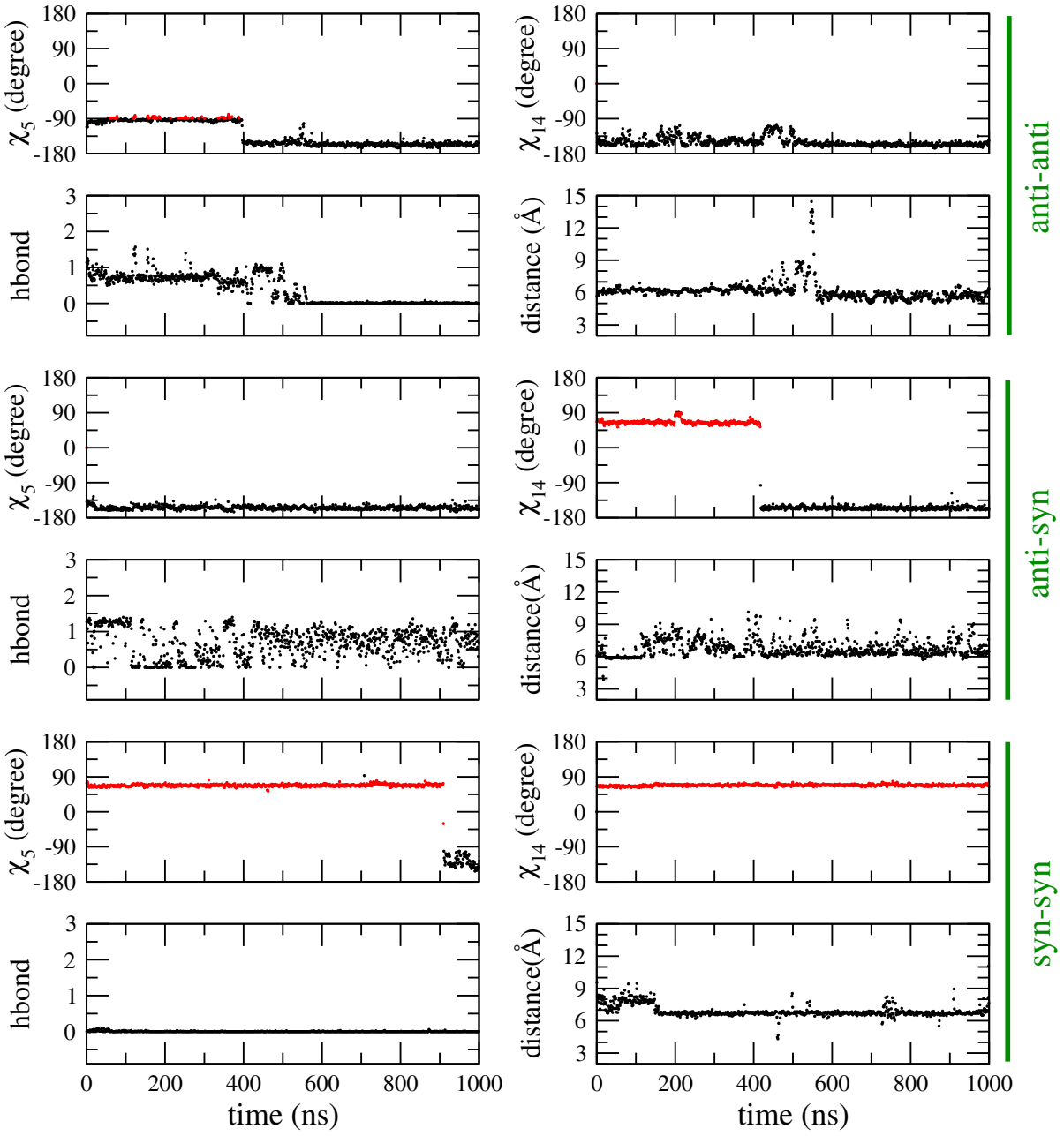


Figure D.11 Characterization of the internal C·C mismatch in RNA-GCC3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

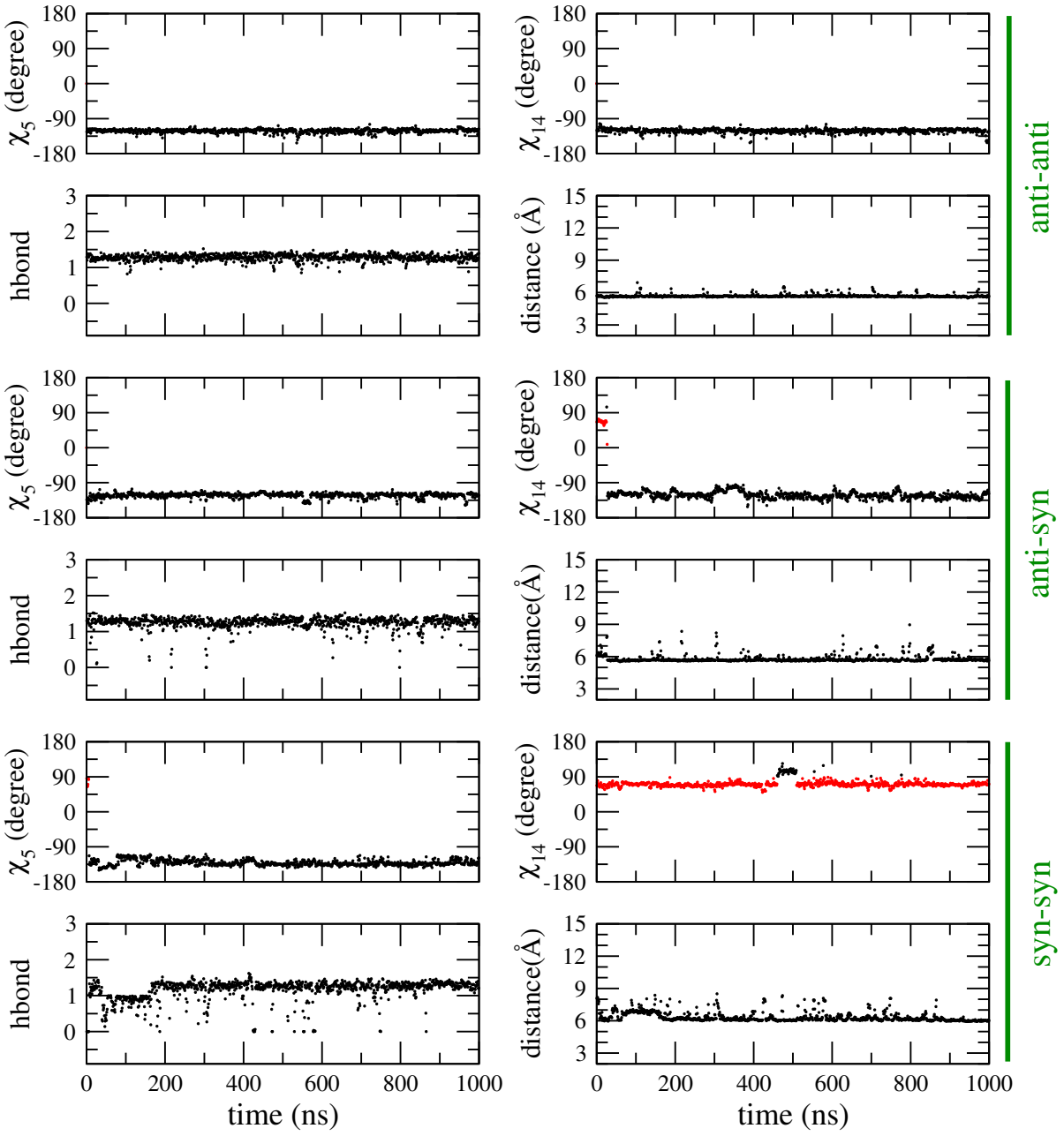


Figure D.12 Characterization of the internal C·C mismatch in DNA-GCC3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

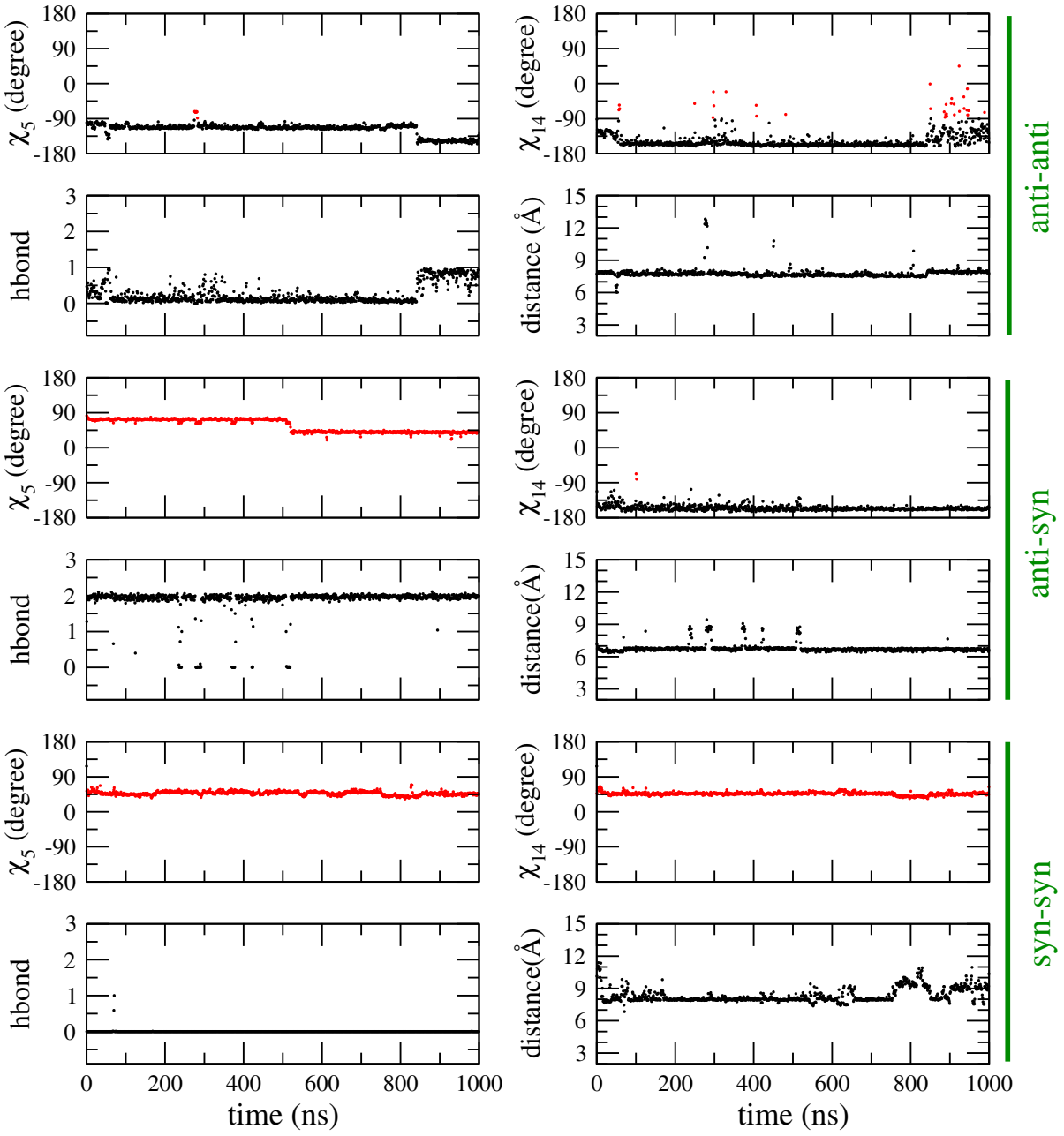


Figure D.13 Characterization of the internal G·G mismatch in RNA-CGG3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

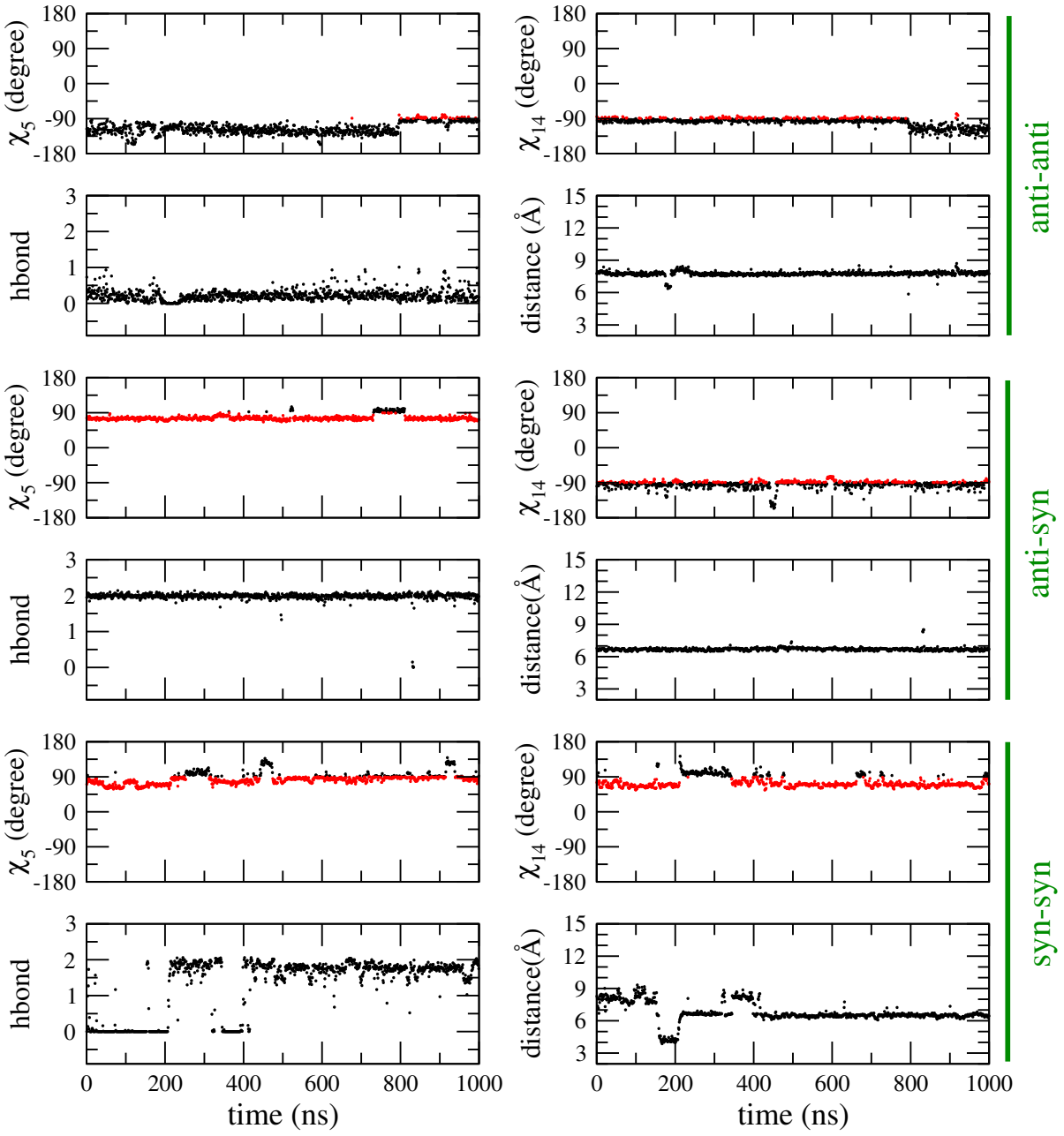


Figure D.14 Characterization of the internal G·G mismatch in DNA-CGG3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

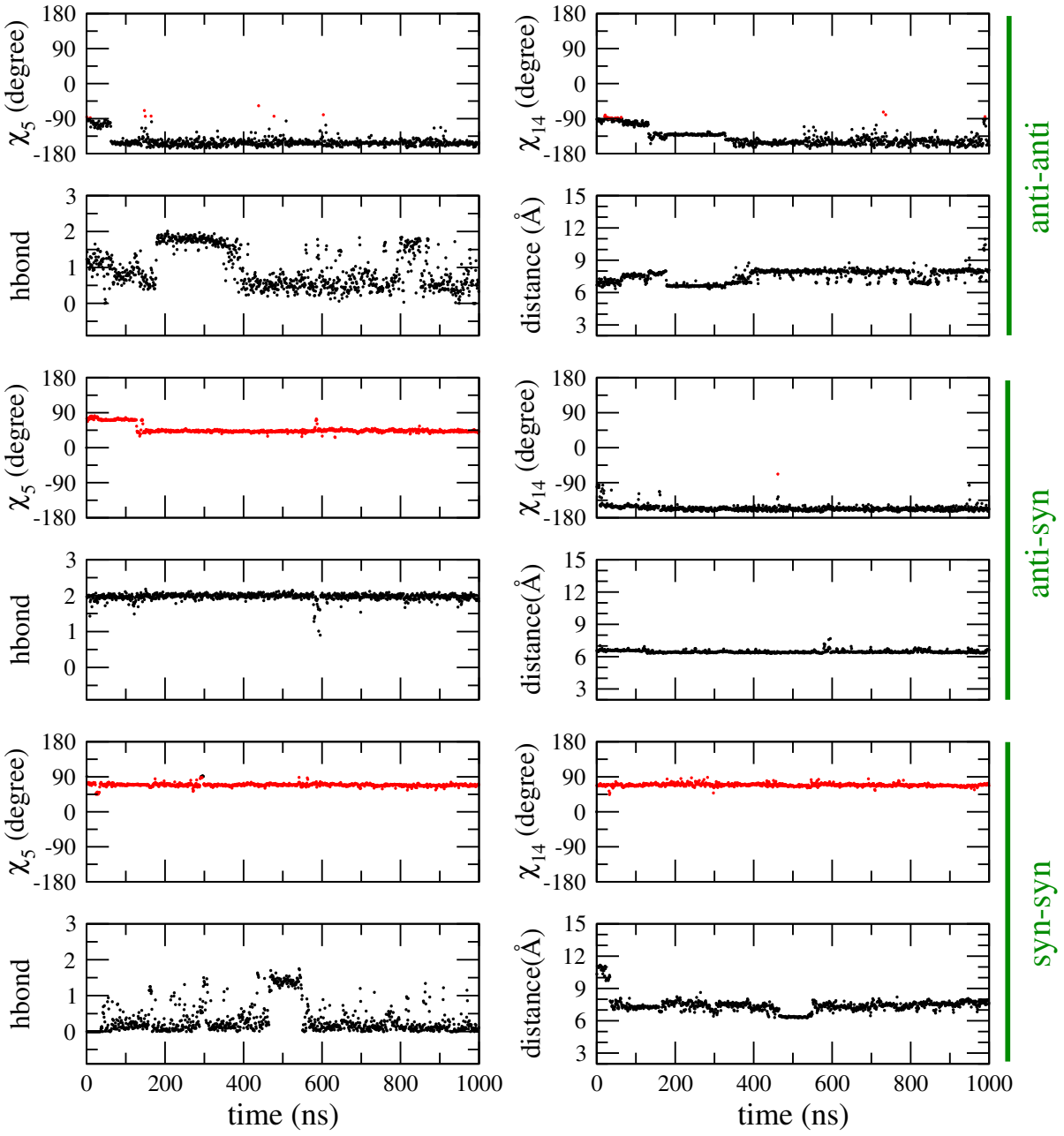


Figure D.15 Characterization of the internal G·G mismatch in RNA-GGC3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

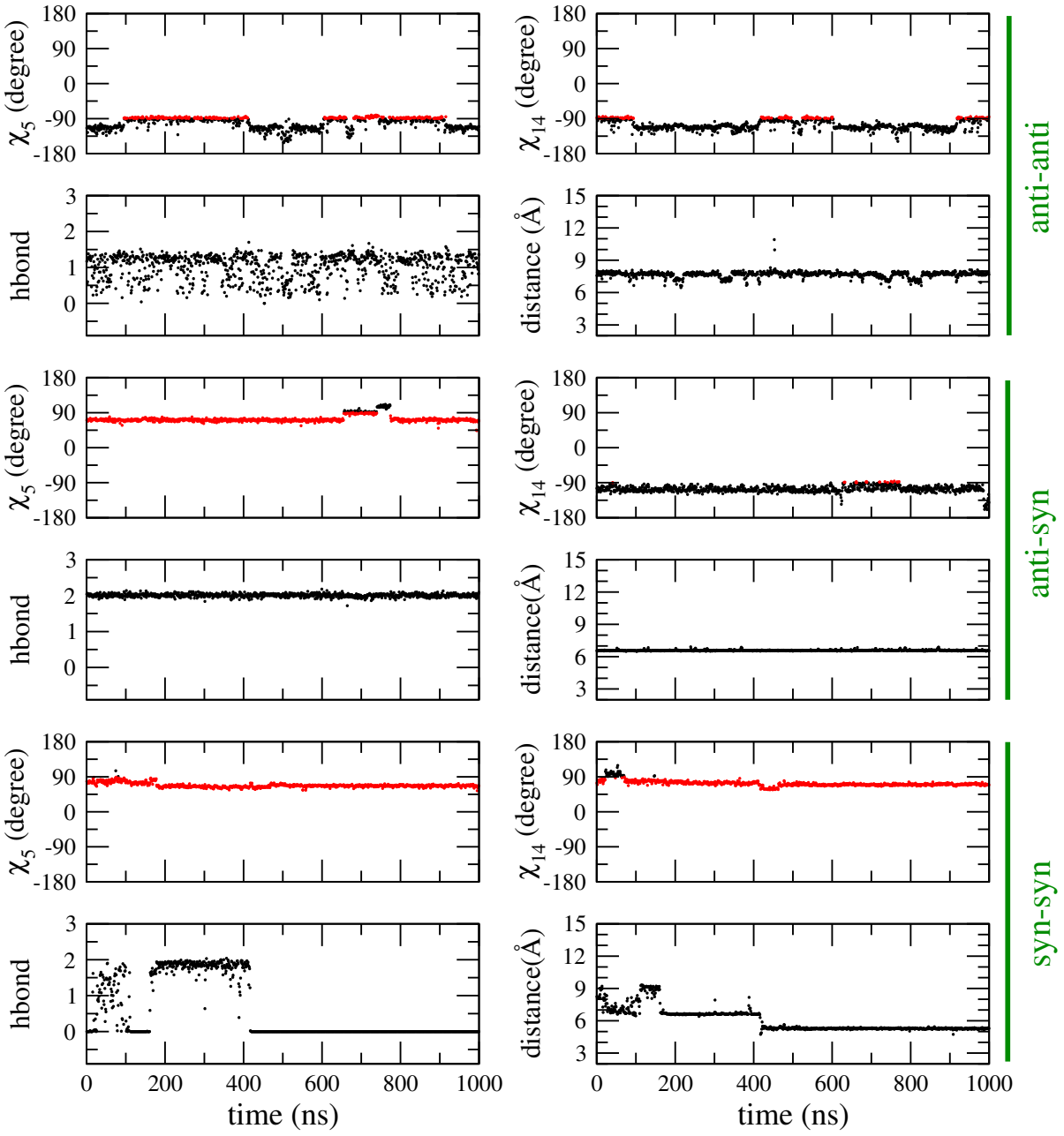


Figure D.16 Characterization of the internal G·G mismatch in DNA-GGC3. Shown are the torsion χ angles, the hydrogen bond number (hbond) and the distance between the centers of mass of the bases in the mismatch. In the χ_5 and χ_{14} plots, anti and syn conformations are drawn in black and red colors, respectively. The data was averaged for every 100 ps.

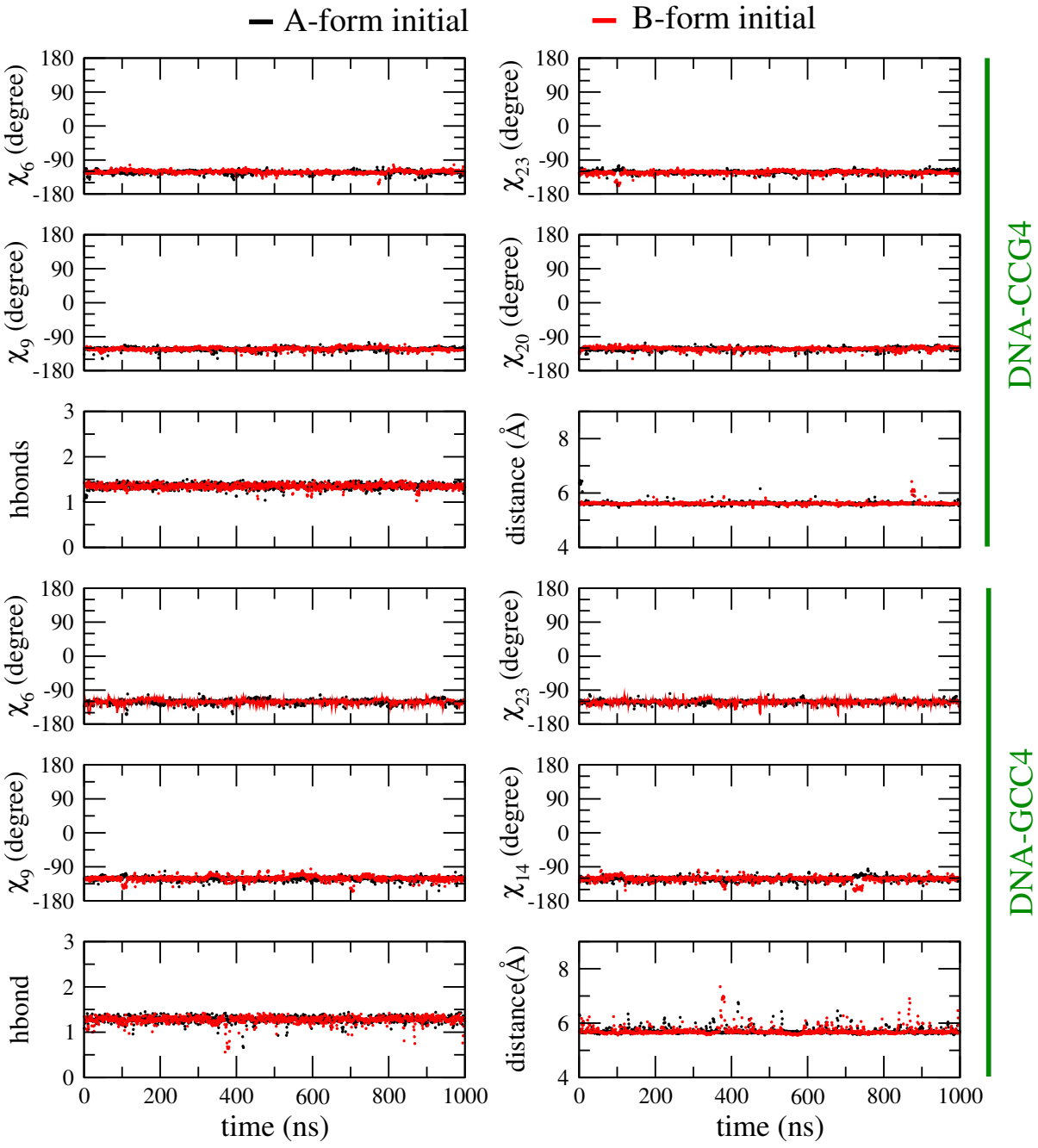


Figure D.17 Time dependence of χ_6, χ_9, χ_{20}, χ_{23}, hydrogen bond and the distance between the centers of mass of the mismatch bases in the DNA four C-C mismatch models.

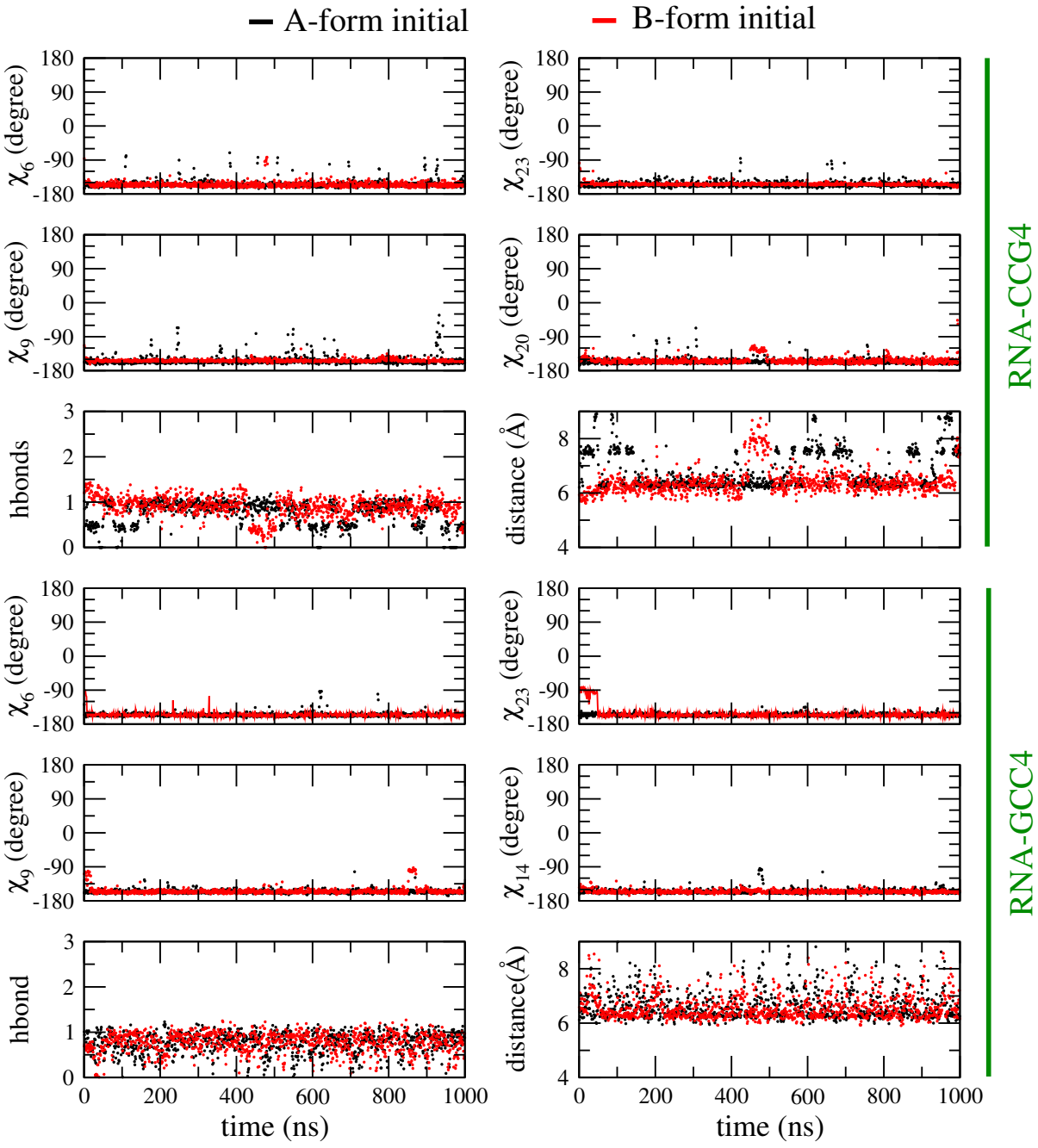


Figure D.18 Time dependence of χ_6, χ_9, χ_{20}, χ_{23}, hydrogen bond and the distance between the centers of mass of the mismatch bases in the RNA four C-C mismatch models.

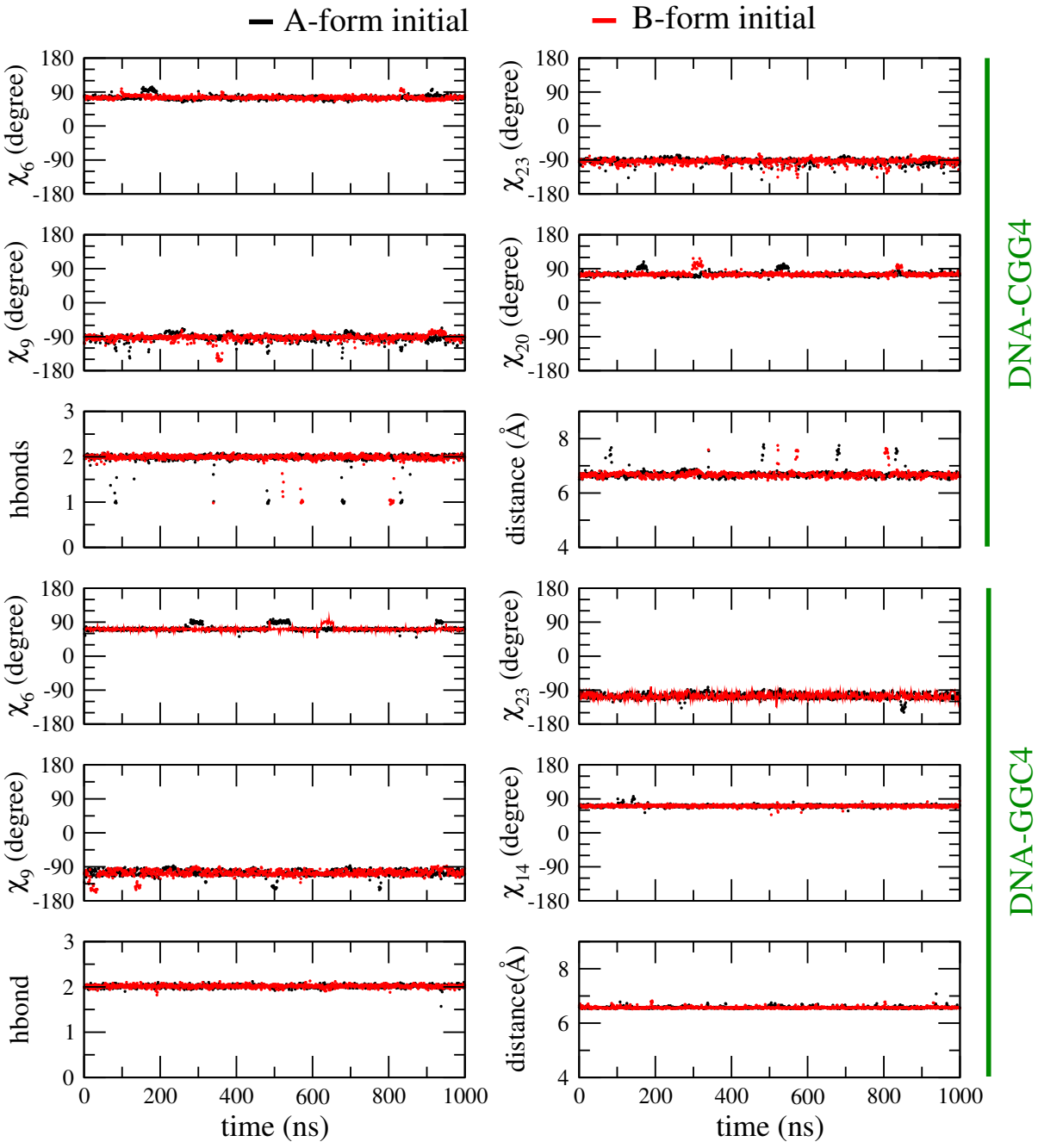


Figure D.19 Time dependence of χ_6, χ_9, χ_{20}, χ_{23}, hydrogen bond and distance between the centers of mass of the mismatch bases in the DNA four G-G mismatch models.

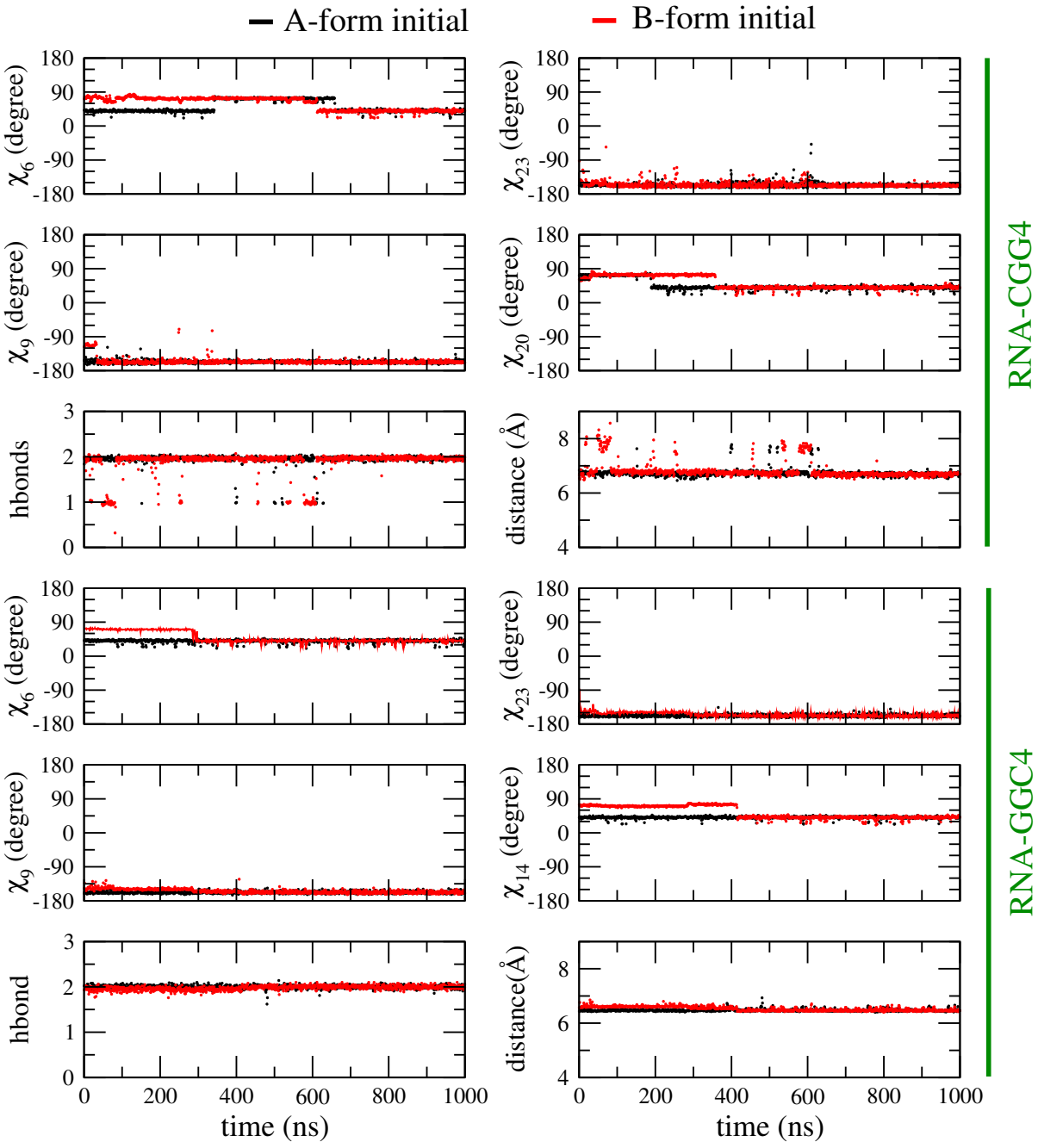


Figure D.20 Time dependence of χ_6, χ_9, χ_{20}, χ_{23}, hydrogen bond and distance between the centers of mass of the mismatch bases in the RNA four G-G mismatch models.

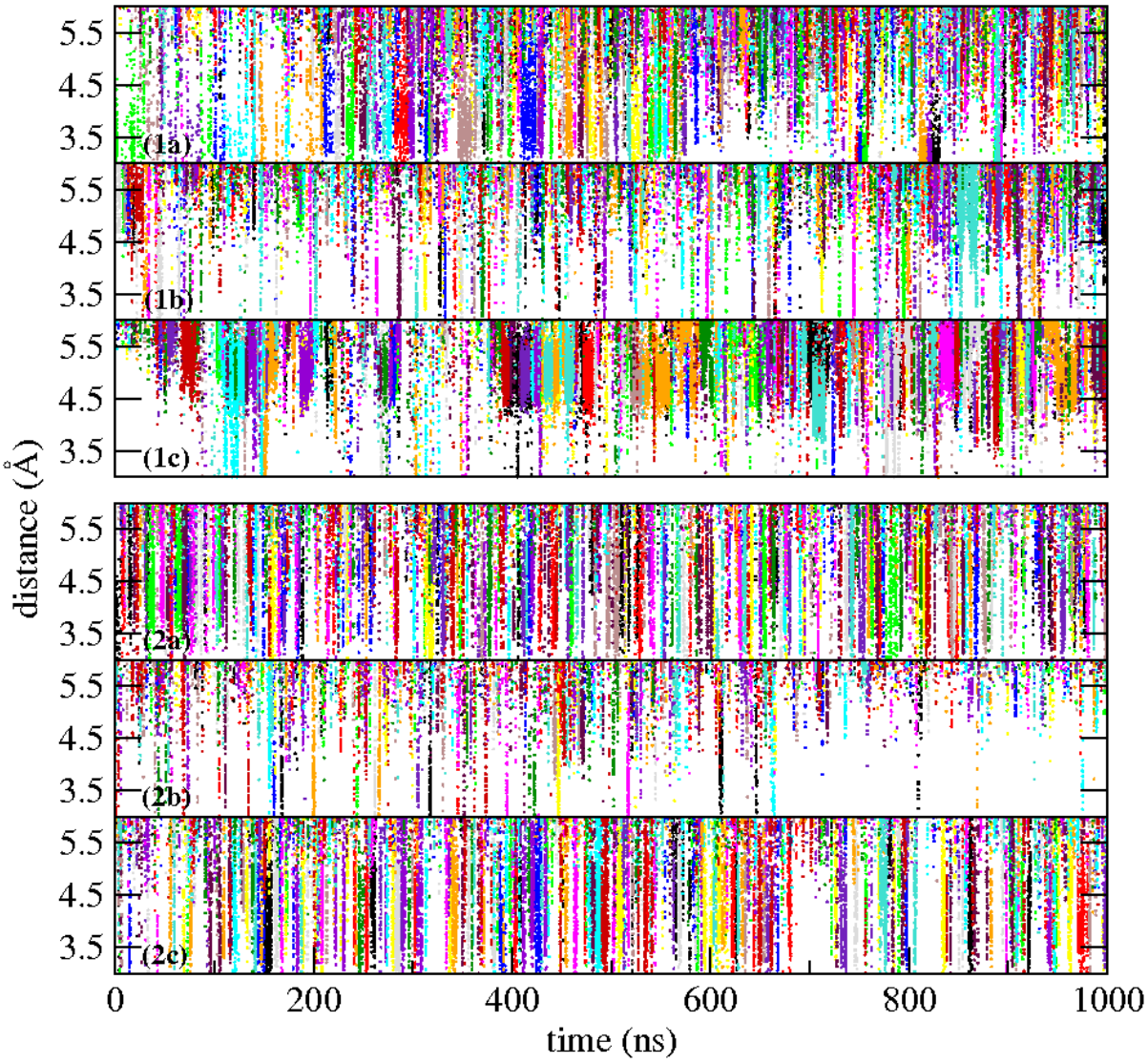


Figure D.21 Distance between Na^+ ions and the center of mass of the C-C mismatches. The single mismatch duplexes are RNA-CCG1 (top) and RNA-GCC1 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

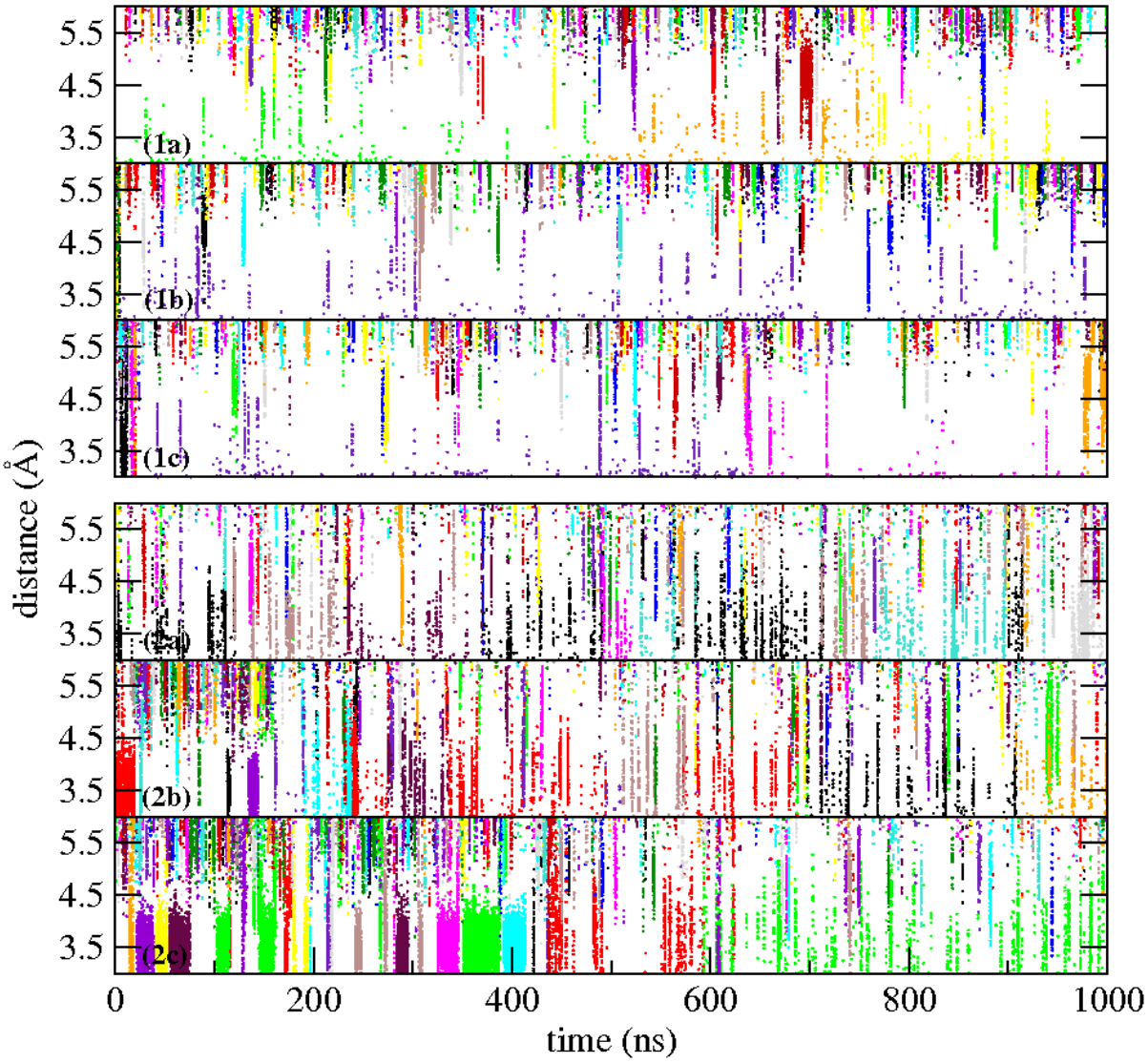


Figure D.22 Distance between Na^+ ions and the center of mass of the C-C mismatches. The single mismatch duplexes are DNA-CCG1 (top) and DNA-GCC1 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

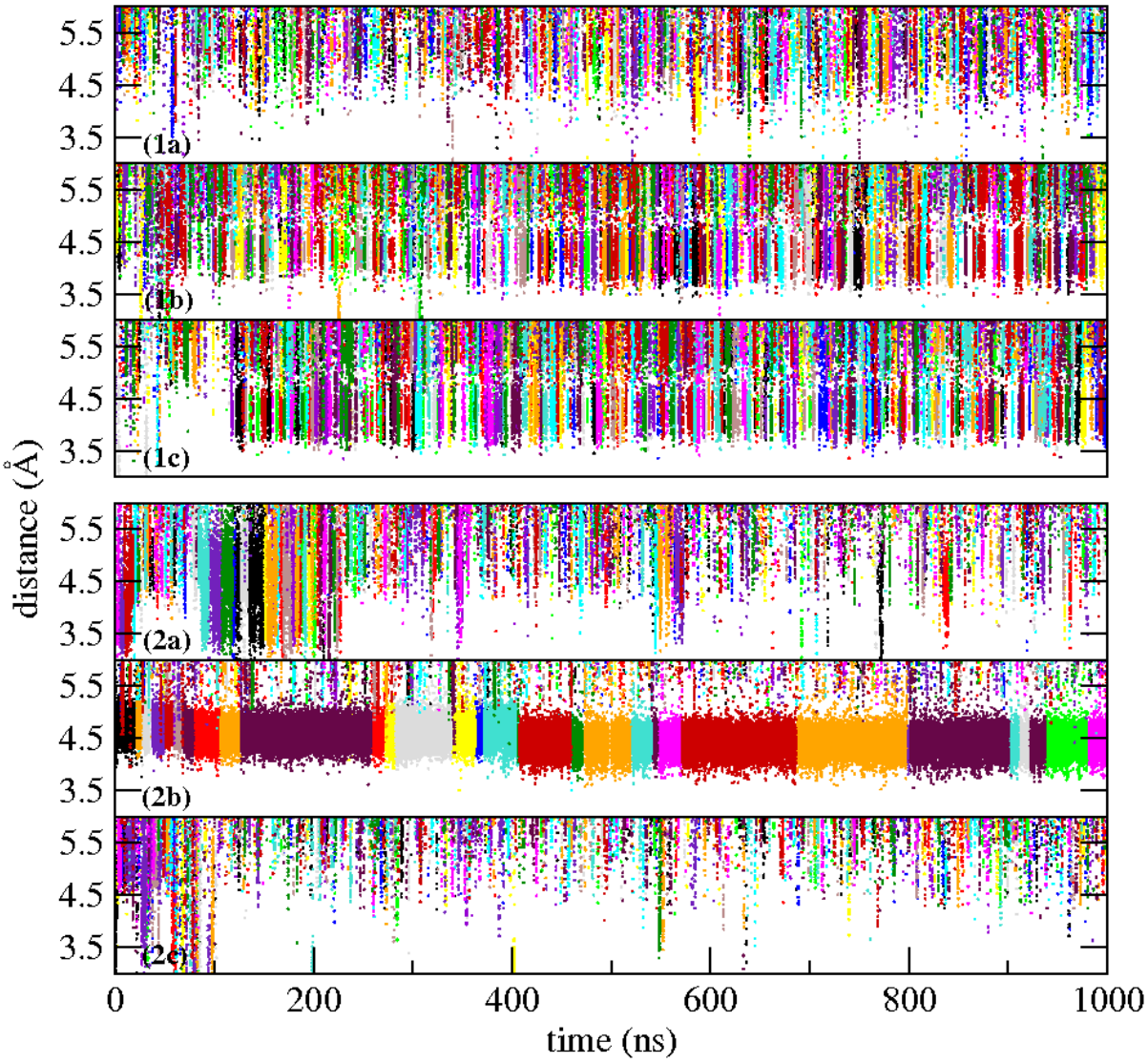


Figure D.23 Distance between Na^+ ions and the center of mass of the G-G mismatches. The single mismatch duplexes are RNA-CGG1 (top) and RNA-GGC1 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

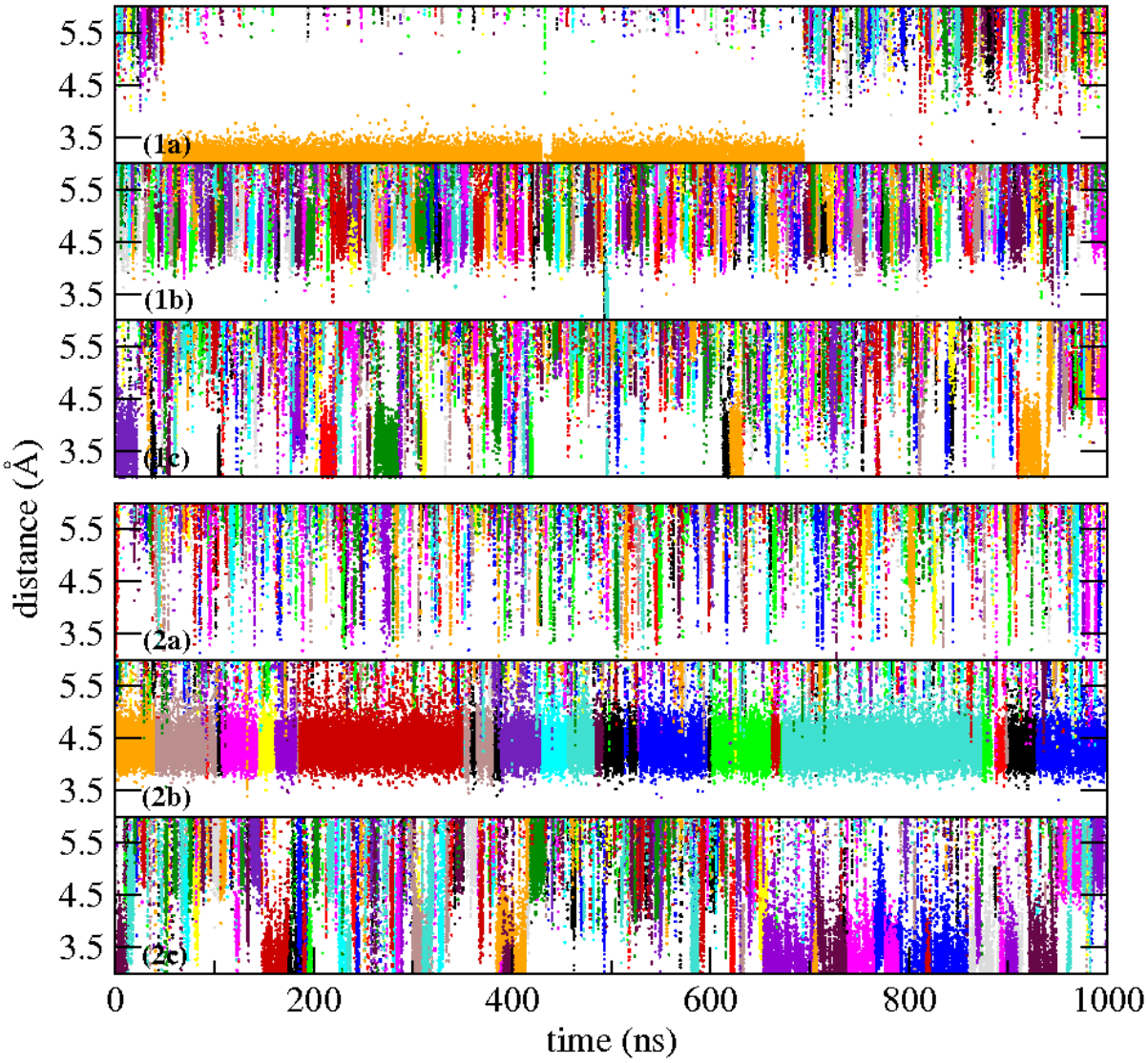


Figure D.24 Distance between Na^+ ions and the center of mass of the G-G mismatches. The single mismatch duplexes are DNA-CGG1 (top) and DNA-GGC1 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

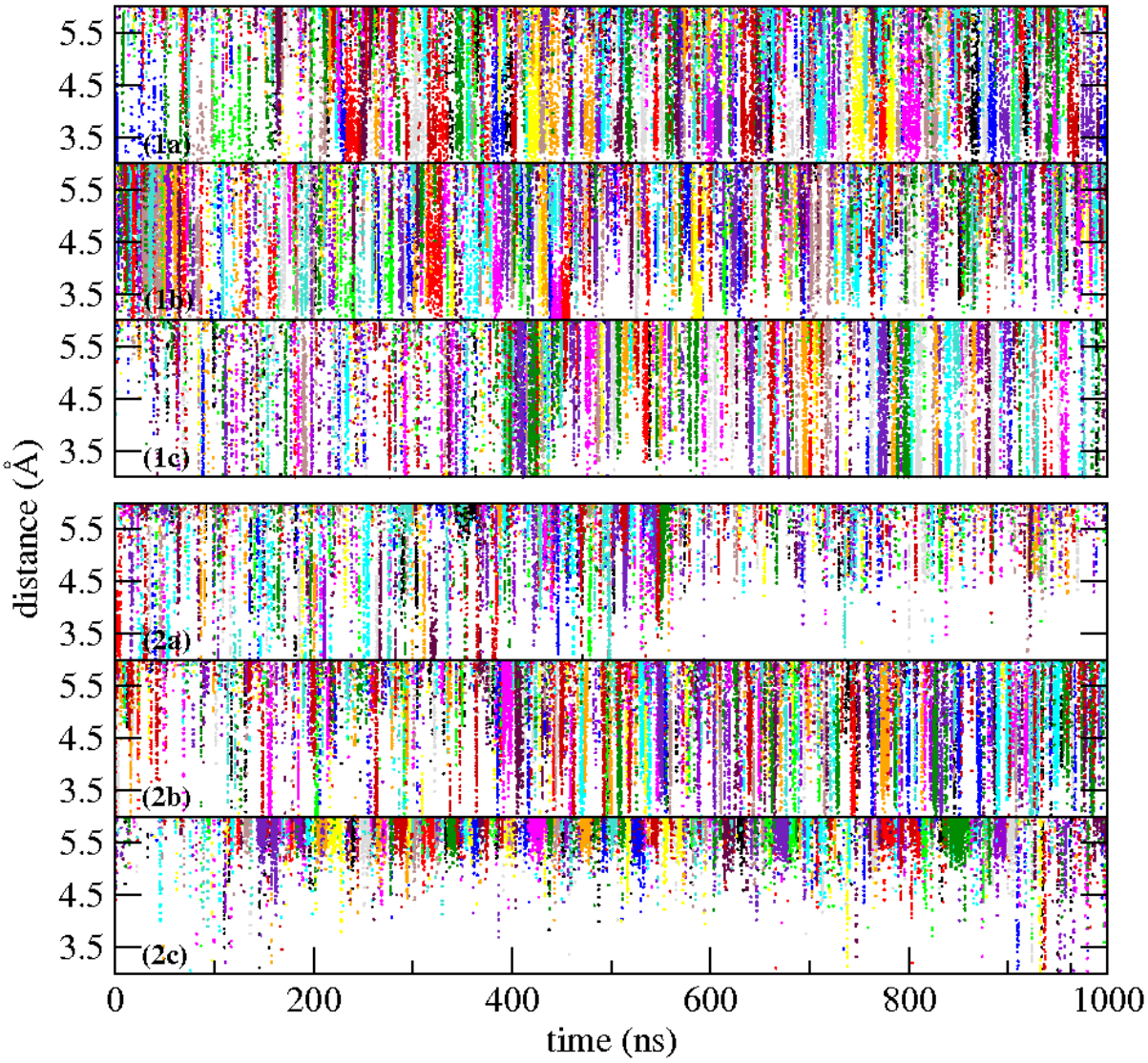


Figure D.25 Distance between Na^+ ions and the center of mass of the C-C mismatches. The three mismatch duplexes are RNA-CCG3 (top) and RNA-GCC3 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

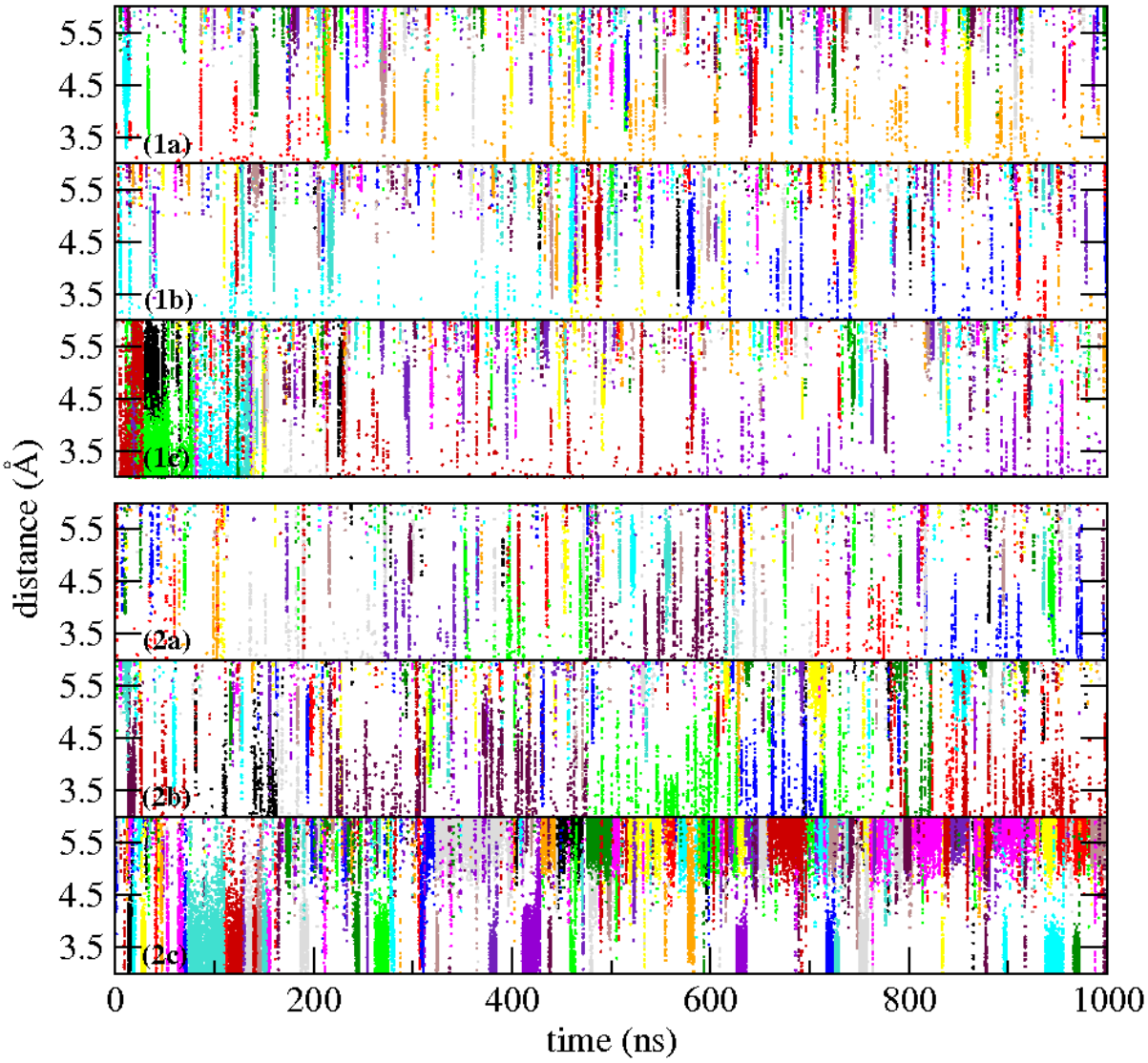


Figure D.26 Distance between Na^+ ions and the center of mass of the C-C mismatches. The three mismatch duplexes are DNA-CCG3 (top) and DNA-GCC3 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

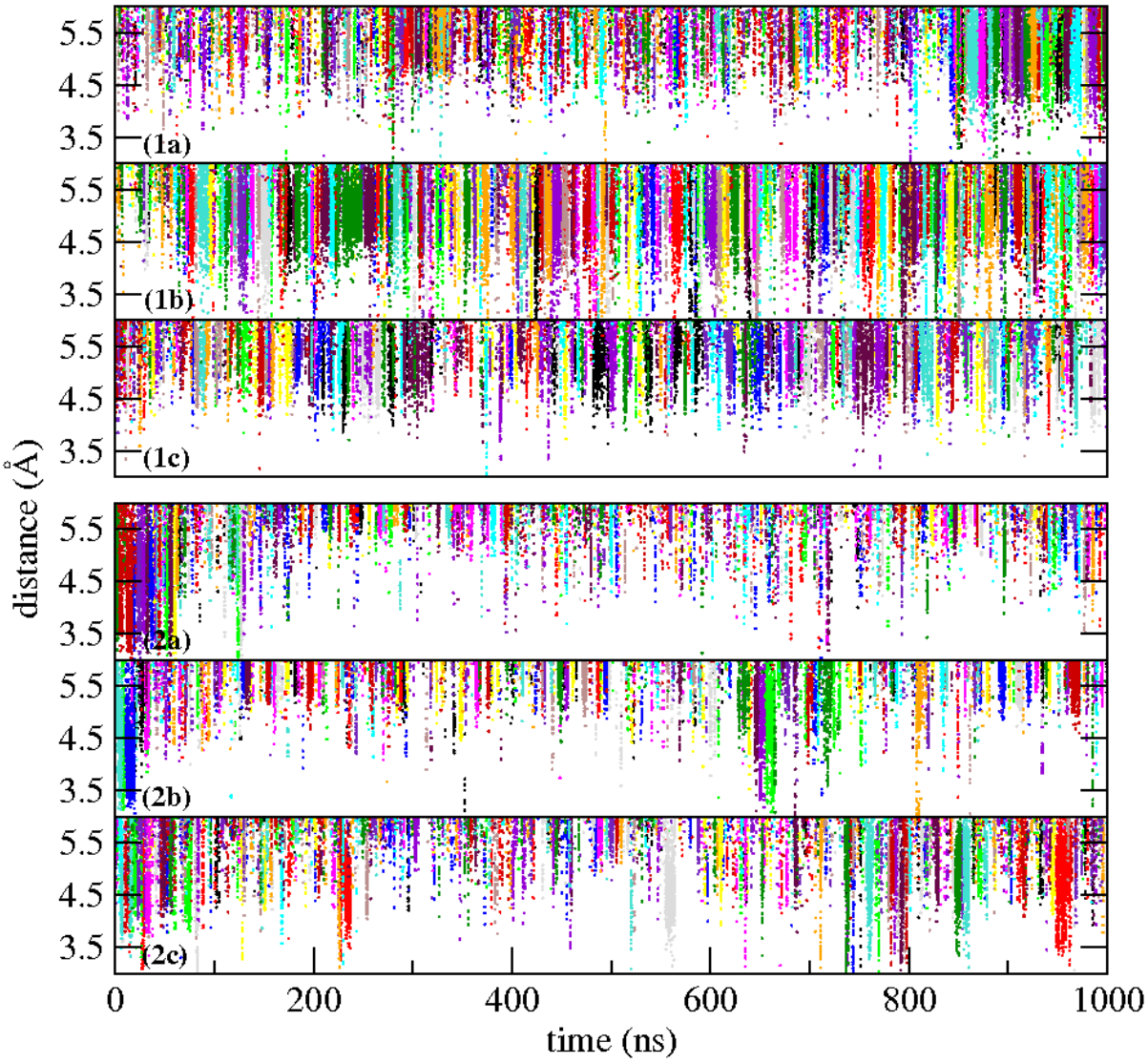


Figure D.27 Distance between Na^+ ions and the center of mass of the G-G mismatches. The three mismatch duplexes are RNA-CGG3 (top) and RNA-GGC3 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

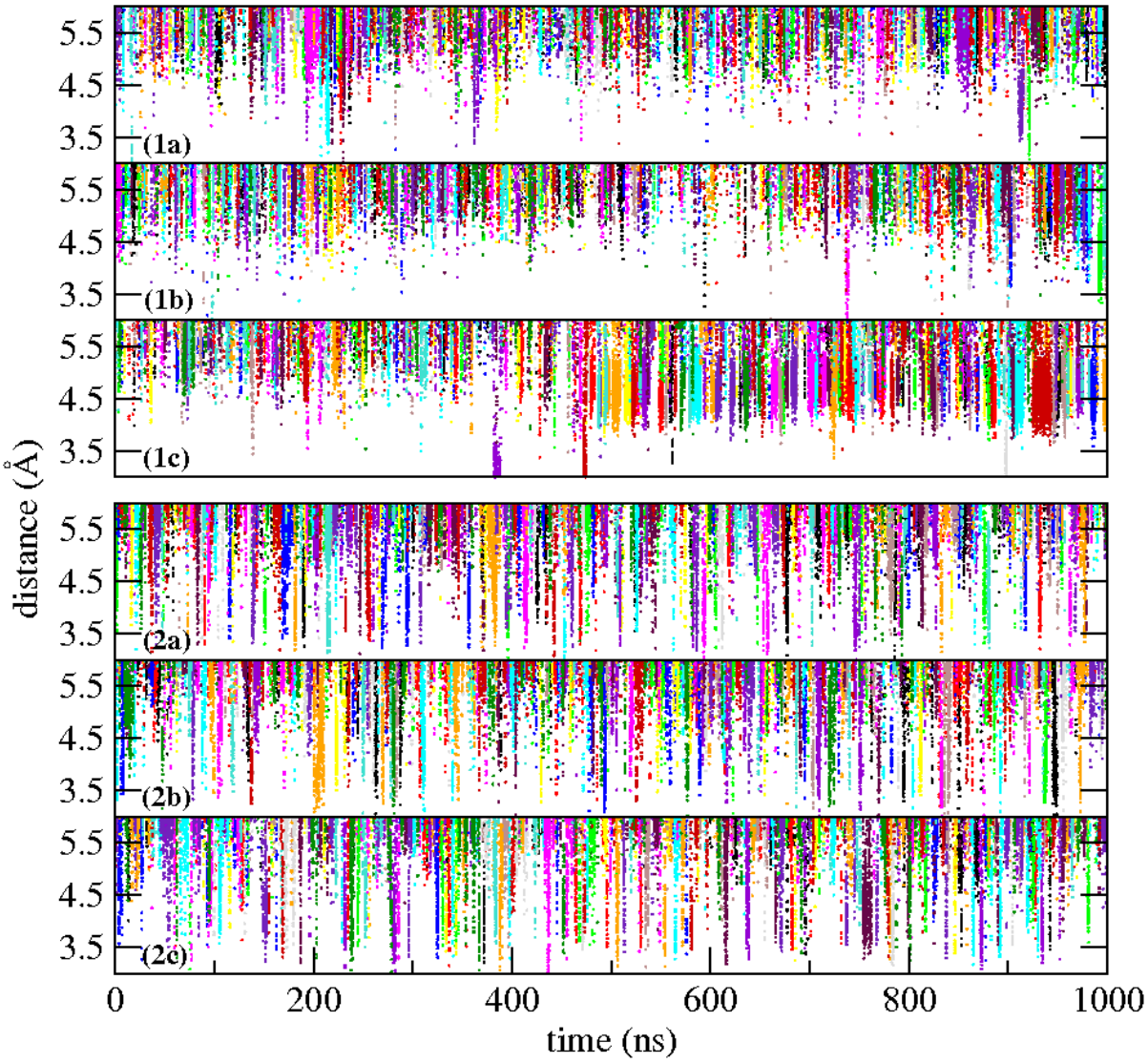


Figure D.28 Distance between Na^+ ions and the center of mass of the G-G mismatches. The three mismatch duplexes are DNA-CGG3 (top) and DNA-GGC3 (bottom) For each duplex, the initial conformations for the top, middle and bottom panels are anti-anti, anti-syn, and syn-syn respectively. Different colors represent different ions to show the binding time for individual ions.

Appendix E

E-motif formed by extrahelical cytosine bases in DNA homoduplexes of trinucleotide and hexanucleotide repeats - Supporting information

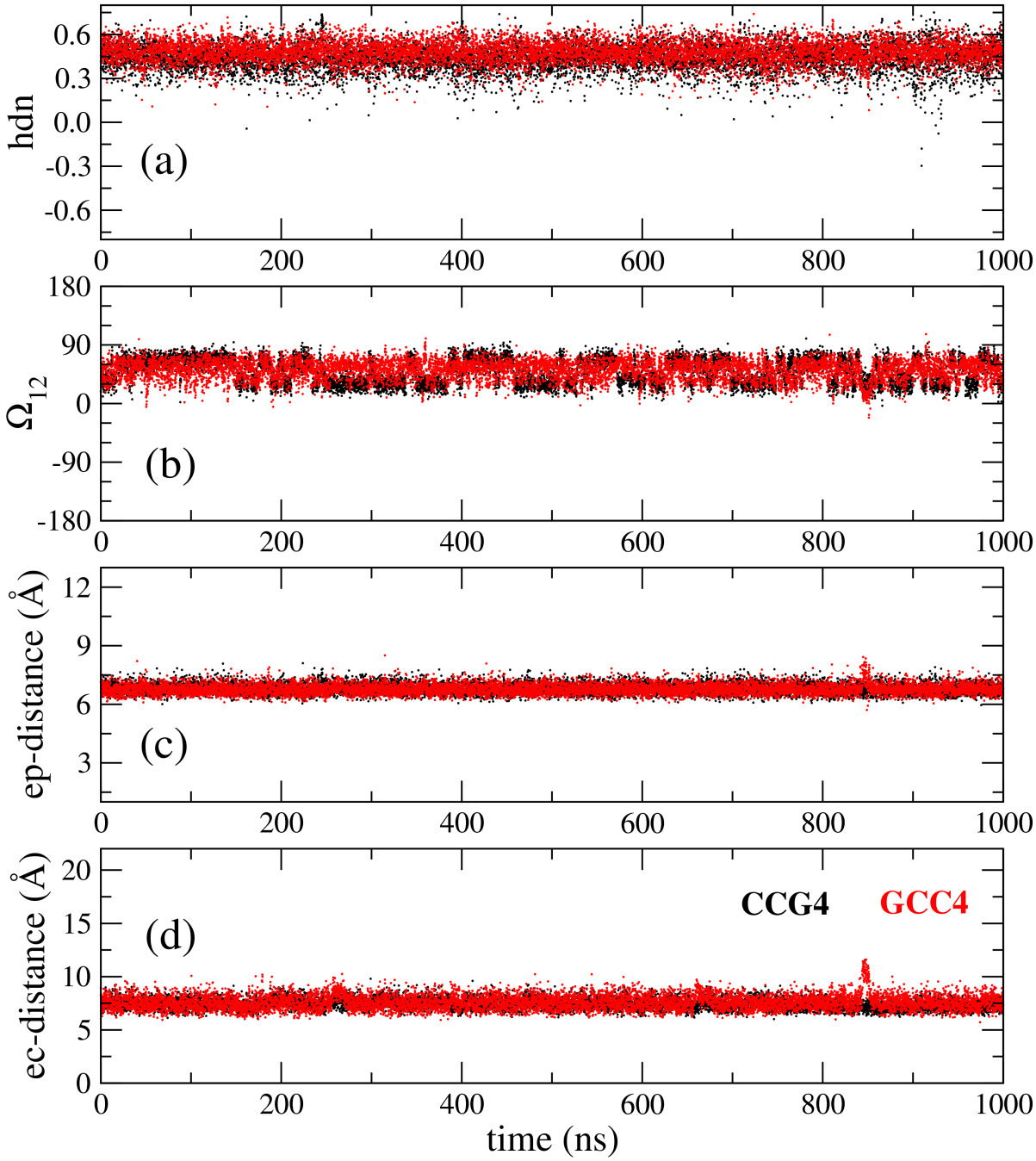


Figure E.1 Time dependence of quantities characterizing the transition to an e-motif. This shows 1 μs simulation under the BSC1 force field for duplexes CCG4 (black) and GCC4 (red). (a) Partial handedness of the C_{12}-C_{17} mismatch; (b) pseudodihedral angle (Ω_{12}) describing the base unstacking of C_{12} with respect to the helical axis; (c) center-of-mass distance (ep-distance) between basepairs 11-16 and 13-18; (d) "e-motif distance" (ec-distance), defined as the distance between the N4 atom of C_{12} and the O2 atom of C_{10} in GCC4 or the N3 atom of G_{10} in CCG4.

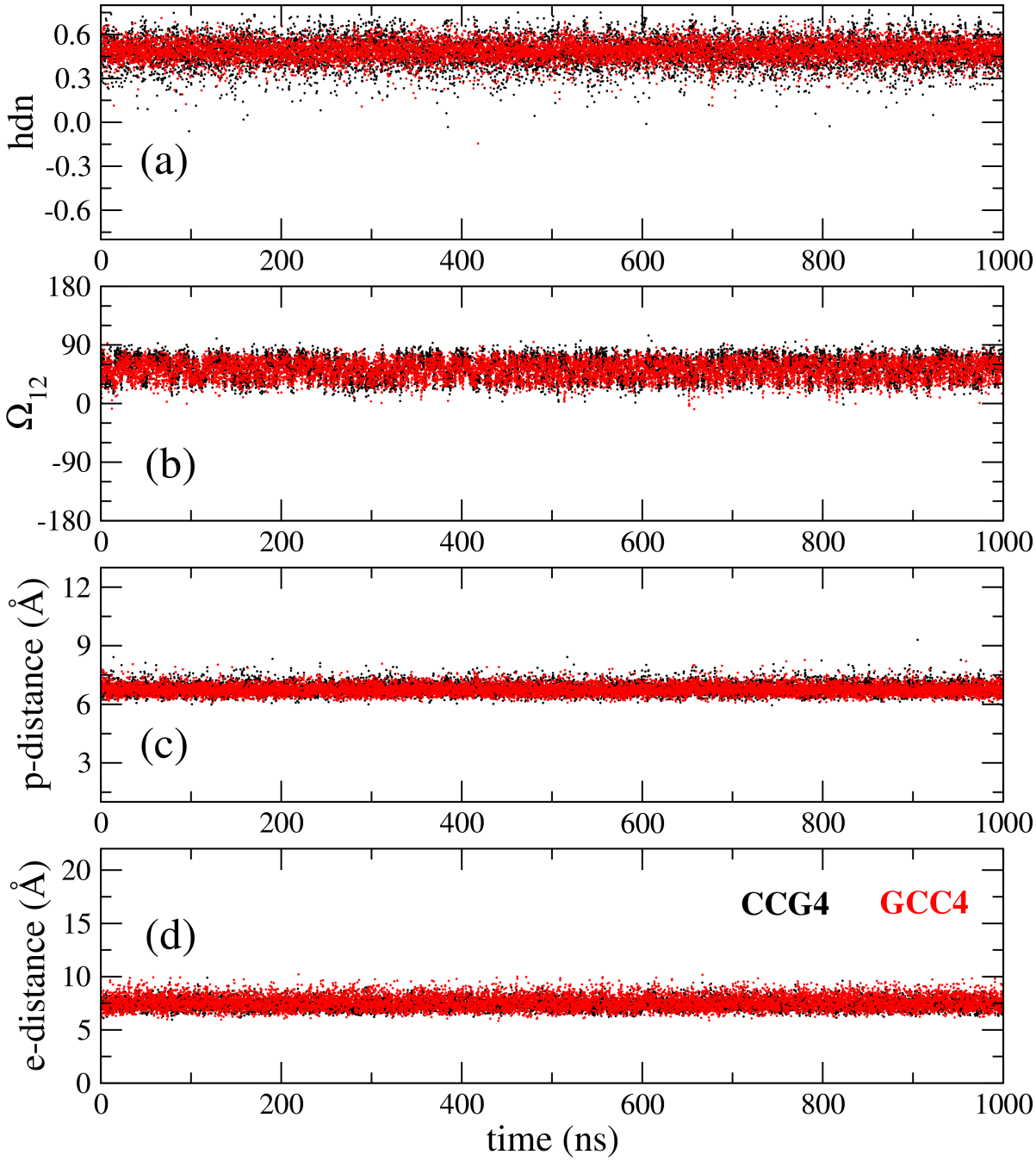


Figure E.2 Time dependence of quantities characterizing the transition to an e-motif. This shows 1 μs simulation under the OL15 force field for duplexes CCG4 (black) and GCC4 (red). (a) Partial handedness of the C_{12}-C_{17} mismatch; (b) pseudodihedral angle (Ω_{12}) describing the base unstacking of C_{12} with respect to the helical axis; (c) center-of-mass distance (ep-distance) between basepairs 11-16 and 13-18; (d) "e-motif distance" (ec-distance), defined as the distance between the N4 atom of C_{12} and the O2 atom of C_{10} in GCC4 or the N3 atom of G_{10} in CCG4.

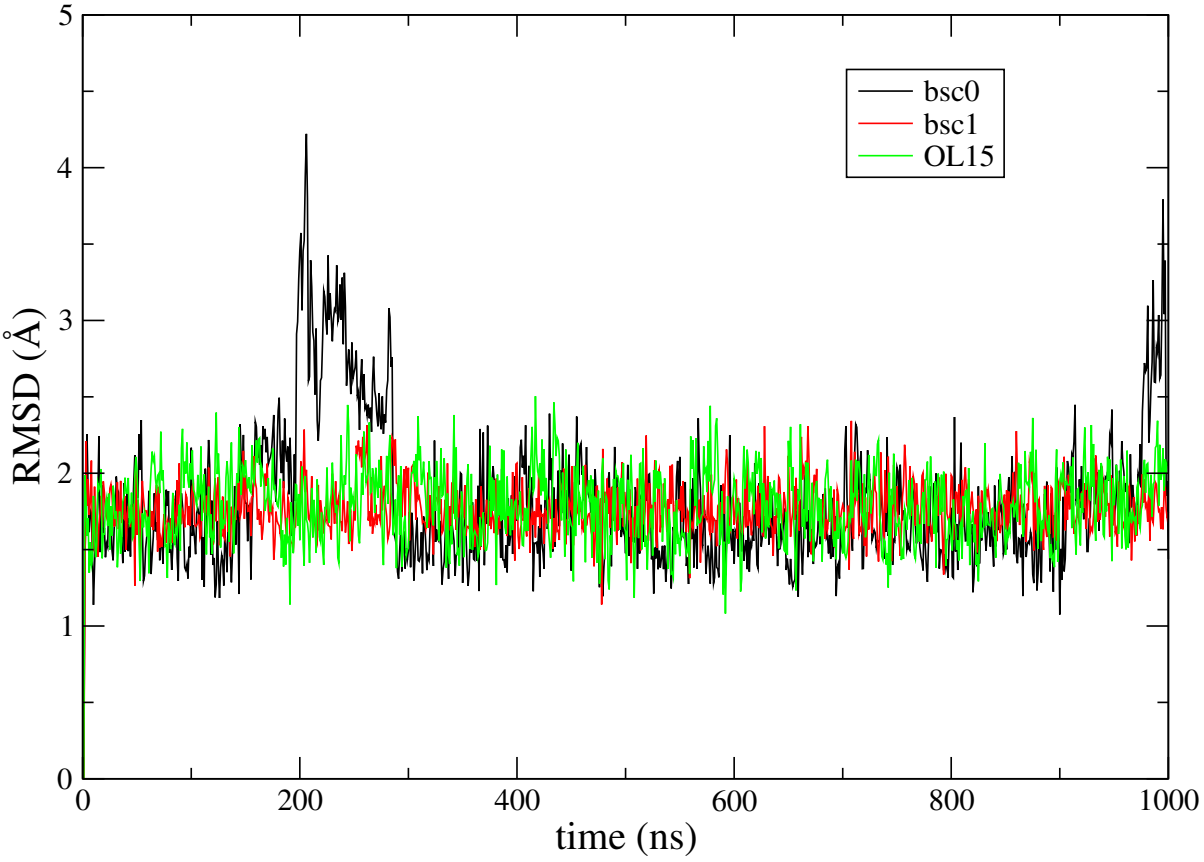


Figure E.3 RMSD of the middle 14 residues around the e-motif (R5-R11 and R20-R26) for the GCC5$_{emotif}$ duplex (Fig.1 (c)) under the three force fields.

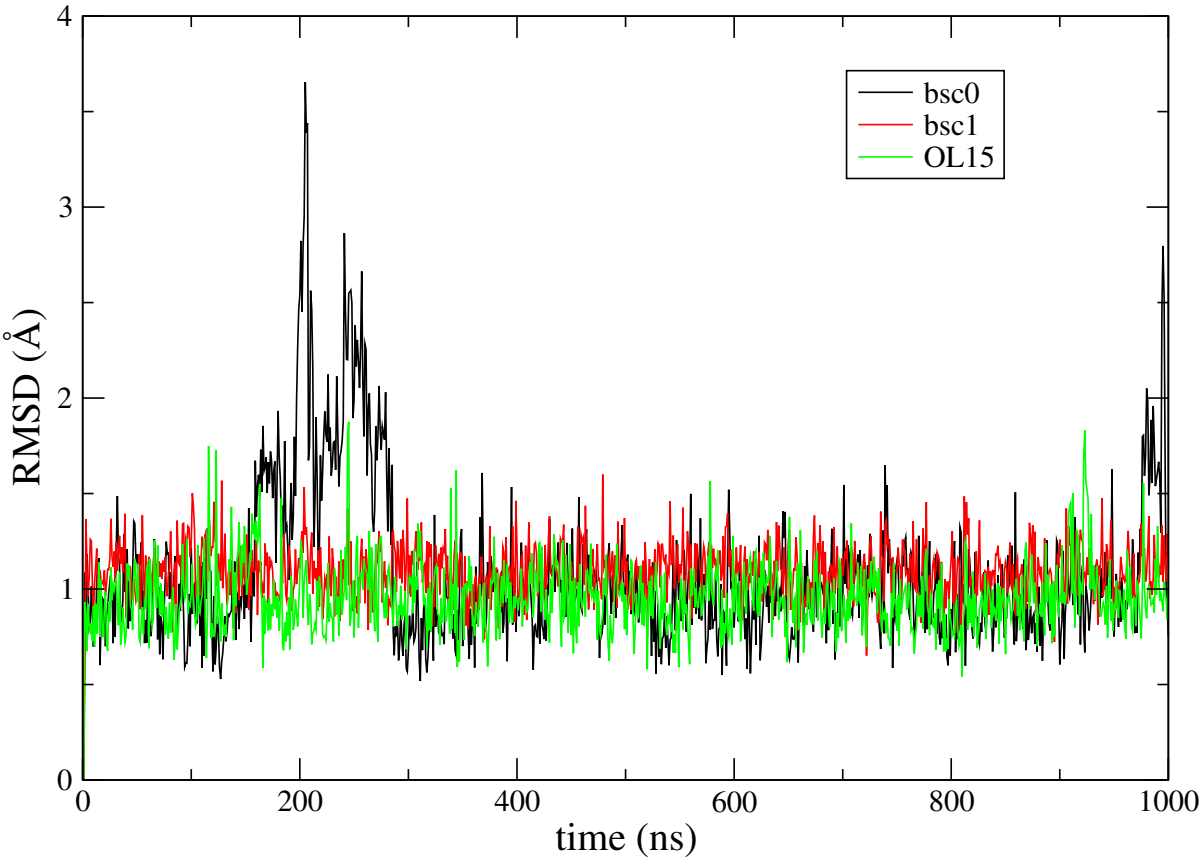


Figure E.4 RMSD of the bases participating in the pseudo GpC step for the GCC5$_{emotif}$ duplex (Fig.1 (c)) under the three force fields.

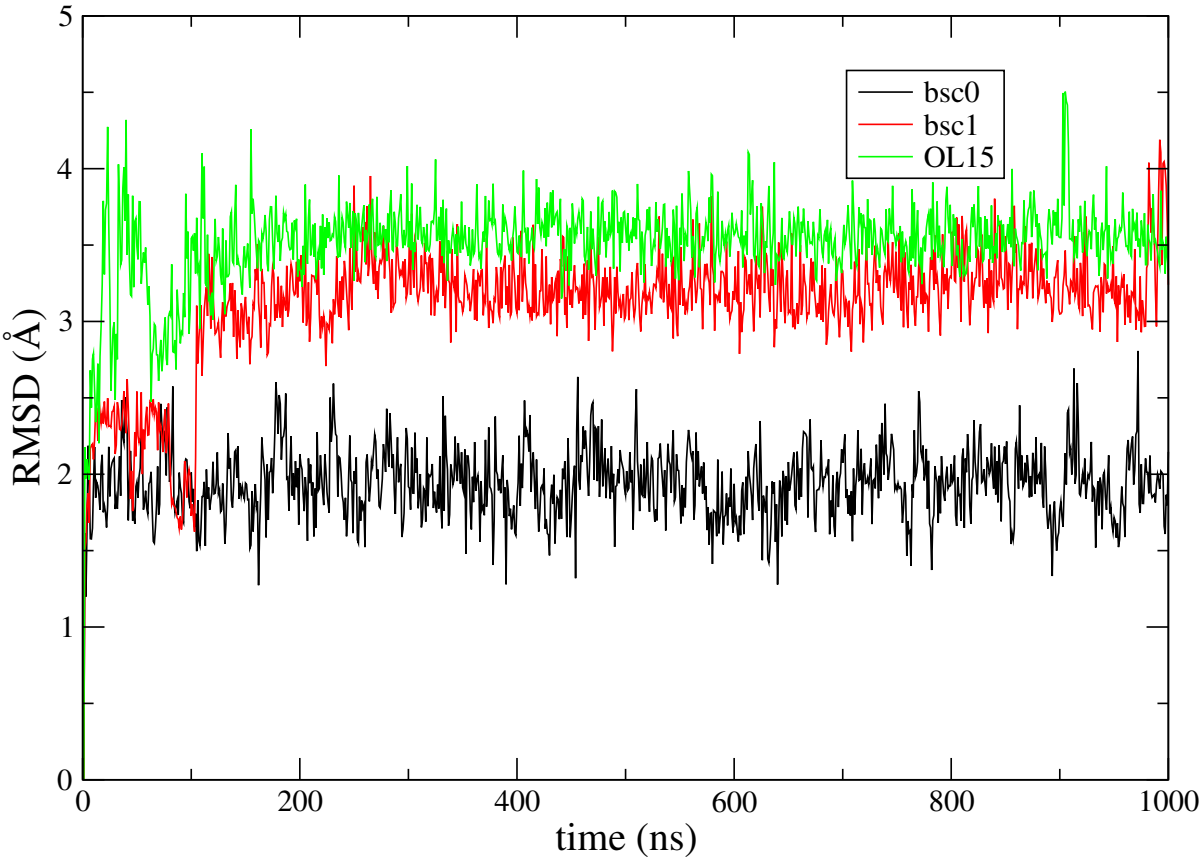


Figure E.5 RMSD of the middle 14 residues around the e-motif (R5-R11 and R20-R26) for the CCG5$_{emotif}$ duplex (Fig.1 (d)) under the three force fields.

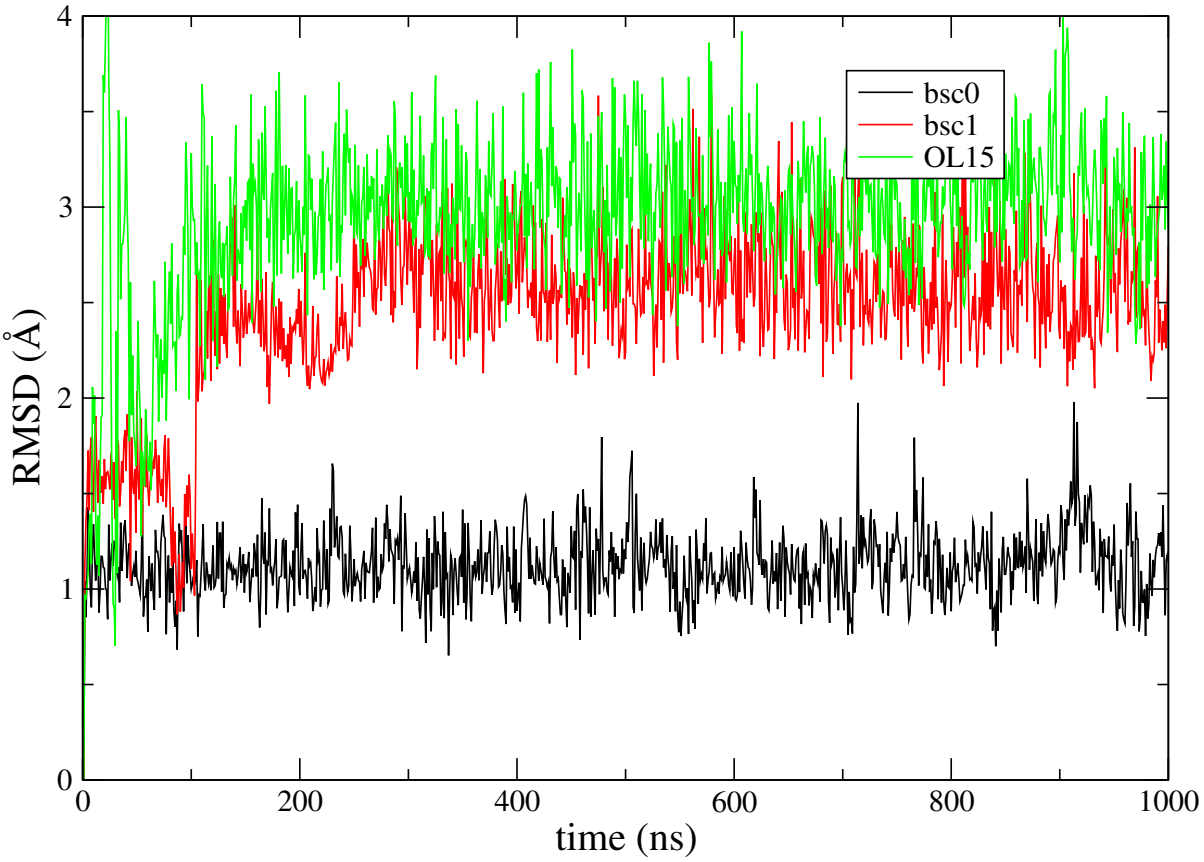


Figure E.6 RMSD of the bases participating in the pseudo CpG step for the CCG5$_{emotif}$ duplex (Fig.1 (d)) under the three force fields.

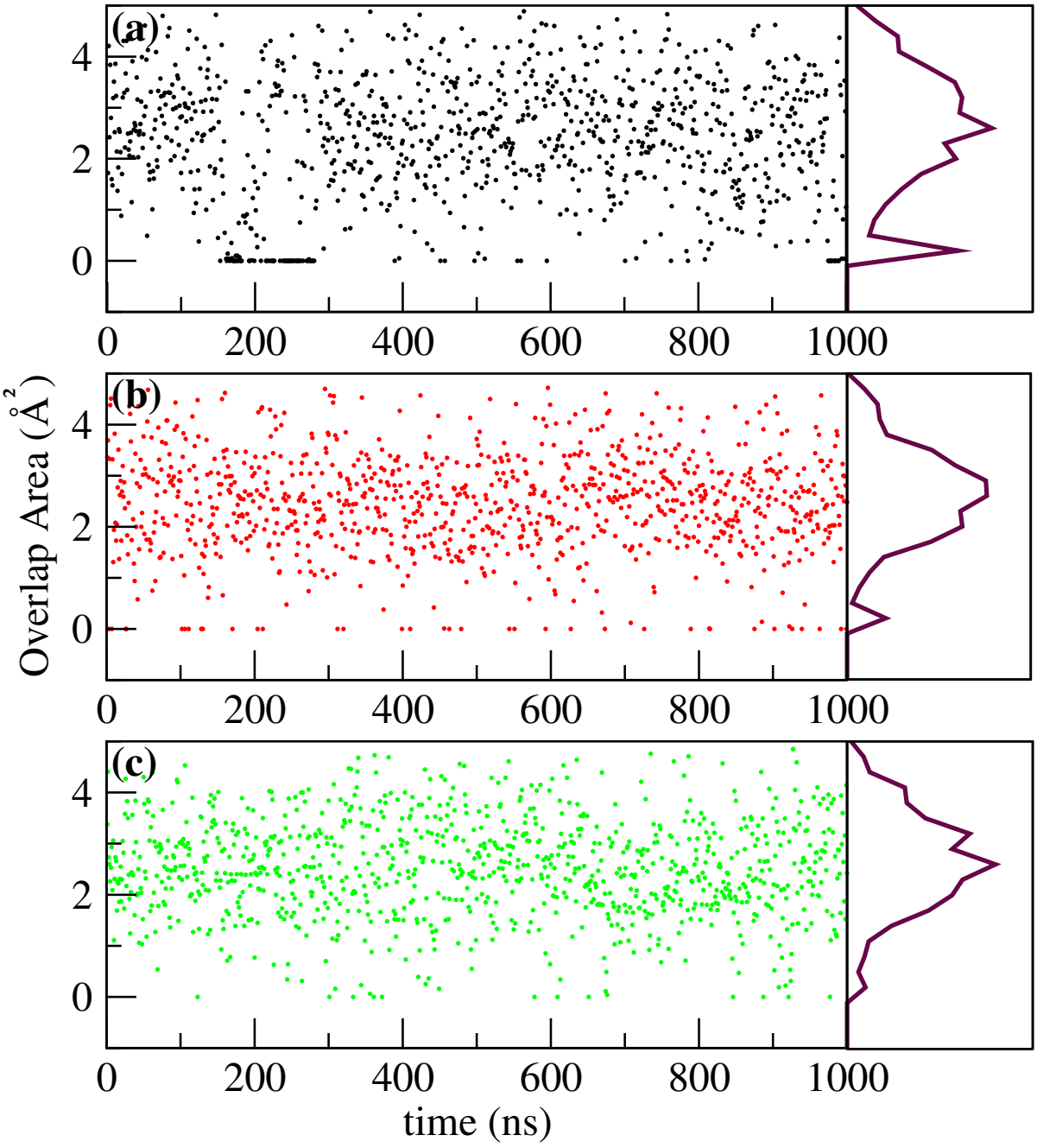


Figure E.7 Overlap areas of the basepair ring atoms of the pseudo GpC step for 1 microsecond run of GCC5$_{emotif}$ in three force fields. (a) BSC0 (black); (b) BSC1 (red); (c) OL15 (green). The subfigures on the right side show the distribution functions.

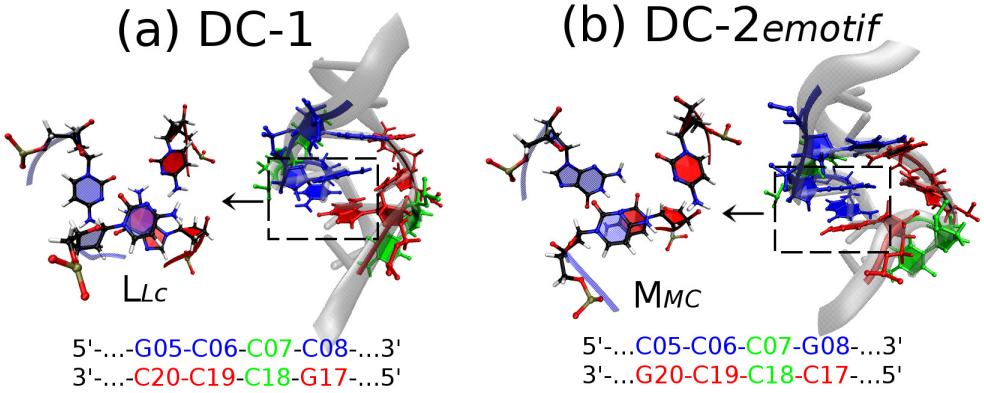


Figure E.8 Pseudo step stacking in an e-motif for hexanucleotide repeats. (a) L_{LC} step in DC-1; (b) M_{MC} step in DC-2_{emotif}. Different bases are drawn in different colors, consistent with the colors in the sequence below. Extra-helical C bases are in green. Left and right subfigures show views perpendicular and parallel to the helical axis, respectively.

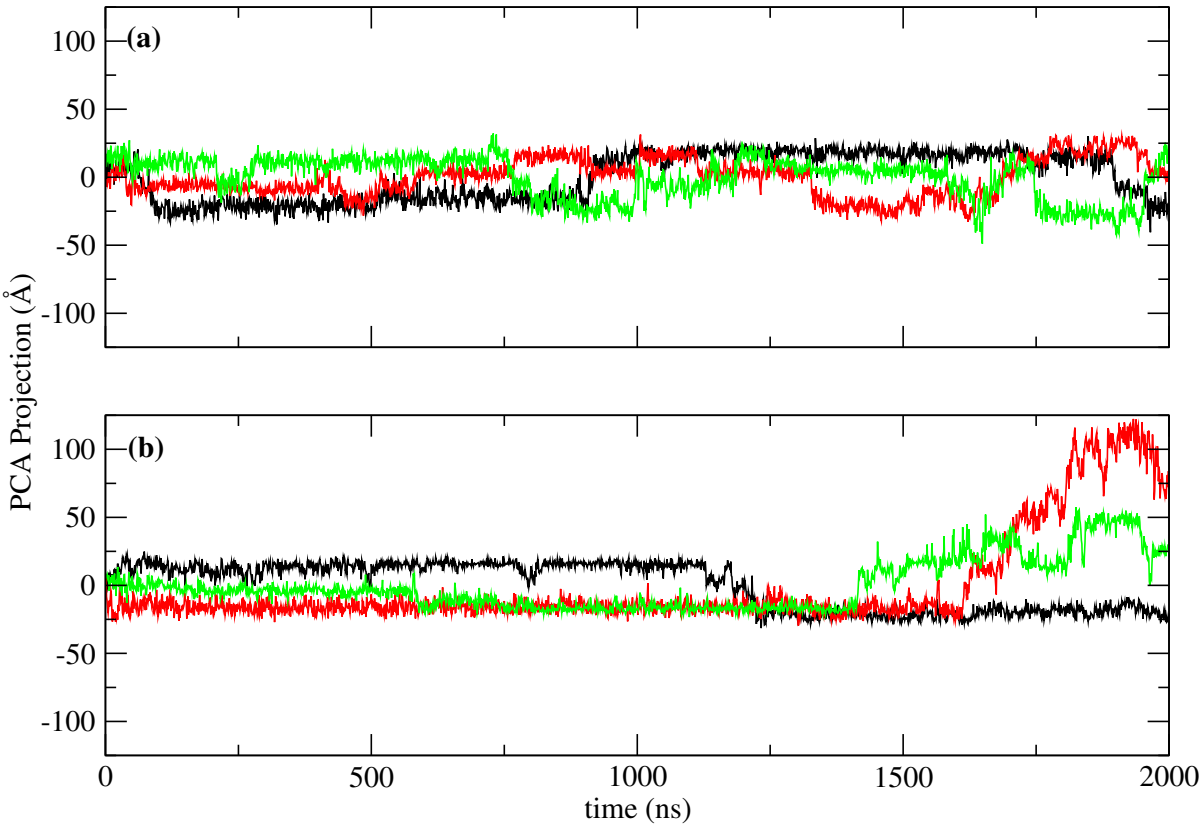


Figure E.9 Projection along the direction of the first eigenvector from a PCA analysis of the middle nucleotides (residues 4-12,18-26). (a) GCC4$_{extended}$ (b) CCG4$_{extended}$. Different colors represent different force fields: BSC0 (black); BSC1(red); OL15 (green). (a) shows ordinary fluctuations in all three force fields while (b) shows that there are conformational transitions in BSC1 and OL15.