Estudio teórico de formas inusuales y modificadas de los ácidos nucleicos

Ignacio Faustino Pló

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- NoComercial – SenseObraDerivada 3.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - NoComercial – SinObraDerivada 3.0. España de Creative* Commons**.

This doctoral thesis is licensed under the ***Creative Commons Attribution-NonCommercial-NoDerivs 3.0. Spain License***.

Universidad de Barcelona

Facultad de Biología

Departamento de Bioquímica

Estudio teórico de formas inusuales y modificadas de los ácidos nucleicos

Ignacio Faustino Pló

2013

PROGRAMA DE DOCTORAT DE BIOMEDICINA
TESIS REALIZADA EN EL INSTITUTO DE INVESTIGACIÓN BIOMÉDICA, BARCELONA

ESTUDIO TEÓRICO DE FORMAS INUSUALES Y MODIFICADAS DE LOS ÁCIDOS NUCLEICOS

Memoria presentada por Ignacio Faustino Pló para optar al título de doctor por la Universidad de Barcelona realizada en el Institute for Research in Biomedicine, Barcelona.

DIRECTOR

Modesto Orozco López

DOCTORANDO

Ignacio Faustino Pló

2013

Contents

List of Figures

List of Tables

1 Introduction

There is a driving force
more powerful than steam,
electricity and atomic energy: the will.
Albert Einstein (1879-1955)

It has been just 70 years since DNA is understood as the main genetic material, the code that makes us who we are. The beginning of the studies on nucleic acids dates back to the nineteenth century when in 1868 Friedrich Miescher found a substance with phosphorus bands found in pus human cells which he called nuclein. However, it was Altman who succeeded in 1889 in isolating a strongly acidic material which he called nucleic acid. In the following years, several chemists helped to discover the main elements and functional groups present in the mononucleotides (A for adenine, C for cytosine, G for guanine and T for thymine) and its linearity in its primary sequence.

In 1953 the joined efforts made by James Watson and Francis Crick (Watson & Crick 1953) using the X-ray fiber diffraction patterns generated by Franklin, Wilkins and colleagues (Franklin & Gosling 1953; Wilkins et al. 1953) led to the determination of the average secondary structure of DNA and opened the way to the most exciting era of modern science. Soon enough after that discovery, Watson found that keto/amino tautomeric forms were able to satisfy Chargaff observations (1950) by pairing adenine with thymine and guanine with cytosine through hydrogen bonds. The semiconservative replication generated by the strand recognition and synthesis of a complementary strand was rapidly verified by Meselson and Stahl (Meselson et al. 1957) confirming double stranded DNA as the carrier of genetic information that is transferred from one generation to the next.

The discovery of messenger and transfer RNAs came to clarify how DNA information is converted into functional proteins. These advances in molecular biology constituted the central dogma which shows the general pathway of information in three steps. First, the replication of DNA assures an identical copy of DNA molecules in the next generation from parental DNA. Second, the transcription process by which some parts of the DNA are copied into RNA. And the third step which is called translation process in which messenger RNA molecules are translated into proteins in the ribosomes.

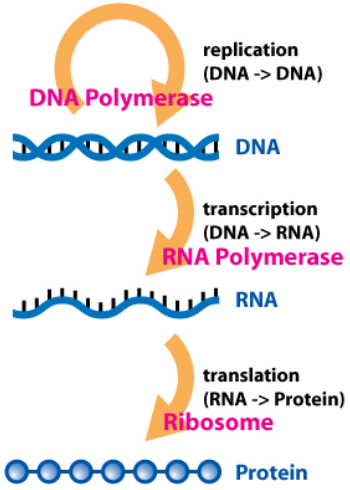


Figure 1.1: Central dogma in molecular biology. The first step, in DNA replication, DNA polymerase generates a copy of genomic DNA. In the second step, RNA polymerase generates RNA copies of specific regions of DNA. And in the third step, proteins are created from RNA molecules in ribosomes during the translation process.

Of the approx. 23000 genes of the human genome, only some of them are expressed in the cell at a time depending on cellular or organism needs. The cellular concentration of proteins is carefully regulated by different mechanisms by controlling both mRNA and protein levels. Transcription regulation is mediated by protein-DNA interactions, involving the recognition of specific sequences called promoters near points at which RNA synthesis starts. Mutations in the nucleotide sequence of promoter regions or in distant enhancer sequences affect the binding affinity of RNA polymerases and transcription factors and decrease promoter functionality.

The accessibility to regulatory sequences is restricted in eukaryotes by the chromatin structure (Luger 2003). The essential part of chromatin is one and a half turn of DNA wrapped around the histone core called nucleosome. The loose DNA regions which, are more accessible to the transcription machinery, constitutes the euchromatin regions while the tightly packed DNA regions which, contain those genes that are

poorly transcribed, form the heterochromatin regions. Among the latter, some regions belong to the constitutive heterochromatin which correspond to less active regions in all cells and, facultative heterochromatin which, contains regions that are active in some cells but not in others and are often associated to differentiation and morphogenesis processes. Hypersensibility studies with DNaseI (Crawford et al. 2006) and MNase (Weiner et al. 2010) have allowed the approximate identification of those accessible regions.

DNA and histone modifications regulate chromatin compactness. For example, acetylation in specific lysine residues by histone acetyltransferases (HATs) activates chromatin for transcription by reducing affinity of histones for DNA. Additionally, 5-methylation of cytosine residues of CpG sequences is related to inactive gene transcription. These mechanisms of gene regulation are known as epigenetics and have a dramatic effect on the individual phenotype (Feinberg 2007) (Figure 1.2). More recently, long non-coding RNA molecules have been found to regulate transcription by guiding the site-specific recruitment of chromatin-modifying enzymes (Koziol & Rinn 2010).

Gene expression is also regulated by small RNA molecules in higher eukaryotes. For example, micro-RNAs (miRNAs) are transcribed as precursor RNAs about 70 nucleotides long forming intramolecular interactions in hairpin-like structures. These precursors are digested by endonucleases which yield short RNA duplexes of 20 to 25 nucleotides long. Dicer, one of the best characterized endonucleases, cuts miRNAs which directs cleavage of homologous mRNA via an RNA-induced silencing complex (RISC) (Lee et al. 2004) leading to degradation of the RNA and inhibition of mRNA expression. The existence of this mechanism called RNA interference (RNAi) (Agami 2002; Hannon 2002; Cerutti 2003) provides an interesting strategy for biotechnological and biomedical applications, since double stranded RNAs can be con structure and introduced into a cell and lead to target mRNA degradation through RISC cleavage.

Although the right-handed double strand B-DNA is the reference structure under physiological conditions, DNA is dynamic and can adopt a number of alternative structures such as hairpins, triplexes, cruciforms, left-handed Z-forms, tetraplexes, A-motifs, etc. (Choi & Majima 2011; Du et al. 2013). Expandable repetitive sequences, which account for more than 50% of the total genomic DNA, can potentially form these non-B-DNA structures. For example, triplexes structures can be form through interaction of a single-strand (triplex-forming oligonucleotide, TFO) with the major groove of the double helix by Hoogsteen or reverse Hoogsteen hydrogen bonding interactions with purine bases of the Watson-Crick base pairs. The formation of these multistrand structures offers a sequence specific control of gene expression through antigene strategy

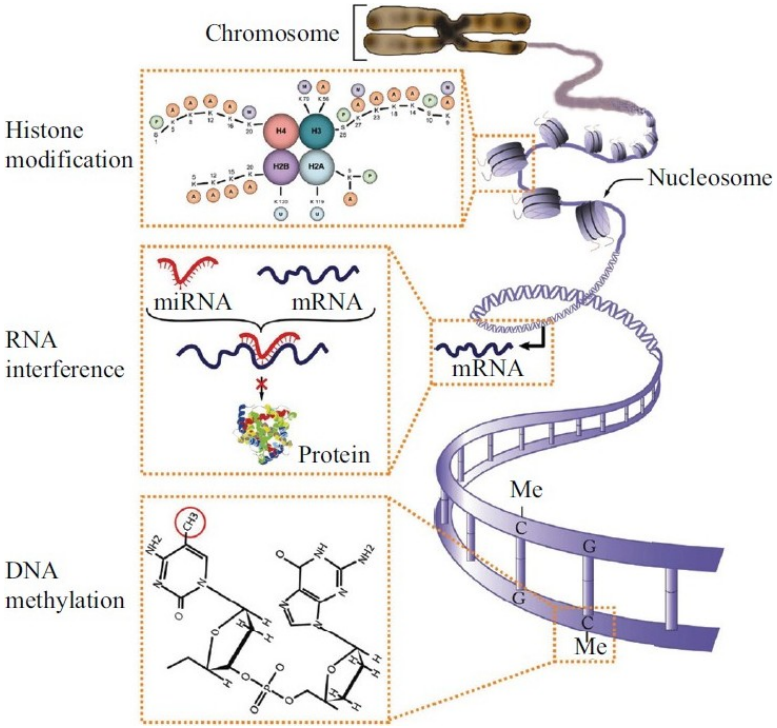


Figure 1.2: Schematic of the mechanisms of epigenetic regulation. DNA methylation, histone modifications, and RNA-mediated gene silencing constitute three distinct mechanisms of epigenetic regulation. DNA methylation is a covalent modification of the cytosine (C) that is located 5' to a guanine (G) in a CpG dinucleotide. Histone (chromatin) modifications refer to covalent post-translational modifications of N-terminal tails of four core histones (H3, H4, H2A, and H2B). The most recent mechanism of epigenetic inheritance involves RNAs. (From Sawan et al. 2008).

(Hélène et al. 1992) and might be formed in living cells (Wang et al. 1995; Majumdar et al. 1998; Vasquez & Glazer 2002; Goñi et al. 2006).

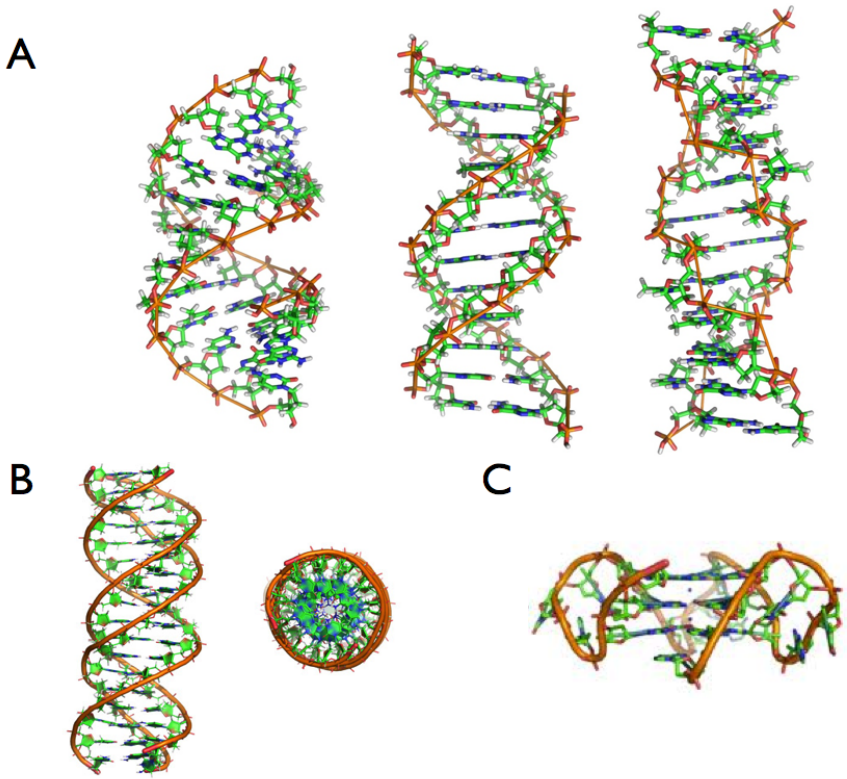


Figure 1.3: A-DNA (right), B-DNA (center) and Z-DNA polymorphisms (A), DNA triplex showing the third strand in the major groove of the double strand of the DNA (B), and G-quadruplex with sodium ions in the central channel (C).

G-quadruplexes have been found in telomeric DNA (chromosome endings) attracting remarkable attention in the last decade. From the sequence point of view, these regions are clearly guanine-riched and have a tendency to form unusual supramolecular structures through hydrogen bonding interactions. Initially thought to have little relevance, such structures have been discovered to form *in vivo* (Biffi et al. 2013) in human telomeric DNA. Currently, the study is focused on the stabilization of G-quadruplexes by binding of small molecule ligands since quadruplex formation inhibits effectively telomerase activity, an enzyme which is activated in tumor cells to evade death cell and stabilizing telomere length. Additionally, the structural diversity in folded topologies available to guanine-rich nucleic acid repeat sequences have made four-stranded G-quadruplex structures the focus of both basic and applied research, from cancer biology and novel therapeutics through to nanoelectronics (Collie & Parkinson 2011).

So far, epigenetic modifications of DNA and post-transcriptional modified RNA

molecules have shown to have crucial role in transcriptional regulation (Sabin et al. 2013). In addition, modifications of nucleic acids can be used for very interesting applications such as nucleic acid sequencing and labelling, in antigene or in antisense strategies (Blackburn et al. 2006). These modifications can be placed both in the backbone or nucleobase moieties (Robles et al. 2002; Blackburn et al. 2006), to improve specificity and affinity while creating nuclease-resistant oligos to silence the expression of undesired overexpressed genes. Backbone modifications, such as peptide nucleic acids (PNAs) (Nielsen et al. 1991; Egholm et al. 1993), have shown the power of triplex-based technology in biochemical, genetic and medical applications by improving on one hand nuclease resistance, and on the other hand, by enhancing triplex stability. In addition, modified nucleosides have been used as antimetabolites, compounds that prevent the DNA synthesis, and therefore, used as anti-cancer drugs (Christopherson et al. 2002; Galmarini et al. 2002). For example, oligonucleotides containing 6-thioguanine have been used extensively for studying RNA and DNA-protein interactions since the thioketocarbonyl group is a weaker hydrogen bond acceptor than a carbonyl group. Besides this, 6-thioguanine absorbs a long wavelength that allows to photoactivate specifically the formation of covalent cross-links for studying 3D structures (Verma & Eckstein 1998).

The physical description of DNA (structure and dynamic properties) is not easy to derive form the sequence, being necessary to use structural biophysical methods, mainly X-ray diffraction and nuclear magnetic resonance (NMR). The first allows to analyze the crystal structure, while the second allows to explore accessible conformations in experimental conditions. Interactions between different molecules and DNA involve specific contacts that can be studied by means of these experimental techniques. However, DNA has sequence-dependence features, such as groove dimensions, flexibility, intrinsic curvature, etc., which do influence in the interaction mode with other molecules. These physical properties that depend on the sequence are known as indirect recognition, and in some cases this becomes more important than direct recognition.

Then, not only the structure is important for functionality, but much of this is explained by conformational changes. Therefore, it is essential to consider not only the structure but also the dynamic behavior of nucleic acids, which are not easily accessible by experimental techniques. A theoretical tool that complements very well the experimental data is the molecular dynamics (MD), which explores conformational space along time (Levitt 1983; Orozco et al. 2003). Through these studies it is possible to study in atomic detail the dynamics of nucleic acid in a solvation environment (ions and water). The study of structural and dynamic aspects allows, among other applications,

simple models extrapolated to predict the characteristics of different sequences and its affinity to adopt certain types of structures.

1.1 Objectives of this PhD

In this thesis quantum mechanics and classical mechanics techniques have been used to study the tautomeric properties of modified nucleobases with potential application to antisense and biotechnology fields. On the other hand, the study of the mechanical properties of double stranded DNA and RNA (dsDNA versus dsRNA) and their sequence dependency have been addressed.

In practice, the work has been divided into the following sections:

1.1.1 Theoretical study of sulfur modified thymines

1. Study of tautomeric preferences of 2- and 4-thioketothymines both in gas phase and in solvated environments.

2. Determination of the structural influence of the most stable tautomeric forms for 2- and 4-thioketothymines in dsDNA structures with guanine or adenine as complementary base.

3. Determination of the thermodynamic preferences for T $\rightarrow$ S (S = 2- or 4-thioketothymine) mutations with respect of the complementary base.

4. Determination of the role of minor tautomeric forms in the G·S mismatch.

1.1.2 Theoretical study of seleno modified guanine

1. Study of tautomeric preferences of 6-selenoguanine (6SeG) both in gas phase and in solvated environments.

2. Determination of the structural influence of 6-selenoguanine in double, triple stranded DNA and in G-quadruplex structures.

3. Determination of relative stabilities for the G $\rightarrow$ 6SeG mutation in the DNA systems aforementioned and their comparison with available experimental data.

4. Exploring the conductimetric properties of 6-selenoguanine.

1.1.3 Inhibition of 3'-exonucleases with dimeric N-ethyl-N modified nucleosides

1. Structural study of inhibition of 3'-exonucleases family for the improvement of antisense strategies with modified siRNAs.
2. Experimental validation of these N-ethyl-N modified nucleosides both with nuclease digestion assays and cellular RNA activity measurements.

1.1.4 Flexibility study of dsRNA molecules

1. Determination of the consensus achieved by the two most used nucleic acids force fields (AMBER and CHARMM).
2. Characterization of the sequence-dependence properties for dsRNA molecules.
3. Evaluation of the mechanical properties of dsRNA and comparison with dsDNA.

1.2 References

Agami, R. 2002. RNAi and related mechanisms and their potential use for therapy. Curr. Opin. Chem. Biol. 6, 829–834.

Blackburn, G.M. et al., 2006. Nucleic acids in chemistry and biology Royal Society of Chemistry, Royal Society of Chemistry.

Biffi, G. et al., 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. Nature Chemistry.

Cerutti, H. 2003. RNA interference: traveling in the cell and gaining functions? Trends Genet. 19, 9–46.

Chargaff, E., 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia, 6(6), pp.201–209.

Choi, J. & Majima, T., 2011. Conformational changes of non-B DNA. Chemical Society Reviews, 40(12), p.5893.

Collie, G.W. & Parkinson, G.N., 2011. The application of DNA and RNA G-quadruplexes to therapeutic medicines. Chemical Society Reviews, 40(12), pp.5867–5892.

Crawford, G.E. et al., 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Research, 16(1), pp.123–131.

Du, X. et al., 2013. The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in Escherichia coli. Nucleic Acids Research, 41(12), pp.5965–5977.

Egholm, M. et al., 1993. PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. Nature, 365(6446), pp.566–568.

Feinberg, A.P., 2007. Phenotypic plasticity and the epigenetics of human disease. Nature, 447(7143), pp.433–440.

Franklin, R.E. & Gosling, R., 1953. Molecular Configuration in Sodium Thymonucleate. Nature, 171(4356), pp.740–741.

Goñi, J.R. et al., 2006. Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. BMC Genomics, 7(1), p.63.

Hannon, G.J. 2002. RNA interference. Nature 418, 244–251.

Hélène, C., Thuong, N.T. & Harel-Bellan, A., 1992. Control of gene expression by triple helix-forming oligonucleotides. The antigene strategy. Annals of the New York Academy of Sciences, 660, pp.27–36.

Koziol, M.J. & Rinn, J.L., 2010. RNA traffic control of chromatin complexes. Current Opinion in Genetics & Development, 20(2), pp.142–148.

Lee, Y.S. et al., 2004. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. Cell, 117(1), pp.69–81.

Levitt, M. (1983). Computer simulation of DNA double-helix dynamics. Cold Spring Harb Symp Quant Biol 47 Pt 1, 251-62.

Luger, K. 2003. Structure and dynamic behavior of nucleosomes. Curr. Opin. Genet. Dev. 13, 127–135.

Majumdar, A. et al., 1998. Targeted gene knockout mediated by triple helix forming oligonucleotides. Nature Genetics, 20(2), pp.212–214.

Meselson, M., Stahl, F.W. & Vinograd, J., 1957. Equilibrium sedimentation of macromolecules in density gradients. Proceedings of the National Academy of Sciences of the United States of America, 43(7), pp.581–588.

Nielsen, P.E. et al., 1991. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. Science, 254(5037), pp.1497–1500.

Orozco, M., Perez, A., Noy, A. & Luque, F. J. (2003). Theoretical methods for the simulation of nucleic acids. Chem Soc Rev 32, 350-64.

Sabin, L.R., Delás, M.J. & Hannon, G.J., 2013. Dogma derailed: the many influences of RNA on the genome. Molecular Cell, 49(5), pp.783–794.

Sawan, C. et al., 2008. Epigenetic drivers and genetic passengers on the road to cancer. Mutation research, 642(1-2), pp.1–13.

Vasquez, K.M. & Glazer, P.M., 2002. Triplex-forming oligonucleotides: principles and applications. Quarterly Reviews of Biophysics, 35(01).

Verma, S. & Eckstein, F., 1998. Modified oligonucleotides: synthesis and strategy for users. Annual review of biochemistry, 67, pp.99–134.

Wang, G. et al., 1995. Targeted mutagenesis in mammalian cells mediated by intracellular triple helix formation. Molecular and cellular biology, 15(3), pp.1759–1768.

Watson, J.D. & Crick, F.H., 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171(4356), pp.737–738.

Weiner, A. et al., 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Research, 20(1), pp.90–100.

Wilkins, M.H.F., Stokes, A.R. & Wilson, H.R., 1953. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. Nature, 171(4356), pp.738–740.

2

Methods for Nucleic Acids Modeling

I am one of those who think like Nobel,
that humanity will draw more good
than evil from new discoveries.
Marie Curie (1867-1934)

Biomolecular modeling and simulation has progressed focusing in the study of molecular structures and their functions supported on model building and the latest improvements in computation. Nowadays, the computational approach allow us to address part of the complex biological questions. The computational methods include *ab initio* and semi-empirical quantum mechanics (QM), docking, molecular dynamics (MD), Monte Carlo (MC), homology modeling, free energy and solvation methods, enhanced sampling and pathways methods, structure/activity relationships (SAR) and many other methods. The choice of one over the other depends on the purpose of the study, the size of the considered system, the computer resources available and the accuracy required for such study and in practice, on the resources available. During the course of this work a wide variety of these methods have been used.

2.1 Quantum Mechanics (QM)

Quantum mechanical techniques allow the study of the structure, chemical reactivity, and electronic properties by solving in principle the Schrödinger equation. The objective of the quantum methodology is the interpretation and the prediction of the molecular structure and the chemical reactivity from the quantum formulation of the molecular physics. Although these techniques where initially developed for the study of small systems in gas phase, nowadays they have been successfully applied to fairly

large molecules and their use in molecular biology is continuously increasing. There are many observables that can be derived from QM calculations. Thermodynamic properties, molecular spectra, relative stability estimations between conformers or molecular properties of transition states in chemical reactions represent only some of those applications. The accuracy of the results usually depends on the choice of the level of theory. This section will briefly describe the theoretical basis, methods and applications of QM techniques.

The first principles on which modern physics are based state that a system is described by their wavefunction (Ψ) in which all the needed information from the system is included. So the energy and other related properties of a molecule may be obtained by solving the Schrödinger equation:

$$\hat{H}\Psi(x) = E\Psi(x) \tag{2.1}$$

where $\hat{H}$ is the hamiltonian operator and E is the energy of the system. The hamiltonian operator includes the kinetic energy of electrons and nuclei and the electron-electron, nucleus-nucleus and nucleus-electron interaction potentials. Since the exact solution for this equation is only possible for a molecule with no more than one electron, approximations to the solution have to be applied. Usually a compromise must be taken between accuracy and speed. Among the different approximations, the most important is the Born-Oppenheimer (BO) approximation that separates the motion of electrons and nuclei, that is, in a molecule the nuclei are essentially stationary compared to the electrons. This assumption simplifies the application of the Schrödinger equation to molecules by allowing us to focus on the electronic energy and add in the nuclear repulsion energy later. Other approximations differentiate on the Hamiltonian, the method used to solve the system equation and the function obtained.

Based on these approximations, the different methods can be classified as *ab initio*, semiempirical and based on the density functional theory:

- *ab initio* methods are based on the formalism of Quantum Mechanics from first principles without additional parameters. They include the Hartree-Fock method and those that include electron correlation e.g., configuration interaction (CI) methods, based on Møller-Plesset perturbation theory (MP), and coupled cluster (CC) methods.

- *semiempirical* methods which apply parameters derived from experimental data to simplify the calculation of an approximate form of the same Schrödinger equation.

- *density functional theory* (DFT) based methods that meet the functional energy directly from the electron density using Kohn-Sham equations and avoids the wavefunction calculation.

2.1.1 *Ab initio* methods

2.1.1.1 Hartree-Fock method

The Hartree-Fock method (here after HF) is important to understand the following methods which chronologically came after it and is based on the variation principle. This principle states that the HF energy is necessarily above the energy which would result upon solution of the Schrödinger equation. In HF approach, individual electrons are confined to functions termed molecular orbitals (MO), each of which is determined by assuming that the electron is moving within an average field of all the other electrons. Every molecular orbital (χ_i), termed as a spin orbital, is the product of a spatial function (φ) and a spin function, (α or β). The total wavefunction is written as an antisymmetrized product (the Slater determinant) of one-electron wavefunctions:

$$\chi_i(q_1) = \varphi_i(r_1)\alpha(\omega_1) \quad \text{or} \quad \chi_i(q_1) = \varphi_i(r_1)\beta(\omega_1) \tag{2.2}$$

$$\Psi(q_1, q_2, ..., q_N) = |\chi_i(q_1), \chi_j(q_2), ..., \chi_k(q_N)\rangle \tag{2.3}$$

This determinant places N electrons in N spin orbitals, but it does not specified which one of the $N!$ possible electronic combinations describes the polyelectronic system thus, each of them contributes to the wavefunction. According to the Pauli exclusion principle, the determinant is cancelled if there are more than one electron in one spin orbital.

The solution of the Schrödinger equation is obtained by a process referred to as a self-consistent-field (SCF) procedure. All SCF procedures lead to equations of the form:

$$f(i)\chi(q_i) = \epsilon\chi(q_i) \tag{2.4}$$

where the Fock operator *f(i)* for a system with M nuclei can be written as:

$$f(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \upsilon^{eff}(i) \tag{2.5}$$

where Z_A is the nuclear charge of atom A, r_{iA} is the distance between electron i and nucleus A. The first term corresponds to the kinetic energy of electron i, the second term corresponds to the interaction between the nuclei and the third term describes the electronic repulsion as a effective potential created by the rest of electrons. The

nature of the effective potential v^{eff} depends on the SCF methodology.

In practice, each MO can be expressed as a linear combination of a finite set (basis set) of prescribed functions known as basis functions, ϕ:

$$\psi = \sum_n c_n \phi_n \tag{2.6}$$

where ϕ_n are the valence atomic orbitals and c_n are the unknown molecular orbital coefficients. Since ϕ_n are usually centered at the nuclear positions, they are referred to as atomic orbitals and thus, the equation 2.6 is termed as the linear combination of atomic orbitals (LCAO) approximation.

The Hartree-Fock and LCAO approximations, taken together and applied to the HF equation, lead to the Roothaan-Hall (Roothaan 1951) equations:

$$FC = \epsilon SC \tag{2.7}$$

where ϵ are orbital energies, C are the coefficients of the molecular orbitals, S is the overlap matrix, and F is the Fock matrix, which is analogous to the Hamiltonian in the Schrödinger equation. As the basis function gets bigger, the resulting energy tends to the so-called Hartree-Fock limit.

Basis functions The basis set is the set of mathematical functions from which the wavefunction is built (equation 2.6). In 1930, John C. Slater was the first to calculate orbitals using basis sets. The functions, called Slater orbitals (STO, Slater-type orbital), are described by the function depending on spherical coordinates:

$$\phi_i(\alpha, n, l, m; r, \theta, \phi) = N r^{n-1} e^{-\alpha r} Y_{l,m}(\theta, \phi) \tag{2.8}$$

where N is the normalization constant, α is the orbital exponent, r is the radius in angstroms, θ and ϕ are spherical coordinates, and $Y_{l,m}(\theta, \phi)$ is the angular momentum part. The n, l, and m are quantum numbers: principal, angular momentum, and magnetic, respectively.

In the 1950s, Frank Boys (Boys 1950) suggested a modification to the wavefunction by introducing Gaussian type functions, which contain the exponential $e^{-\alpha r^2}$, rather than the $e^{-\alpha r}$ of the STOs. Besides this, the expression of the angular momentum is replaced by a simple function of cartesian coordinates and the pre-exponential factor r^{n-1} of the STO function is dropped in the GTO function. Such functions are easier to

evaluate than STOs functions. A cartesian Gaussian centered can be represented as:

$$g(\alpha, l, m, n; x, y, z) = Ne^{-\alpha r^2} x^l y^m z^n \tag{2.9}$$

where i, j, and k are nonnegative integers, α is a positive orbital exponent, x_a, y_a, and z_a are Cartesian coordinates with the origin at the nucleus and N is the cartesian Gaussian normalization constant. However, since GTO functions deviate from hydrogen-like functions in areas very close to and far away from the atomic nucleus, a linear combination of GTOs is usually used to describe each STO. In this direction, Pople and coworkers (Hehre 1969) optimized the parameters (exponents and coefficients) to better reproduce STOs functions as a linear combination of GTOs for a wide variety of atoms.

The wavefunction can be described using different sets of basis functions. In a minimal basis set, one basis function is selected for every atomic orbital that is required to describe the free atom while considering two or three basis functions will lead to double-zeta (DZ) or triple-zeta (TZ) basis sets, respectively. The split-valence (SV) basis sets, introduced by Pople and coworkers in 1970s (Hariharan & Pople 1973; Hehre et al. 1986), use one function for orbitals that are not in the valence shell and 2 functions for those in the valence shell. In DZ, which is normally treated as the general split valence basis sets, each atomic orbital function is split up into two basis functions. A better description of the long-range tails of the orbitals can be obtained by adding polarized (p- or d-type basis functions that are added to describe the distortion of s or p orbitals, respectively) and diffuse functions for molecules with heteroatoms, anions, and electronically excited molecules.

Other basis sets that are used for post HF calculations (following section) include shells of polarization functions (d, f, g, etc.) that can yield convergence of the electronic energy to the complete basis set limit. Examples of these correlation-consistent basis sets are cc-pVDZ, cc-pVTZ, and cc-pVQZ (Dunning 1989) which correspond to correlation consistent polarized valence doble, triple, and quadruple zeta, respectively.

2.1.1.2 Post Hartree-Fock methods

Although Hartree-Fock theory gives reasonable values for many purposes, there are others, like relative energies or even geometry optimizations, for which electron correlation is needed, i.e., the basic approximation of HF method of *independence* of electronic movement is not acceptable. Methods that do include electronic correlation can be mainly classified into three groups: those based on the Møller-Plesset perturbation theory (MPn, where n stands for the order of correction), configuration interaction (CI) methods, and those based on coupled cluster theory (CC):

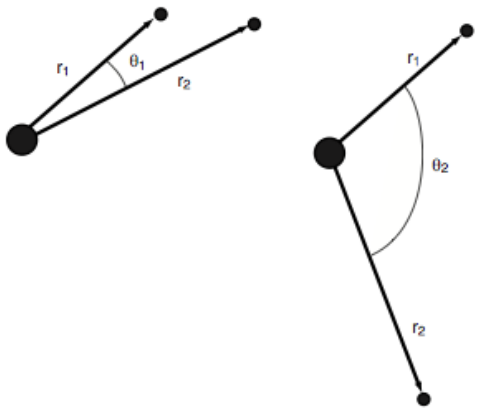


Figure 2.1: Two schematic situations for a nucleus (big point) and the electrons around (small points). The lack of electron correlations within HF theory would not distinguish between these two situations.

1. **Møller-Plesset Perturbation Theory based methods**: MPn methods do treat real molecules as a perturbed Hartree-Fock system and include electron correlation as if a disturbance or perturbation of the HF wavefunctions (Møller & Plesset 1934; Binkley & Pople 1975). The second-order Møller-Plesset (MP2) energy can be written as the sum of the HF energy or (MP1 energy) and a correction term that accounts for the electronic interaction:

$$\hat{H} = \hat{H}_0 + \lambda \hat{V} \tag{2.10}$$

where $\hat{V}$ is the perturbation and λ is a dimensionless parameter which takes a value from 0 to 1. Therefore, the exact wavefunction and energy equations can be written in terms of the Hartree-Fock wavefunctions and energies:

$$E = E^0 + \lambda E^1 + \lambda^2 E^2 + \lambda^3 E^3 + \ldots \Psi = \Psi_0 + \lambda \Psi^1 + \lambda^2 \Psi^2 + \lambda^3 \Psi^3 + \ldots \tag{2.11}$$

where E^n and Ψ^n terms correspond to the n-corrections for the energy and the wavefunction respectively. The sum of E^0 and E^1 terms will result in the HF energy itself so the next terms correspond to the second, third, fourth-corrections and so on. It must be said that, although MPn methods lead to good results, they do not follow the variational principle and it is not uncommon to get lower energies than the exact value with e.g., MP2 calculations.

2. **Configuration interaction method**: This method is based on the idea of

adding to the HF wavefunction the corresponding HF wavefunctions that represent the promotion of electrons from the occupied molecular orbitals to the virtual ones (Shavitt 1998). The different electronic configurations that represent each of the different excited states are contained in the actual wavefunction, and therefore, it can be thought as the resulting of the interactions between the different configurations. A full CI with a large basis set calculation will yield the exact energies of all electronic states, the ground and many of the excited states. Since this kind of calculation will require a high computational cost even for medium size systems, the truncation to singles and doubles (CISD) is accepted as a balanced solution between accuracy and CPU costs. But the most popular implementation of the CI method are probably multiconfigurational SCF (MCSCF) and its variant complete active space SCF (CASSCF), and the coupled-cluster (CC) and related quadratic CI (QCI) methods.

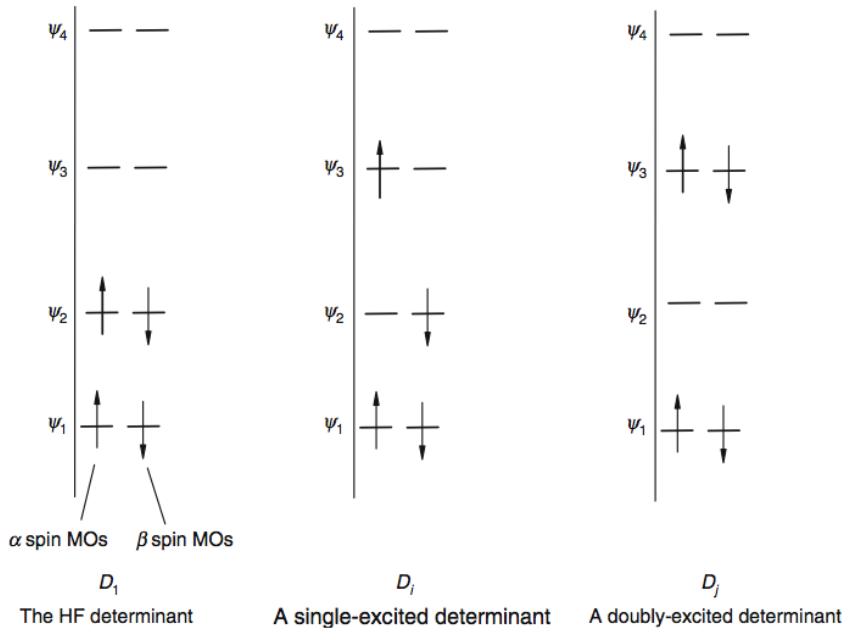


Figure 2.2: Configuration interaction (CI): schematic plot of the promotion of electrons from the occupied MOs (corresponding to the Hartree-Fock determinant). The result corresponds to several wavefunctions, one to the ground state wavefunction and the rest to excited states (From (Lewars 2011)).

3. **Coupled cluster methods**: These correlation methods use linear combinations of many determinants for the inclusion of the electronic correction and in this particular, they are very similar to the methods of configuration interaction (CI). They include several orders which include excitation terms resulting in different nomenclatures, e. g., S stands for single excitation, D for double excitation and T

for triple excitation. Besides this, calculations may include perturbative excitations which are named in parentheses. Thus, a CCSD(T) calculation corresponds to a triple excitations calculation which are perturbingly included with single and double expansions (Watts et al. 1992).

Coupled cluster calculations usually provide higher accuracy than MP4 computations (Tsuzuki et al. 1998). In addition, for some cases in which convergence is slowly reached with MPn calculations, CC methods have been successfully proved to be faster (Jenkins et al. 2000).

Practical considerations, recommendations and applications In conclusion, it is worth noting that *ab initio* methods give fairly accurate results, but at a very high computational cost. The smaller the approximations used in the calculation the faster the convergence to the exact solution and in a comparative manner the accuracy of the results obtained from those methods in closed shell systems should follow this order:

$$HF << MP2 < MP4 \approx CCSD < CCSD(T) < CCSDT \tag{2.12}$$

Hartree-Fock calculations, although they are insufficient for high-quality calculations, they are useful for base-level approximations for stable molecules and some transition states.

Ab initio methods allow the exact calculation, not only of the geometries and the corresponding energies, but also to predict the frequencies in the infrared and Raman spectra, the force constants and study the nature of the geometries in the potential energy surface of the molecule under study.

2.1.2 Semiempirical methods

Semiempirical methods are mainly based on the Hartree-Fock theory with the difference that these approximations include some empirical parameters derived from experimental data, or high-level *ab initio* that help, on one hand, to correct the deficiencies introduced by HF calculations and, on the other hand, to speed the calculations of bigger systems. Nowadays, semiempirical calculations on large proteins or nucleic acids can be routinely performed (Dixon & Merz 1997; Wang et al. 2004).

Semiempirical calculations yield acceptable solutions for geometry optimizations and single point calculations with respect to experimental data, however, those results may become unreliable whenever the molecules are dissimilar to the ones of the training set.

Semiempirical methods can also be used for the calculation of the vibrational modes and transition structures but usually less accurately than *ab initio* methods. Today, the most popular semiempirical methods are Austin method 1 (AM1; Dewar et al. 1985) and parametric method 3 (PM3; Stewart 1989) which have been parameterized to reproduce experimental thermodynamic measurements like heats of formation.

2.1.3 Density Functional Theory based methods

Nowadays, density functional theory (DFT) methods are not only quite popular but they are continuously renewing and are becoming the default in many QM studies. Part of this interest has grown due to the fact that these calculations involve less computational cost than correlation methods with similar accuracy in comparison with, for example, MP2 calculations.

The DFT methods do not intend to solve the Schrödinger equation, but try to obtain the energy of the system by using the Hohenberg-Kohn theorem (Hohenberg 1964) which states that the ground state energy of a polielectronic system can be calculated exactly from its electron density ($E(\rho)$) and thus, replacing the use of the corresponding wavefunction to determine the exact energy. However, the problem stands on the mathematical relationship that relates the electron density function with energy since this is unknown and, therefore, it is necessary to adopt other approaches. The most common formalism is the self-consistent (SCF-DFT) developed by Kohn and Sham (Kohn & Sham 1965), which divides the electronic energy of the system into several additive terms:

$$E = E_T + E_V + E_J + E_{XC} \tag{2.13}$$

where E_T, the kinetic energy is coming from the motion of electrons; E_V which includes the terms that describe the potential energy of the nucleus-electron attraction and the repulsion between pairs of nuclei; E_J representing the electron-electron repulsion; and E_{XC} which represents the exchange-correlation term that includes the residual electron-electron interactions. Except for E_T, all terms depend on the total electron density, ρ.

Although original Slater approximation did only include electron exchange but not correlation, within modern DFT methods E^{XC} is normally divided into two parts or functional, one which corresponds to the exchange energy and another that corresponds to the correlation energy:

$$E^{XC}(\rho) = E^{X}(\rho) + E^{C}(\rho) \tag{2.14}$$

There exists two main approaches for the calculation of the exchange-correlation energy:

- local density models with functionals based on the local density approximation (LDA) (Vosko et al. 1980) and,
- non-local or gradient-corrected models which use well functionals based on the generalized gradient approximation (GGA) (Perdew & Y. Wang 1992) or the Hartree-Fock exchange as a component.
- hybrid methods which combine functionals from other methods with the Hartree-Fock exchange integrals. For example, the B3LYP method, one of the first, and still most used, DFT hybrid methods comes from the use of the Becke hybrid functional (Becke 1988) of three parameters for the exchange term and the Lee-Yang-Parr functional for the correlation term.

In recent years, DFT methods have been adapted to calculate excited states with methods such TD-DFT (time-dependent DFT). On the other hand, there still remains the inaccurate description of dispersion forces and their inability to systematically solve the errors of the method.

Non-covalent interactions are of special importance in the stability and formation of biological structures. Hydrogen bonds and weak dispersion forces accumulate in large structures and therefore should be accounted for. Hydrogen bonds are reasonably described within DFT however, commonly used functionals fail to describe long-range dispersion. Dispersion corrected schemes were developed by S. Grimme for BLYP and PBE GGA functionals, and are, because of their correction, named BLYP-D, BP86-D, and PBE-D (Grimme 2004; Grimme 2006). The correction is done by augmenting these functionals with an empirical correction for long-range dispersion effects. Within this correction it is chosen as such that that normal bonds, which are well below the van der Waals distance, are not affected by the correction. Alternatively, the M06-2X functional (Zhao 2007), which has been constructed to recover accurate *ab initio* (MP2 and CCSD(T)) data on dispersion complexes, includes dispersion correction. This directness implies that no correction term is added to the energy calculated with the M06-2X functional.

Practical considerations, recommendations and applications Nowadays, there are many functionals and a continuos effort in developing new density functionals which

lead this type of methods to the highest places in accurate *ab initio* calculations. Major efforts are made in the improvement of numerical integration involved in the functional and, as every program does it differently, it is not uncommon to find differences in the energies obtained from identical geometries. DFT calculations are applicable provided that no excessive accuracy is required and they work particularly well for organic molecules of moderate size (50-100 atoms). In general, gradient-corrected or hybrid methods yield the most accurate results. Dispersion-corrected functionals must be chosen when dealing with non-covalent interactions. For these cases, BLYP-D or M06-2X functionals have been shown to reproduce very accurate *ab initio* data (Fonseca Guerra et al. 2010).

2.1.4 Application of QM calculations

Some of the basic applications of these methods are the prediction of molecular geometries, the relative energies of related molecular species, the spectrum of a molecule, molecular reactivities, molecular electronic potentials, or accurate description of the frontier energy levels (HOMO and LUMO).

2.1.4.1 Structural and thermodynamical properties calculation

QM methods allow the identification and localization of single points in the potential energy surface. In order to obtain those geometries, there are several optimization algorithms which typically rely on the use of analytical derivatives.

The frequency analysis is used to determine the nature of a stationary point found by a geometry optimization thus, a minimum on the potential energy surface is characterized by positive frequencies so, no imaginary frequencies. Frequency analysis include a variety of results such as frequencies, intensities, the associated normal modes, the zero point energy of the structure and various thermochemical properties. Thermal corrections include change in energy due to nuclear vibration and rotation at room temperature.

Free energies (G) calculated at some finite temperature include enthalpic (H) and entropic (S) contributions:

$$G = H - TS \tag{2.15}$$

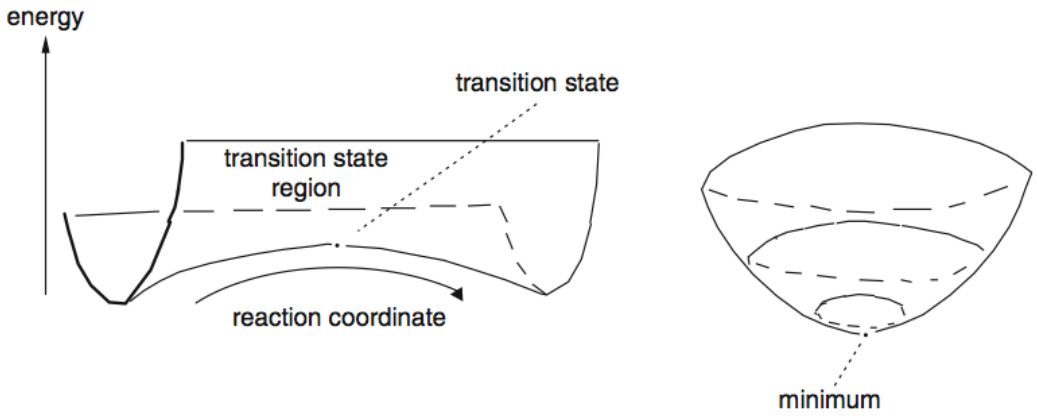


Figure 2.3: A transition state or saddle point and a minimum. Both states have zero first derivatives but can be distinguished by their second derivatives. Thus, for a minimum the second derivative must be positive while for a saddle point at least one of the frequencies should be negative.

where the change in enthalpy from 0K to a finite temperature (T) is given by:

$$H = H_{trans}(T) + H_{rot}(T) + \Delta H_{vib}(T) + RT \quad (2.16)$$

$$\Delta H_{vib}(T) = H_{vib}(T) - H_{vib}(0) = H_{vib}(T) - ZPE \quad (2.17)$$

$H_{trans}(T)$, $H_{rot}(T)$, and $H_{vib}(T)$ are the translational, rotational, and vibrational enthalpic contributions (typically computed using harmonic oscillations) at temperature T, respectively, and R is the gas constant. And the absolute entropy may be expressed as a sum of terms:

$$S = S_{trans} + S_{rot} + S_{vib} + S_{el} \quad (2.18)$$

where S_{trans}, S_{rot}, S_{vib}, and S_{el} are the translational, rotational, vibrational, and electronic entropic contributions typically approximated using the harmonic oscillator model.

2.1.4.2 Population analysis based on the basis functions

Atom charges and bond orders cannot be estimated by any experimental measurements because they are not associated to any specific physical observables. Actually, there is no exact way to determine the number of electrons that *belong* to an atom or which ones are being shared between two or more. However, there are several approaches to the approximated assignment of atom charges and bond orders and usually they

implied to fragment the electron density defined by:

$$\rho(r) = 2\sum_{i=1}^{N/2} |\psi_i(r)|^2 \tag{2.19}$$

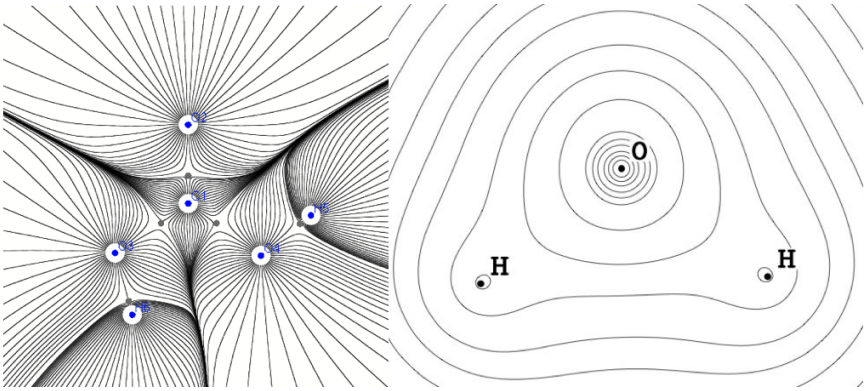


Figure 2.4: The electron density can be visualized in different ways. Electronic density is usually displayed as a gradient vector paths (left) or as a contour map (right). In the case of gradient vectors, the lines are perpendicular to the electron density contours while, in a contour map, the points that belong to the same line correspond to points with the equal density.

Among the current fractional approaches to derive charges, the Mulliken scheme is probably the most popular due to its simplicity (Mulliken 1955). The fundamental assumption used by the Mulliken scheme for partitioning the wavefunction is that the overlap between two orbitals is shared equally, and for this reason, tends to cause an unrealistic buildup of charge density on electropositive atoms. Additionally, the method based on the Bader's atoms in molecules (AIM) theory (Bader 1985) is more accurate than Mulliken approach. Bader approximation is based on the gradient vector path concept which, relates to the electron density, and gives a more intuitive idea about the molecular electronic topology. The density can be obtained from either experimental (e.g., X-ray diffraction studies) or theoretical sources. The AIM theory analyze the density using the gradient vector:

$$\nabla\rho = \frac{\partial\rho}{\partial x}\hat{i} + \frac{\partial\rho}{\partial y}\hat{j} + \frac{\partial\rho}{\partial z}\hat{k} \tag{2.20}$$

As seen in figure 2.4, gradient paths are typically traced from infinity and terminate at a nucleus which, in the electron density, corresponds to a local maximum. In addition, there are points in the electron density whose first derivative is zero ($\nabla\rho = 0$). These points are called critical points and can be characterized in terms of its second derivatives: nuclear critical points (NCP, which correspond to nuclei), bond critical

points (BCP, which are localized between bonded atoms), ring critical points (RCP, which are localized in the center of a benzene ring), and cage critical points (CCP, which are placed inside a molecule with aromatic surfaces).

In general, the charges derived from AIM theory have been shown to be relatively invariant to the basis set (Wiberg & Rablen 1993).

2.1.4.3 Population analysis based on the electrostatic potential (ESP)

The electrostatic potential, ϕ(r), is exactly defined as the work to bring a unit point positive charge from infinity to a point r (Soliva et al. 1997). Therefore, a positive ESP can be thought as repulsive effect experienced by a *proton* at the point r while a negative ESP will mean that the same proton will *feel* an attractive influence. Unlike the electron density, the electrostatic potential results from the contributions from nuclei ($\phi_{nucl}(r)$)and electrons ($\phi_{elec}(r)$) and is precisely defined as the expected value of the r^{-1} operator.

$$\phi_{nucl}(r) = \sum_{A=1}^{M} \frac{Z_A}{|r - R_A|} \tag{2.21}$$

$$\phi_{elec}(r) = -\int \frac{dr\prime\rho(r)}{|r - r\prime|} \tag{2.22}$$

$$\phi(r) = \phi_{nucl}(r) + \phi_{elec}(r) \tag{2.23}$$

where Z_A is the atom A nuclear charge, $\mathbf{R}_A$ is the position of atom A and the electronic contribution can be obtained from the integral of the electron density. The calculation of ESP is affected by the choice of basis set or by the inclusion of electron correlation (Soliva et al. 1997).

Electrostatic potentials represent an qualitative measure for molecular reactivity since they include the information about the electronic and nuclear charge distribution of a molecule. MEPs are used to derive partial atomic charges in force field parameterization (section 2.2.1.2).

2.1.4.4 Interaction energies and the basis set superposition error (BSSE)

In general, one can calculate the energy of interaction between two atoms or molecules (A and B) using the *supermolecular approach* (Chalasinski 1994), where the interaction energy is computed by subtracting the sum of A and B monomers energies from the

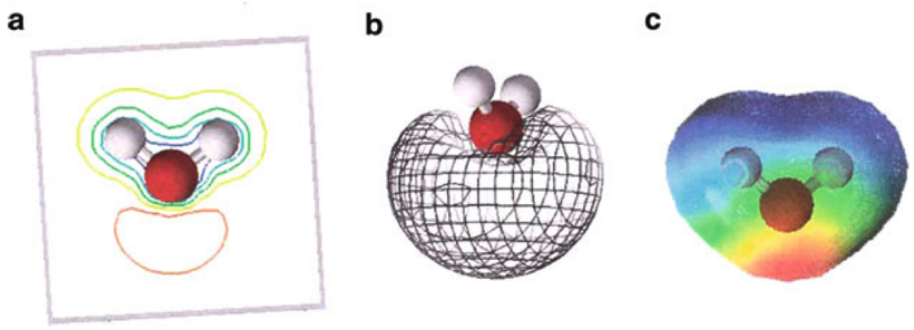


Figure 2.5: Different ways of depicting electrostatic potentials for water. Electrostatic potentials are commonly plotted as contour maps (a), as a surface itself (b) or as a van der Waals surface. Code color goes from minimum negative ESP values (red color) to maximum positive ESP values (blue color) (From Lewars 2011).

energy of the dimer AB:

$$E_{int} = E_{AB}^{AB} - E_{A}^{A/B} - E_{B}^{A/B} \tag{2.24}$$

where the superscripts indicate the basis set used for the calculation with different geometries (subscripts).

However, this scheme assumes the use of a complete basis set, which usually does not corresponds to reality. For finite basis sets, interaction energy is overestimated due to a pseudo extension of the basis sets (the basis set superposition error, BSSE). To solve this, the energies of the monomers are computed with the basis set of the dimer, a strategy which is known as the *counterpoise correction* (CP) (Boys & Bernardi 1970). Additionally, since the geometry of A alone is not the same as when interacting with B, a correction of this geometric distortion must be included, resulting in the following equations for calculating the interaction energy:

$$E_{int}(CP) = E_{AB}^{AB}(A) - E_{A}^{AB}(A) + E_{AB}^{AB}(B) - E_{B}^{AB}(B) \tag{2.25}$$

$$E_{int}^{CP} = E_{int} - E_{int}(CP) \tag{2.26}$$

where all energies are calculated with the same basis set used for the dimer.

In the DNA context, interaction energies are often computed at the base pair level or base pair step level. For example, the system consists of two stacked base pairs AB and CD, the total stacking energy $\Delta E^{AB//CD}$ (// denotes stacking) can be written as

the difference in electronic energy of the complex and the hydrogen bonded base pairs:

$$\Delta E^{AB//CD} = E^{AB//CD} - (E^{AB} + E^{CD}) \tag{2.27}$$

where all individual terms are corrected for basis set superposition error.

Complementary to the supermolecular approach, methods based on the *symmetry-adapted perturbation theory* allow the calculation of interaction energies (Eisenschitz 1930). These methods start with isolated monomers and treat the interactions between them as small perturbations of the system:

$$H_{AB} = H_A + H_B + V = H_0 + V \tag{2.28}$$

where the unperturbed operator $H_0 = H_A + H_B$ is chosen as the sum of the monomer electronic Hamiltonians H_A and H_B and V is the intermolecular interaction operator which accounts for Coulombic intermonomer interactions. This method allows to separate the intermolecular interactions by their contributions thus, the total interaction energy is computed as the sum of various energy contributions: electrostatic, exchange repulsion, induction, and dispersion. Currently, the SAPT(DFT) version allows calculations of rather big systems (about 100 atoms) with high accuracy (Jeziorski 1994; Szalewicz 2012).

2.1.4.5 Complete Basis Set Extrapolation

This scheme allows infinite basis set calculations which can be technically done by extrapolating from calculations using a finite basis set to the complete basis set (CBS) limit. The Helgaker and Truhlar schemes (Helgaker et al. 1997; Truhlar 1998) are some robust extrapolation schemes in the literature which allow very accurate calculations with an optimal amount of computational effort:

$$E_X^{HF} = E_{CBS}^{HF} + Ae^{-\alpha X} and\ E_X^{corr} = E_{CBS}^{corr} + BX^{-3} \tag{2.29}$$

$$E_X^{HF} = E_{CBS}^{HF} + BX^{-\alpha} and\ E_X^{corr} = E_{CBS}^{corr} + BX^{-\beta} \tag{2.30}$$

where E_X and E_{CBS} are energies for the basis set with the largest angular momentum X and for the complete basis set, respectively, and α and β are parameters fitted by the authors. The method of Helgaker et al. is likely more conservative and is available also for basis sets larger than aug-cc-pVTZ, while Truhlar's scheme for VTZ-quality

basis set is closer to the CBS limit.

In some cases where the highest accuracy is requested, some corrections for higher order correlation effects must be included. The difference between MP2 and CCSD(T) interaction energies has been shown to have a small basis set dependence. Thus, a good solution is usually to add this difference ($\Delta E^{CCSD(T)}$ - ΔE^{MP2}) calculated at the same level of theory to the CBS-MP2 energy:

$$\Delta E_{CBS}^{CCSD(T)} = \Delta E_{CBS}^{MP2} + (\Delta E^{CCSD(T)} - \Delta E^{MP2}) \tag{2.31}$$

2.1.5 Solvation methods

Most systems of interest are in solution and not in gas phase, and the solvent is known to have a crucial role in defining the properties of solute (spectra, vibrational frequencies, geometries, electronic properties, etc.). The study of solvated systems requires a higher computational resources since the number of molecules is increased from gas phase calculations. For this reason, a variety of approximations have been developed for those systems, where solvent is expressed as a continuum. Those approaches consider the solute as the center part of the calculation and the solvent as a continuous polarizable environment.

The free energy of solvation, ΔG_{sol} is defined as the work needed to pass 1M solute from gas phase to solution. This free energy can be expressed as the sum of three contributions: cavitation (ΔG_{cav}), van der Waals (ΔG_{vW}), and electrostatic (ΔG_{ele}):

$$\Delta G_{sol} = \Delta G_{cav} + \Delta G_{vW} + \Delta G_{ele} \tag{2.32}$$

The cavitation term corresponds to the work needed to create a volume cavity for the solute accommodation in the middle of the solvent and the van der Waals term corresponds to the repulsion-dispersion interactions between the solvent and the solute. The electrostatic term represents the charge-charge interaction between the solute and the solvent due to the rearrangement of solvent molecules to counteract against solute charge distribution.

During this PhD the main method for the evaluation of these contributions to the free energy of solvation has been the polarizable continuum method (MST).

MST method This method is based on the continuum polarizable model developed by Miertus, Scrocco and Tomasi (Miertuš et al. 1981; Miertuš & Tomasi 1982)

and parameterized for several solvents (water, chloroform, carbon tetrachloride and n-octanol) at the HF *ab initio* level and at the AM1 and PM3 semiempirical levels. In this method, the cavitation term is determined using the Pierotti's scaled particle theory (Pierotti 1976) adapted to molecular-shaped cavities by means of the Claverie procedure (Claverie 1978). Thus, the cavitation free energy is calculated by means of a series of powers of the radii of spheres R_{MS} which includes the solute (M) and excludes the centers of the solvent molecules around (S).

$$\Delta G_{cav} = K_0 + K_1 R_{MS} + K_2 R_{MS}^2 + K_3 R_{MS}^3 R_{MS} = R_M + R_S \tag{2.33}$$

where R_{MS} is the sum of the radii of atom i in the solute M and the solvent molecule S, and the coefficients K are expressed in terms of properties of the solvent and of the solution, such as the molecular radius and number density of the solvent, pressure and temperature. The final cavitation free energy is thus the sum of every atom:

$$\Delta G_{cav} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{cav,i}^P \tag{2.34}$$

where S_i is the solvent exposed surface area for each atom i, S_T is the sum of surfaces for each atom of the system, and $\Delta G_{cav,i}^P$ is the cavitation free energy for atom i according Pierotti's formula.

The van der Waals free energy is calculated using the linear relationship with the molecular surface area where ζ_i denotes de hardness of atom i determined from the experimental free energies of solvation for neutral molecules:

$$\Delta G_{vdw} = \sum_{i=1}^{N} \zeta_i S_i \tag{2.35}$$

where S_i corresponds to the solvent-exposed surface of atom i which results from scaling the van der Waals radii by a solvent-dependent factor which adopts values of 1.25, 1.50, 1.60, and 1.80 for the solvation of neutral compounds in water (Bachs et al. 1994), octanol (Curutchet et al. 2001), chloroform (Luque et al. 1996), and carbon tetrachloride (Luque et al. 1996), respectively.

Finally, the electrostatic contribution is determined by introducing the effect of the solvent reaction field on the solute as a perturbation operator to the solute Hamiltonian:

$$(H^0 + V_R)\psi = E\psi \tag{2.36}$$

The perturbation operator V_R describes the charge distribution induced on the cavity surface, generated by the polarization of the solute to the solvent. In practice, V_R is calculated by dividing the cavity surface in M small enough elements (S_i) that the charge distribution is assumed as constant inside the cavity. Therefore, V_R can be expressed in terms of polarization point charges q_i distributed over the cavity surface and they can be determined by means of Laplace equation:

$$V_R = \int_S \frac{\sigma(s)}{|r_0 - r|} dS = \sum_{i=1}^{M} \frac{\sigma(S_i)S_i}{|r_0 - r|} = \sum_{i=1}^{M} \frac{q_i}{|r_0 - r|} \tag{2.37}$$

where M is the number of elements on surface i, S_i is the area of each element, σ is the electronic density in that area and q_i is the total charge at the surface i.

The electrostatic component to solvation free energy is obtained by this equation:

$$\Delta G_{ele} = \left\langle \psi^{sol} \left| H^0 + \frac{1}{2} V_R \right| \psi^{sol} \right\rangle - \langle \psi^0 | H^0 | \psi^0 \rangle \tag{2.38}$$

where ^{0} and ^{sol} superindexes stand for gas phase and solution states.

2.2 Classical Mechanics (CM)

CM methods allow the energy calculation of very large systems through the application of classical equations for the description of bonds, angles, torsions and non-covalent interactions.

Molecular mechanics treats the atoms not as a set of atomic orbitals, but as points in space with their coordinates and partial atomic charges that are connected together by simple potentials. Unlike QM methods, these links are needed to specify whether single bonds, double, etc. are connecting the atoms and where these atoms can form bond, angles or torsions. Therefore, the underlying concept of these methods is more like a model of balls and bars, where internal motions are restricted by the parameters on which they depend. The combination of these parameters, derived from both experimental and QM estimations, along with the energy equation of the system is called the force field. The name stands for the fact that the first derivative of the energy with respect to the coordinates of atoms corresponds to the forces to which the atoms are subjected.

2.2.1 The force field

Nowadays, there are several force fields differing in the number of terms in the energy equation, the complexity of those terms and the parameterization used to obtain the force constants. A general equation for the potential energy of a molecule can be written as an additive contribution of the bonding terms (bonds, angles and dihedrals) and the non-bonding interactions:

$$E = \sum_{bond} E_{stretching} + \sum_{angle} E_{bending} + \sum_{dihedrals} E_{torsion} + \sum_{pairs} E_{nonbond} \tag{2.39}$$

2.2.1.1 Bonding Terms

Bonding *stretching* term Considering bonds as springs that bind atoms as classical particles, the energy expression is often described by a harmonic oscillator equation so proportional to the square of the difference between the distance between atoms and their distance in equilibrium:

$$E_{stretch} = k_{stretch}(l - l_{eq})^2 \tag{2.40}$$

where the constant $k_{stretch}$ is actually one-half the actual force constant of the bond according to Hooke's law. Therefore, the bigger the $k_{stretch}$ the more difficult it is to move the atoms from the equilibrium distance.

Angle *bending* term The energy related to the angle connecting three bonded atoms is also proportional to the square of the difference formed by the angle between three consecutive atoms and its angle in the equilibrium:

$$E_{bend} = k_{bend}(a - a_{eq})^2 \tag{2.41}$$

where k_{bend} corresponds to one-half the bending force constant, a is the angle at any time and a_{eq} the equilibrium value.

***Torsional* term** Describes the periodicity of the potential energy for intramolecular rotations. There are two types: proper and improper torsions. Proper torsions correspond to four atoms sequentially bonded while improper ones are associated to relative

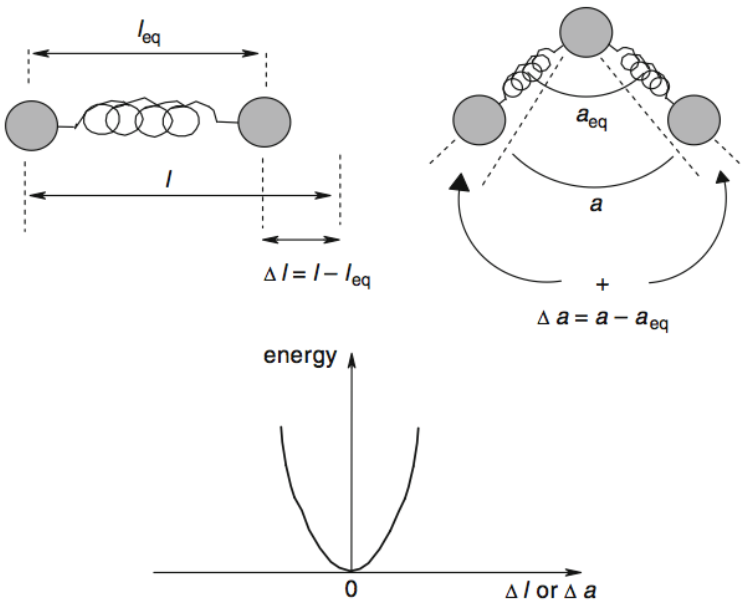


Figure 2.6: The separation of the equilibrium position both in the bond distances and angles in an increase of the energy of the molecule.

movements of one atom respect to the other three in the same plane. The energy expression is usually defined as a sine or cosine function or combination of both. The energy expression has typically the form:

$$E_{torsion} = \sum_{torsions} \sum_{n} \frac{V_n}{2}(1 + cos(n\chi - \gamma)) \tag{2.42}$$

where V_n is the force constant of each term of the Fourier series, n is the multiplicity (the number of function minima when rotating 360 °), χ corresponds to the torsion angle and γ_i is the phase angle or the angle corresponding to the energy minimum. In the case of improper torsions, the periodicity is usually two.

2.2.1.2 Non-bonding terms

Electrostatics interactions term Force fields may or may not include an electrostatic contribution to the total potential energy expression. This term represents the attraction or repulsion between two non-bonded atoms due to their atomic charge generally represented as two point charges. The electrostatic interaction between atoms is calculated using the Coulomb's law:

$$E_{elec} = \sum_{i,j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.43}$$

where the partial atomic charges associated with each atom are represented by q_i y q_j, ϵ_0 represents the dielectric constant in vacuum and r_{ij} is the distance between the two interacting atoms in angstroms.

Atomic charges are typically derived from ESPs through the so-called RESP (restrained electrostatic potential) fitting strategy (Momany 1978; Cox & Williams 1981). Within the RESP approach, ESP atomic-centered charges are fitted to reproduce the molecular electrostatic potential (MEP) derived from QM calculation. The MEP is calculated at a large number of points defined on three-dimensional surfaces around the molecule of interest. Currently, derivation of RESP charges can be done in an automatic and straightforward way (Dupradeau et al. 2010; Vanquelef et al. 2011).

van der Waals term Includes dispersion-repulsion interactions between atoms not directly linked so that when two atoms get closer the van der Waals energy increases. When the distance between them is equal to the sum of the van der Waals radii of the two atoms then the attraction is optimal, and thus, the van der Waals energy will correspond to a minimum. For shorter distances, the van der Waals energy increases drastically for that pair of atoms while for longer distances converge to zero. The mathematical relationship of this contribution is typically expressed as a Lennard-Jones 12-6 potential:

$$E_{vdW} = \sum_{i,j<i} \left[\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - \left(\frac{B_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.44}$$

where A_{ij} y B_{ij} include energy parameters that determine the depth of the potential energy minimum that depends on van der Waals radii of each pair of atoms, and r_{ij} represents the distance between the non-bonded atom centers.

2.2.2 Force field parameterization

In the mathematical expressions explained so far, force field parameters are determined to reproduce experimental data and theoretical estimations. Some parameters, like bond and angle parameters, can be parameterized independently. However, non-bonding and torsion parameters present a non-negligible coupling and, therefore, they

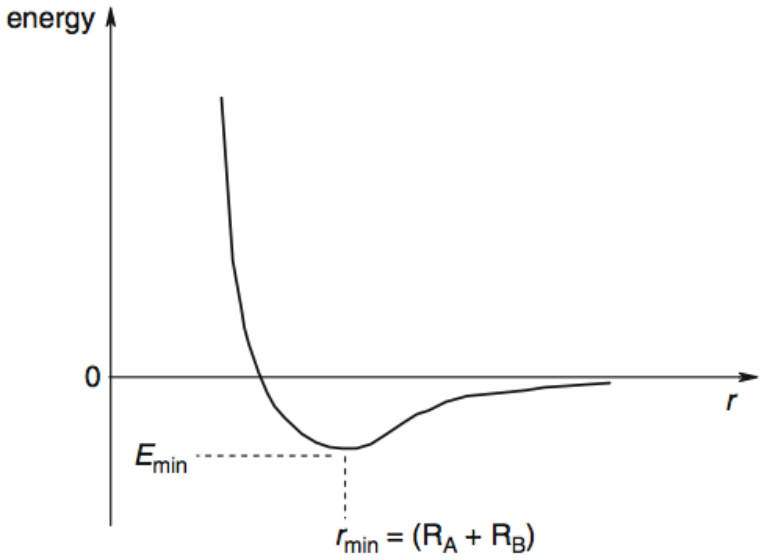


Figure 2.7: Representation of energy between two unbound atoms A and B. The energy minimum corresponds to the distance of the sum of the van der Waals radii corresponding to two atoms. The posterior approach below the minimum distance rapidly destabilizes the interaction.

cannot be independently parameterized. Usually, atomic charges are derived from potential electrostatic QM calculations and van der Waals parameters are adjusted to reproduce QM energy interactions. Since torsional parameters are difficult to extract from experimental data, force constants and phase angles are adjusted to reproduce the rotational profiles obtained from high-level *ab initio* calculations.

Since force field-based methods have to deal sometimes with new atoms or molecules, usually, the first option is to search in the literature for those parameters. There are online databases for common prosthetic factors and other ions and organic molecules like the one maintained by Bryce group http://www.pharmacy.manchester.ac.uk/bryce/amber/.

Current force fields account for the majority of the commonest organic compounds and, as previously explained, the effort of different research groups around the world enlarge the possibilities of these force field-based methods.

2.2.3 Main force fields for the study of biomolecules

There are several force fields that are similar and these undergo periodic updates over time. They differ in the selection of the atom types and their parameterization avoiding the transferability of parameters from one force field to another. Some of the most popular force fields are presented as follows:

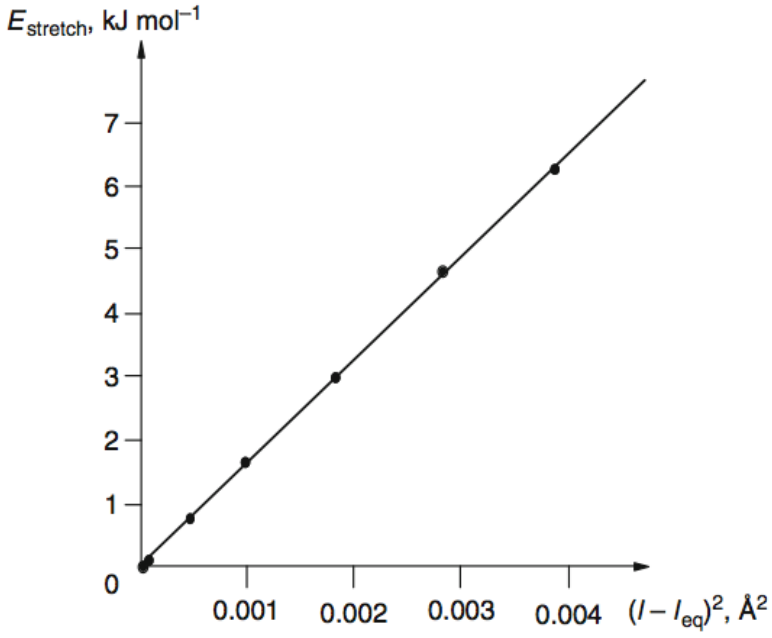


Figure 2.8: Linear representation of the *stretching* energy and the square of the difference link lengths for calculating the force constant associated.

- AMBER (http://ambermd.org/): developed originally in the laboratory of Peter Kollman, has now become one of the preferred options for the simulation of proteins and nucleic acids. It is remarkable the high degree of correlation of some properties derived from simulations with experimental data, especially in the case of the treatment of electrostatic interactions. The expression of the potential energy term for adding a hydrogen bond interactions and the use of torsional general parameters, i.e. in general torsional potential are defined by the two central atoms and not by the four dihedral atoms. The result is an expression of the potential energy with five terms:

$$\begin{aligned} V &= \sum_{bonds} \frac{k_l}{2}(l - l_{eq})^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_{eq})^2 + \sum_{dihedral} \sum_{n} \frac{V_n}{2}[1 + cos(n\omega - \gamma)] + \\ &+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \end{aligned} \tag{2.45}$$

- CHARMM (http://www.charmm.org/): developed originally by the laboratory of Martin Karplus. Together with AMBER is one of the two most commonly used force fields in the field of nucleic acids and proteins. In the mathematical expression level of the force field, the energy term associated with the hydrogen

bonds is implicitly in terms of the combination of electrostatic and van der Waals and also the interaction parameters are calculated in a different way as is done for those included in the AMBER force field. The potential energy equation has the following form:

$$\begin{aligned} V &= \sum_{bonds} \frac{k_l}{2}(l-l_{eq})^2 + \sum_{angles} \frac{k_\theta}{2}(\theta-\theta_{eq})^2 + \\ &+ \sum_{Urey-Bradley} \frac{k_{UB}}{2}(l_{1-3}-l_{1-3,eq})^2 + \sum_{dihedral} \sum_{n} \frac{V_n}{2}[1+cos(n\omega-\gamma)] + \\ &+ \sum_{improper} \frac{k_\omega}{2}(\omega-\omega_{eq})^2 + \sum_{residue} V_{CMAP}(\Phi,\Psi) + \\ &+ \sum_{i=1}^{N}\sum_{j=i+1}^{N} \left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \end{aligned} \tag{2.46}$$

where, unlike the expression used in AMBER, it mainly includes two different terms. First, the additional bonded term is the two-body Urey-Bradley term, which extends over all 1-3 bonds and second, the CMAP term (MacKerell et al. 2004; MacKerell et al. 2004) is a cross term which is a function of two sequential protein backbone dihedrals (Φ and Ψ). CMAP originates from differences observed between classically derived 2D Φ/Ψ peptide free energy surfaces using CHARMM22 force field and those of experiment. Therefore, CMAP accounts for correction of the classical map to match that of a QM calculated map.

- GROMOS (http://www.gromacs.org/): the GROMOS force field (Groningen molecular simulation) originally developed in the laboratory of Wilfred van Gunsteren, albeit intended for general application, has not received so far wide acceptance for nucleic acids simulations. It is however, very popular for proteins and membrane simulations. Its mathematical formalism (Oostenbrink et al. 2004) is:

$$\begin{aligned} V &= \sum_{bonds} \frac{1}{4}K_b[b^2-b_0^2]^2 + \sum_{bond\ angles} \frac{1}{2}K_\theta[cos\theta-cos\theta_0]^2 + \\ &+ \sum_{improper} \frac{1}{2}K_\zeta[\zeta-\zeta_0]^2 + \sum_{dihedral} K_\varphi[1+cos(\delta)cos(m\varphi)] + \\ &+ \sum_{pairs\ i,j} \left(\frac{C12_{ij}}{r_{ij}^{12}} - \frac{C6_{ij}}{r_{ij}^{6}}\right) + \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1}\frac{1}{r_{ij}} + \\ &+ \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1}\frac{-\frac{1}{2}C_{rf}r_{rf}^2}{R_{rf}^3} + \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1}\frac{-\left(1-\frac{1}{2}C_{rf}\right)}{R_{rf}} \end{aligned} \tag{2.47}$$

which includes, in this order, bond, angle, improper, torsional terms for bonded interactions, and van der Waals (using a Lennard-Jones 12-6 interaction function) and electrostatic terms for non-bonded interactions. In fact, electrostatic interactions consist of three contributions: direct Coulombic interactions, a reaction-field contribution, and a distance-independent reaction-field term.

2.2.4 Force field based methods

The result of a classical calculation is the potential energy associated with a configuration of atoms in the system. This information can subsequently be processed using different algorithms, which results in a range of computational methods. They include Molecular Mechanics (MM), Molecular Dynamics (MD), and Monte Carlo (MC).

2.2.4.1 Molecular mechanics (MM)

The simplest case of use of Molecular Mechanics method is the calculation of the energy minimization, this type of calculation is used to find the energy minimum or the nearest stationary point to the starting geometry on the potential energy surface or, if it is an optimized geometry, to calculate the relative energies among different optimized conformations. Minimization calculations are also performed to allow the system to find a *relaxed* geometry without strong interactions or steric clashes as a previous step before starting more complex simulations.

The commonest algorithms for finding the energy minimum are the steepest descent (SD) and the conjugate gradient. In the steepest descent, the searching algorithm explores the negative gradient direction so lowering the function value and getting close to a minimum and for this reason, SD is fast specially when the starting point is far from the minimum. On the other hand, The conjugate gradient algorithm searches not along the gradient direction but along a *conjugate* direction from the previous search and therefore, it is more efficient when being close to the minimum (Leach 2001). Combination of both methods is typically used to find the energy minimum.

2.2.4.2 Molecular dynamics (MD)

The result of MD simulations is a sequence of positions and velocities as a function of time. The MD method is based on solving the differential equation from the Newton's second law of motion:

$$\frac{\partial^2 x_i}{\partial t^2} = \frac{\vec{F}_i}{m_i} \tag{2.48}$$

where x_i are the coordinates, t is the time, m_i is the mass of nucleus i and $\vec{F}_i$ is the force applied to i. The calculation of the position (x_i) at any time t of a particle with mass m_i for which the applied net force $\vec{F}_i(t)$ is computed from the first derivative of the potential energy ($\partial\vec{F}_i/\partial$ x). Once the new positions and velocities of the particle are known, it is possible to calculate the new potential and kinetic energies of the system so the procedure can be again applied to obtain the corresponding new net forces on each particle ($F(t + \Delta t)$). Since numerical integration is needed, there are several techniques to achieve this goal.

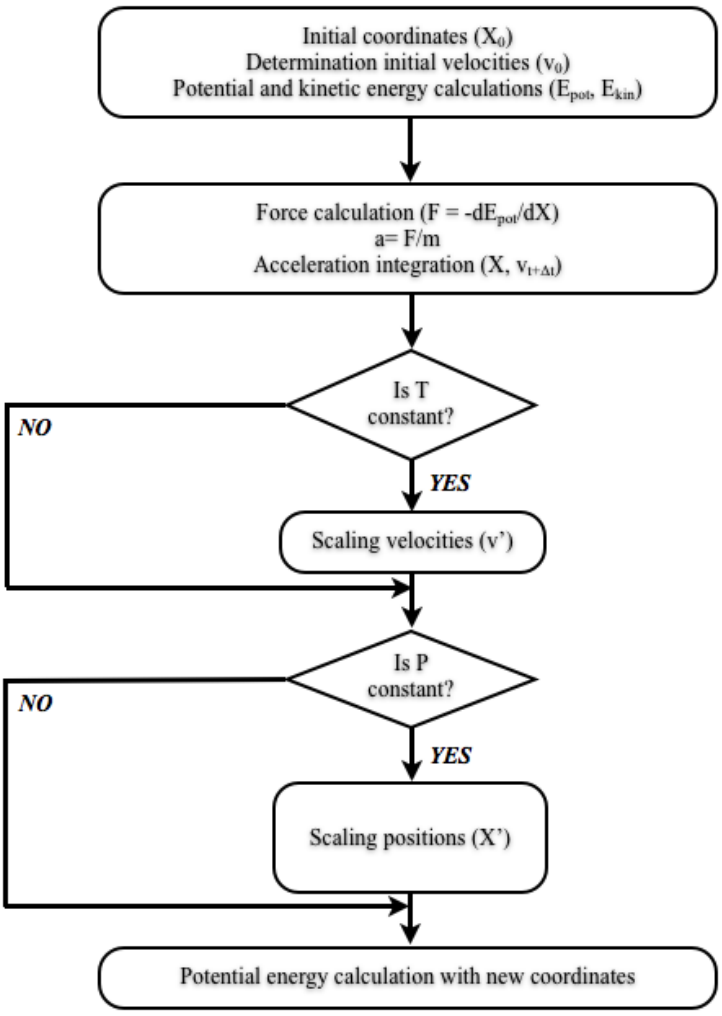


Figure 2.9: Basic algorithm of molecular dynamics.

Numeric integration algorithms The set of differential equations (3N second-order or 6N first-order differential equations) to be solved is transformed into difference

equations. Therefore, a numerical integrator must be chosen, and the algorithm should be consistent (that is, it should reproduce the original differential equations), accurate (the final solution should be close to the real one), efficient (important since the calculation of forces is computationally demanding it should be able to use the larger possible time step in order to save computing time) and, symplectic (so the volume in the phase space and the total energy of the system must be conserved). There are accurate algorithms which properly conserve short-term energy but they suffer from small long-term energy drift. Second-order algorithms like the Verlet algorithm (Verlet 1967) works very well for complex systems.

- The *Verlet* algorithm (Verlet 1967) uses the position and force at time t and the position at time *(t-Δt)* to calculate the new position at time *(t+Δt)*. The Verlet algorithm is obtained from the Taylor expansions of $\vec{r}$*(t+Δt)* and $\vec{r}$*(t-Δt)* about time t:

$$\vec{r}(t+\Delta t) = 2\vec{r_i}(t) - \vec{r_i}(t-\Delta t) + \frac{1}{m}\vec{F_i}(t)(\Delta t)^2 \tag{2.49}$$

  However, at $t = 0$, since the position at time *(-Δt)* is unknown a different Taylor expansion of $\vec{r}$*(t)* or a different algorithm must be use for the initial step. The velocity in the Verlet algorithm is calculated by:

$$\vec{v_i}(t) = \frac{\vec{r_i}(t+\Delta t) - \vec{r_i}(t-\Delta t)}{2\Delta t} \tag{2.50}$$

  So the velocity is always calculated a step behind the position term. Besides this, to correct minor errors in the integration, the Verlet algorithm allows often to rescale the velocities after some steps.

- A modification of the Verlet algorithm is the so-called *leapfrog* algorithm (Hockney & Eastwood 1988) which follows these equations for the calculation of the positions and the velocities:

$$\vec{v_i}\left(t+\frac{\Delta t}{2}\right) = \vec{v_i}\left(t-\frac{\Delta t}{2}\right) + \frac{1}{m_i}\vec{F_i}(t)\Delta t \tag{2.51}$$

$$\vec{r_i}(t+\Delta t) = \vec{r_i}(t) + \vec{v_i}\left(t+\frac{\Delta t}{2}\right)\Delta t \tag{2.52}$$

  The leapfrog algorithm uses the position and the force at time t and the velocity at half a time step *(t-$\frac{\Delta t}{2}$)* to calculate the positions and velocities. The leapfrog algorithm avoid the problem at the initial step and the initial velocities that can be derived from a Boltzmann distribution at a given temperature.

The time step choice The choice of this parameter is essential since a very high value would cause the atoms to move too far along the path, an undesirable problem if one wants to study the motion of molecules. However, when using a very small time step advance of trajectory will be very slow. The solution is usually a time step that is one order of magnitude smaller than the time scale in which the high-frequency movements of the system occur. Those are the bond vibrations and are typically not of interest. Therefore, constraint algorithms are applied to remove those degrees of freedom. Constraint methods can be applied to bonding distances like SHAKE algorithm (Ryckaert et al. 1977) and as well to velocities like RATTLE algorithm (Andersen 1983). Usually, the integration step chosen is on the femtosecond scale (fs, 10^{-15} s), a value fairly distant in time, for example, from the timescales of folding processes of proteins or many biologically relevant conformational changes. The application of constraint algorithms extend up to 2 fs the integration step reducing the computational cost. Alternatively, some authors use multiple time step MD strategies being the most popular RESPA (REference System Propagation Algorithm) (Tuckerman 1992) which classifies into a number of groups the forces within a system according to how rapidly those forces vary over time, associating to each group its own time step.

MD simulation conditions The starting point for any MD simulation is the initial configuration. This typically comes from an x-ray structure, an NMR ensemble or a theoretical model. Besides initial coordinates, a set of initial velocities is defined from a Maxwell-Boltzmann distribution so as the associated kinetic energy is that corresponding to the target temperature. Finally, additional conditions are required like the number of particles (N), the volume (V), the temperature (T), the pressure (P) and the total energy of the system (E). Therefore, depending on these conditions, the ensemble is defined as:

- microcanonical (NVE)
- isothermic-isobaric (NPT)
- canonical (NVT)

where NPT and NVT ensembles are the two most common ensembles used in the field.

In explicit MD simulations, the solute is surrounded by solvent molecules, usually water, and ions neutralizing the net charge of the system and introducing a given ionic strength. Due to molecular motions during simulation, the limits of the finite system can introduce artifactual behavior. In order to avoid this effect, periodic boundary conditions (PBC) are used by replicating the system in 3D space. The shape and the size of the box depend on the geometry of the system. Octahedral and rectangular

shapes are the most used for nucleic acid simulations. In any case, box size must be large enough so as to avoid self-interactions between replicas.

In common NPT and NVT ensembles, the temperature must be constant. The simplest way to control the temperature is by scaling the velocities (Woodcock 1971) at each time step so decreasing them when the system warms up and increasing when the system cools down. Altenatively, by coupling the system to an external heat bath that is fixed at the desired temperature (Berendsen et al. 1984), the system receives or gives heat depending on its temperature. Therefore, velocities are scaled at each step and the rate of change of temperature is proportional to the difference in temperature between the system and the bath. Finally, a third alternative, the so-called Nosé-Hoover thermostat (Nosé 1984; Hoover 1985), includes a friction coefficient that depends on the temperature in motion Newton's equations.

$$\lambda^2 = 1 + \frac{\partial t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right) \tag{2.53}$$

where λ is the scaling factor, ∂t is the length step, τ is the coupling parameter, T_{bath} is the temperature of the heat bath and T is the desired temperature.

In order to fix the pressure to a given value (NPT ensemble), the volume of the system must change. A change in the volume of the system is related to the isothermic compressibility coefficient (κ). Many methods control pressure by scaling the volume. An alternative is to couple the system to a ′pressure bath′, analogous to a heat bath (Berendsen 1984), so the scaling factor, λ, depends on κ and the difference between the pressure of the system and in the *bath*.

$$\lambda = 1 - \kappa\frac{\partial t}{\tau_p}(P - P_{bath}) \tag{2.54}$$

where τ_p represents the relaxation constant, ∂t is the length step, P is the desired pressure and P_{bath} is the pressure of the *bath*. The new positions of each atom are multiplied by $\lambda^{1/3}$.

Current force fields for the study of nucleic acids In the latest years, the improvement of computational resources allow the production of very long simulations (10^{-3} seconds with specific hardware) so the results can be directly compared with experimental measurements. Since these technical jumps occur very rapidly, force field development is currently improving the stability and reliability of MD-derived data. Nowadays, for nucleic acids, the basic Cornell force field (Cornell et al. 1995) in AMBER has been improved with several modifications like the *parm99bsc0* (Pérez et al.

2007) which is the reference force field for nucleic acids simulations. *Parm99bsc0* has been shown to be strikingly good at reproducing the highest quantic calculations for the essential interaction energies that stabilize nucleic acids (Sponer et al. 1996). The level of detail goes so close to reality that the latest force fields are able to recover even the sequence-dependency already present in experimental structural databases (Pérez et al. 2007).

Additional reparameterization for RNA systems has been done more recently. In order to address a tendency of RNA double helices to convert them (after very long simulations) to ladder-like structures, there are currently three different backbone modifications for the glycosidic torsion which have been applied to different non-double stranded RNA systems. From these force field modifications, the more tested is the *chiOL3* (Zgarbová, M. et al. 2011) which lead to a better RNA representation, always used in combination with *parm99bsc0*.

The CHARMM force field has undergone also a notable improvement in the last years. The previous force field, CHARMM27 (MacKerell et al. 2000), although it was shown to correctly represent DNA molecules (Pérez et al. 2008), it gave some instabilities for double stranded A-RNA molecules which were related to uncomplete specific backbone parameterizations and an underestimation of the hydrogen-bonding interactions within the A·U base pair (Faustino et al. 2010). The reparameterization of the dihedral parameters of the 2'-hydroxyl proton of the ribose suppresses those instabilities but still, although less frequently, some breaking events persist after relatively long simulations. Further improvement of the ϵ/ζ backbone dihedrals(Hart et al. 2012) for DNA has been added to the latest force field (CHARMM36 (Denning et al. 2011)) in order to correctly represents the BI/BII backbone conformation balance. Thanks to the latest modifications included in CHARMM36 force field, the description of RNA molecules is now more reliable and comparable to the results obtained from AMBER simulations.

Practical considerations and recommendations It is important to check convergence by analyzing evolution of the system along different simulation windows. It is also required to analyze trajectories to compute parameters which can be compared with experimental values bearing in mind the corresponding time scales.

From all nucleic acids force fields used within MD techniques, *parm99bsc0* has been shown a good ability to reproduce well a variety of nucleic acids (Orozco 2003; Pérez 2012). In any case, it is advisable to remember that a force field is an approximation to the potential energy and that continuum checking is required.

2.2.5 Statistical mechanics

Since standard MD and MC simulations do not sample the entire phase space, thermodynamic properties cannot be accurately calculated from them. For this reason, statistical mechanics provide a framework connecting the micro states of a system to macroscopic thermodynamic quantities by means of probability functions. Among the thermodynamic properties, free energy is probably the most relevant one.

Free energy differences between two states are related to the ratio of probabilities of those states:

$$\Delta G_{ij} = -k_B T ln \frac{Q_i}{Q_j} \tag{2.55}$$

where ΔG_{ij} corresponds to the Helmholtz free energy difference between states i and j, k_B the Boltzmann constant, Q the canonical partition function and T is the temperature in Kelvin. At this point, there are several key assumptions which must be noted for the practical calculation of ΔG. First, the free energy is always calculated as a difference between two states and no absolute free energies are calculated. Second, although masses can change between states i and j only the change in potential energy is included so kinetic energy differences are assumed to be negligible. And third, both states are considered to be at the same temperature and the PV contribution is usually very small. However, even with these assumptions, the use of equation 2.55 is difficult since it requires long enough simulations so as to generate a reversible and equilibrated sampling of states i and j, which is often difficult forcing the use of methods which biases trajectories to generate such sampling.

The three most common methods for calculating free energies differences between two states coupled to MD (or MC) methods are explained below.

- *Zwanzig relationship (Zwanzig 1954)*: historically known as the free energy perturbation (FEP) method, it computes the free energy between two potential energies U_0 and U_1 over a coordinate:

$$\begin{aligned} \Delta G &= G_1 - G_0 = k_B T ln < exp - [k_B T (V_1 - V_0)] >_0 & (2.56) \\ &= k_B T ln < exp - [k_B T \Delta V] >_0 & (2.57) \end{aligned}$$

  where $\langle\rangle_0$ denotes an MD or MC generated ensemble average of ΔV that is sampled using the V_0 potential. Thus, the free energy difference between states 0 and 1 is typically calculated along a pathway defined by an order parameter λ divided in segments or 'windows' from state 0 to 1 through intermediates without

physical meaning, which allows us to define a reverse path between state 0 and 1.

$$H(\lambda) = \lambda H_1 + (1 - \lambda)H_0 \tag{2.58}$$

where λ varies from 0 to 1 and H_λ is the Hamiltonian of the system associated to intermediate λ. Therefore, the free energy difference from states 0 to 1 is the sum of the free energies between each of the λ values.

- *Thermodynamic integration, TI* (Kirkwood 1935; Kirkwood 1968): this method results from the first derivative of the free energy with respect to the λ coordinate which relates the two ending states of the thermodynamical process. Unlike the FEP method, the free energy is calculated by integrating a finite difference of the free energies respect to λ at different values of λ over very small $\partial\lambda$ steps along the route of the chemical transformation:

$$\partial A/\partial\lambda = \left\langle \frac{\partial U(\lambda)}{\partial\lambda} \right\rangle_\lambda \tag{2.59}$$

$$\Delta A = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial\lambda} \right\rangle_\lambda \partial\lambda \tag{2.60}$$

  The parameter λ can be chosen to follow any variation of the system, i.e. a torsion conformational change or an alchemical transformation of a group into another.

- *Bennett acceptance ratio, BAR* (Bennett 1976): The Bennett acceptance ratio (BAR) and the multistate Bennett acceptance ratio (MBAR) methods have been shown to be more efficient than the Zwanzig and TI methods. In 1976, Charles Bennett solved the free energy difference problem between two states by developing the acceptance ratio estimator which is proportional to the degree of overlap between both ensemble states. The BAR method has been remarkably useful in large molecular changes given that a fewer number of intermediate steps are required in comparison with the TI method (Bennett 1976; Shirts & Pande 2005; N. Lu et al. 2003).

  Since the free energy is a state function, the combination of thermodynamic cycles with both FEP, TI or MBAR methods is typically used to calculate the free energy associated to, for example, binding of different ligands to a given receptor (Tembre & McCammon 1984), or solvation (Jorgensen & Ravimohan 1985) or alchemical changes (Jorge et al. 2010). It is important to start by determining the ending states and build a suitable thermodynamic cycle which allows the free energy computation (Figure 2.10).

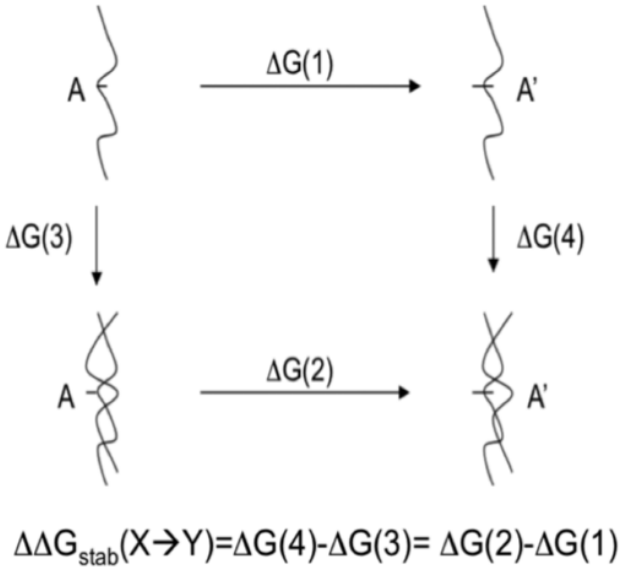


Figure 2.10: Example of thermodynamic cycle used to determine the contribution of mutation A → A' to the stability of the DNA duplex (From Faustino et al. 2009).

2.2.5.1 Free energy pathways via biased simulations

If the free energy barriers are of the same magnitude as the thermal fluctuations, it is feasible to obtain the free energy profiles associated with a given conformational transition directly from classical MD simulations. However, the energy barrier between two states is too big or diffusion is required to go from one state to the other. Some methods attempt to overcome the sampling problem by modifying the potential function so that unfavorable states are sufficiently sampled. Among these biased methods, *umbrella sampling* (Torrie & Valleau 1977) is probably the most used. This method introduces an additional term in the potential energy function, the umbrella potential, along one (or a few) reaction coordinate(s) which takes a harmonic form:

$$U'(x) = U(x) + k(\zeta - \zeta_0)^2 \tag{2.61}$$

where *U(x)* is the potential energy for a given configuration denoted by x, and ζ denotes a specific reaction coordinate. Unlike FEP or TI methods, an umbrella sampling calculation involves a series of mutually overlapping 'windows'. Therefore, the choice of the reaction coordinate is essential to recover the free energy between the two states of interest. Transforming the biased free energy into the corresponding unbiased value is typically performed with the *weighted histogram analysis method* (WHAM) (Kumar et al. 1992; Roux 1995).

WHAM is based on the idea that from a MD or MC simulation it is possible to rebuild the density of visited states by weighting the probability distributions for any parameter of the system. Thus, it is possible to obtain a discrete number of states, then it is possible to create a histogram in which for every bin the probability of finding a specific state can be provided, from which the WHAM equation for K statesis derived:

$$G_i = -\beta^{-1} \ln \sum_{k=1}^{K} \sum_{1}^{N_k} \frac{exp[-\beta U_i]}{\sum_{k=1}^{K} N_k exp[\beta G_k - \beta U_k]} \tag{2.62}$$

where β ($\beta = k_B T$), k corresponds to one of the states from 1 to K states, i corresponds to one of the bins, N_k is the number of counts in histogram bin associated to state k, G_i and U_i are the corresponding free and potential energies for each bin, and G_k and U_k correspond to free and potential energies of k state in bin i.

Other biased simulation techniques, such as *adaptive biasing force (ADP)* (Darve et al. 2008) and *metadynamics* (Laio & Parrinello 2002) are used to calculate free energies associated to a reaction coordinate, referred to as collective variables (CVs). They might be useful when there is difficult to identify the variable driving transition.

Practical considerations and recommendations Although these methods have been shown to yield acceptable solutions to different kind of processes, running free energy methods has several inconveniences. It is very important that at every simulation step the system is able to reach a complete equilibration in order to get accurate energy values at every step of the process. This equilibration length is usually chosen depending on the magnitude of the change to be tracked. It is as well important to check the reversibility of the reaction so the free energies associated to the $0 \rightarrow 1$ and the $1 \rightarrow 0$ processes have to be very similar in absolute value. The lack of reversibility is usually due to hysteresis problems which can be address by reducing the λ difference and/or integrating the energy function at a higher number of steps.

In recent years, the emergence of methods for the calculation of free energy differences has been increased making sometimes a bit of chaotic to find the best technique to address each problem. Careful analysis of the system and multiple tests are required to find the optimal procedure.

2.2.6 Analysis of the results

2.2.6.1 Structural analysis

A MD trajectory contains a detailed information of the temporal fluctuations of the system. This information can be processed globally or locally depending on the level of study. For example, one can calculate a global descriptor such as the **root mean square deviation (RMSD)** with respect to a reference structure, usually the original one or in other cases the average along a part or the entire simulation or only a part of it:

$$RMSd = \left[\frac{1}{M} \sum_{l=1}^{3N} m_l (x_{kl} - x_l)^2 \right]^{1/2} \tag{2.63}$$

where M corresponds to the total mass, m_l is the atomic mass associated to l, $\mathrm{x}_k l$ represents the l-coordinate in the structure k, x_l the reference value in the reference structure and N is the number of atoms.

In nucleic acids, the structural analysis requires the study of geometric parameters which usually include base pair relative positions (3 rotations and 3 translations) backbone torsions and groove volume analysis among others. These **physical parameters** are computed according to the Tsukuba convention (Olson et al. 2001) with a variety of software programs, e.g., Curves+ and 3DNA (Lavery et al. 2009; Lu & Olson 2003) that have been used in this PhD.

Hydrogen bonding and stacking interactions are essential for DNA stability and flexibility. DNA force fields have been developed to carefully account for these interactions and succeeding in getting similar values to QM calculations. We typically compute hydrogen bonding energy from trajectories by using truncated residues i.e. without backbone-sugar atoms, keeping neutrality by compensating the C1' atomic charge for the rest of the nucleobase. Stacking interactions are usually classified in intramolecular and intermolecular interactions while hydrogen bonding interactions involve opposing bases.

2.2.6.2 Dynamic analysis

It is usually worth to find the main internal motions in molecules which imply correlations between several atoms. This can be computed with **essential dynamics** by diagonalizing the covariance matrix as in PCA (principal component analysis):

$$C = cov(x) = \langle (x - \langle x \rangle)(x - \langle x \rangle^T \rangle \tag{2.64}$$

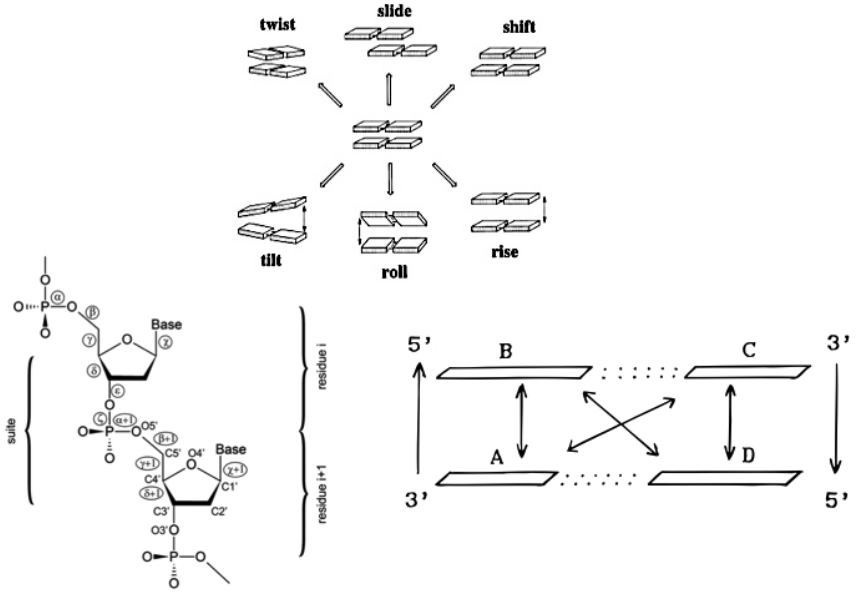


Figure 2.11: Base pair step helical parameters (top), backbone torsions (left bottom) and base-base interactions (hydrogen bonding and stacking) in DNA (right bottom).

where $\langle\rangle$ represents the temporal average value and x are the corresponding cartesian coordinates for every atom in the trajectory usually referred to a common reference structure. The covariance matrix diagonalization produces a set of $3N$ *eigenvectors* and their corresponding *eigenvalues* and from them 6 will represent rigid motions (both translations and rotations) so their value will be zero. The orthogonal eigenvectors which result from the PCA are able, on one side, to represent the total variance of the system and on the other side, their associated eigenvalues represent the weighted contribution to the total variance.

From the computation of the eigenvalues and the eigenvectors is possible to calculate the force constants associated to the different deformation modes:

$$K_i = \frac{k_B T}{\lambda_i} \tag{2.65}$$

where k_B corresponds to the Boltzmann constant, T is the temperature and λ_i is the associated eigenvalue to the i mode in Å^2. Once the corresponding force constant is known, the deformation energy associated to a deformation mode is easy to compute by using an harmonic potential:

$$E_i = \frac{K_i}{2}(\Delta X_i)^2 \tag{2.66}$$

where ΔX_i corresponds to the cartesian deformation of the i mode.

Eigenvalues can be used to estimate the accessible configurational volume, the entropy of the system or many other descriptors of the system dynamics. On the other hand, eigenvectors give information about the kind of movement of the system. A common first approximation is to compute the differences with average coordinates along the most representative vectors i.e, with higher contribution to the total variance. This procedure creates a pseudotrajectory which nearly enclosed the principal movements of the system. These main eigenvectors can be used to compare with similar simulations in order to check how similar two different trajectories are in terms of their essential space. This comparison can be easily computed by scalar product of the two independent sets of eigenvectors received the name of **absolute similarity index** (Hess 2000; Hess 2002):

$$\gamma_{AB} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} (v_i^A \cdot v_j^B)^2 \tag{2.67}$$

where γ_{AB} is the similarity index between trajectories A and B, n is the number of eigenvectors needed to represent the 80-90% of the total system variance (in the case of double strand nucleic acids represents the first 10-20 eigenvectors) and v_i and v_j are the corresponding eigenvectors for A and B. Resulting zero value corresponds to orthogonal motions between trajectories while on the other hand, unity values correspond to high degree of overlap or similarity.

The trajectory conformational variability has been used for the detection of differences not only between trajectories but between different parts of the same trajectory:

$$\kappa_{AB} = 2 \frac{\gamma_{AB}}{(\gamma_{AA}^T + \gamma_{BB}^T)} \tag{2.68}$$

where κ_{AB} is the **relative similarity index**, γ_{XX}^T corresponds to the absolute autosimilarity index for the trajectory X obtained for their first and second halves. It is worth to say that these measurements do not take into account the different contributions of eigenvectors to the total flexibility so they are equally considered. For this reason, more complex methods which include the associated eigenvalues have been

developed (Pérez et al. 2005):

$$\xi_{AB} = \frac{2\sum_{i=1}^{i=z}\sum_{j=1}^{j=z}\left((v_i^A \cdot v_j^B)\frac{exp\left\{-\frac{(\Delta x)^2}{\lambda_i^A}-\frac{(\Delta x)^2}{\lambda_j^B}\right\}}{\sum_{i=1}^{i=z}exp\left\{-\frac{(\Delta x)^2}{\lambda_i^A}\right\}\sum_{j=1}^{j=z}exp\left\{-\frac{(\Delta x)^2}{\lambda_j^B}\right\}}\right)^2}{\sum_{i=1}^{i=z}\left(\frac{exp\left\{-2\frac{(\Delta x)^2}{\lambda_i^A}\right\}}{\left(\sum_{i=1}^{i=z}exp\left\{-\frac{(\Delta x)^2}{\lambda_i^A}\right\}\right)^2}\right)^2+\sum_{j=1}^{j=z}\left(\frac{exp\left\{-2\frac{(\Delta x)^2}{\lambda_j^B}\right\}}{\left(\sum_{j=1}^{j=z}exp\left\{-\frac{(\Delta x)^2}{\lambda_j^B}\right\}\right)^2}\right)^2} \quad (2.69)$$

where λ_i is the eigenvalue associated to the eigenvector i with unit vector v_i. The sum can be calculated for all (z = m) or for a number of eigenvectors (z = n).

Entropy calculation can be done in principle according to Shannon definition, but this requires knowledge of the all M possible states in a finite configurational space (equation 2.67). This creates a clear obstacle since a complete sampling is not possible, forcing the use of approximated methods to compute entropy (Andricioaei & Karplus 2001; Schlitter & Klähn 2003). Both Andricioaei-Karplus and Schlitter methods are based on the computation of the main components derived from the covariance matrix diagonalization. While the Andricioaei-Karplus method (equation 2.68) is based on a quantic harmonic oscillator model, the Schlitter method (equation 2.69) uses a classic-quantic hybrid method. Both methods transform the calculated eigenvalues in entropy values giving similar results in most cases. It is worth to note that entropy values are dependent on the simulation time used to compute the corresponding eigenvalues but this can be avoided by extrapolating the calculated total entropy for different trajectory fragments and fitting some parameters *a posteriori* (equation 2.70) (Harris et al. 2001).

$$S = -k_B\sum_{i=1}^{M}P_i lnP_i \quad (2.70)$$

$$S = k\sum_i \frac{\alpha_i}{e^{\alpha}-1} - ln(1-e^{-\alpha_i}) \quad (2.71)$$

$$S = \frac{k}{2}\sum_{i=1} ln(1+\frac{e^2}{\alpha_i^2}) \quad (2.72)$$

$$S(t) = S_\alpha - \frac{\alpha}{t^\beta} \quad (2.73)$$

where α and β are the fitting parameters to which the different entropy values has been computed and t is the simulation time.

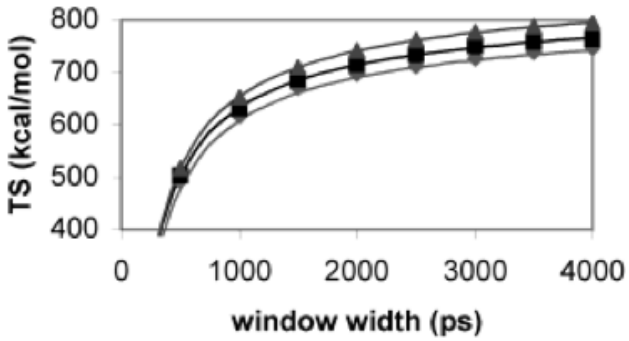


Figure 2.12: Effect of sample window width on the values of the configurational entropies, calculated by the method of Schlitter (Schlitter & Klähn 2003). The points are the experimental values, the lines are the results of the least-squares fit to the function given in Eq. 2.53 (From Harris et al. 2001).

As it has been stated before, internal coordinates can be used to describe the base pair relative position, and the local associated parameters can be related to the DNA local deformability. The **stiffness matrix** can be created by calculating a 6x6 covariance matrix for the 3 translations and 3 rotations. However, since matrix diagonalization is not applied in this case, the resulting vectors are not orthogonal and the out-of-diagonal values are non-zero indicating that cross terms between different helical parameters affect each other:

$$F = k_B T C^{-1} \tag{2.74}$$

where F is the stiffness matrix, C represents the covariance matrix, T is the temperature and k_B the Boltzmann constant. **Deformation energy** can be computed from the derived stiffness constants for any dinucleotide step (neighboring base pairs) (E_{XX}) assuming an harmonic oscillator behavior:

$$E_{XX} = \sum_{i}^{6} \sum_{j}^{6} \frac{k_{ij}}{2} (X_i - \bar{X}_i)(X_j - \bar{X}_j) \tag{2.75}$$

where k_{ij} represents the associated force constant to the i and j helical parameters and $\bar{X}$ corresponds to reference value for the helical parameter and XX the corresponding base pair step.

2.3 References

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

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

Bader, R., 1985. First-Principles Approach to Vibrational Spectroscopy of Biomolecules. Accounts of Chemical Research, 18, pp.9–15.

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

Bennett, C.H., 1976. Efficient estimation of free energy differences from Monte Carlo data. Journal of Computational Physics.

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

Beveridge, D.L. & DiCapua, F.M., 1989. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. Annual Review of Biophysics and Biophysical Chemistry, 18(1), pp.431–492.

Binkley, J.S. & Pople, J.A., 1975. Møller-Plesset theory for atomic ground state energies. International Journal of Quantum Chemistry, 9(2), pp.229–236.

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

Boys, S.F. & Bernardi, F., 1970. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Molecular Physics, 19(4), pp.553–566.

Chalasinski, G. & Szczesniak, M.M., 1994. Origins of Structure and Energetics of van der Waals Clusters from ab Initio Calculations. Chemical Reviews, 94(7), pp.1723–1765.

Claverie, P., 1978. In Intermolecular Interactions: From Diatomics to Biomolecules B. Pullman, ed., J Wiley.

Cornell, W.D. et al., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society, 117(19), pp.5179–5197.

Cox, S.R. & Williams, D.E., 1981. Representation of the molecular electrostatic potential by a net atomic charge model. Journal of Computational Chemistry, 2(3), pp.304–323.

Curutchet, C., Orozco, M. & Luque, F.J., 2001. Solvation in octanol: parametrization of the continuum MST model. Journal of Computational Chemistry, 22(11), pp.1180–1193.

Darve, E., Rodríguez-Gómez, D. & Pohorille, A., 2008. Adaptive biasing force method for scalar and vector free energy calculations. The Journal of Chemical Physics, 128(14), p.144120.

Denning, E.J. et al., 2011. Impact of 2'-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA. Journal of Computational Chemistry, 32(9), pp.1929–1943.

Dewar, M.J. et al., 1985. AM1: a new general purpose quantum mechanical molecular model. Journal of the American Chemical Society, 107(13), pp.3902–3909.

Dixon, S.L. & Merz, K.M., 1997. Fast, accurate semiempirical molecular orbital calculations for macromolecules. The Journal of Chemical Physics, 107(3), pp.879–893.

Dunning, T.H., 1989. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. The Journal of Chemical Physics, 90(2), p.1007.

Dupradeau, F.-Y. et al., 2010. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. Physical Chemistry Chemical Physics, 12(28), pp.7821–7839.

Faustino, I. et al., 2009. Unique tautomeric and recognition properties of thioketothymines? Journal of the American Chemical Society, 131(35), pp.12845–12853.

Faustino, I., Pérez, A. & Orozco, M., 2010. Toward a consensus view of duplex RNA flexibility. Biophysical Journal, 99(6), pp.1876–1885.

Gear, C.W., 1971. The automatic integration of ordinary differential equations. Communications of the ACM, 14(3), pp.176–179.

Grimme, S., 2004. Accurate description of van der Waals complexes by density func-

tional theory including empirical corrections. Journal of Computational Chemistry, 25(12), pp.1463–1473.

Grimme, S., 2006. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. Journal of Computational Chemistry, 27(15), pp.1787–1799.

Fonseca Guerra, C. et al., 2009. Adenine versus guanine quartets in aqueous solution: dispersion-corrected DFT study on the differences in π-stacking and hydrogen-bonding behavior. Theoretical Chemistry Accounts, 125(3-6), pp.245–252.

Hariharan, P.C. & Pople, J.A., 1973. The influence of polarization functions on molecular orbital hydrogenation energies. Theoretical Chemistry Accounts, 28(3), pp.213–222.

Harris, S.A. et al., 2001. Cooperativity in Drug-DNA Recognition: A Molecular Dynamics Study. Journal of the American Chemical Society, 123(50), pp.12658–12663.

Hart, K. et al., 2012. Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. Journal of Chemical Theory and Computation, 8(1), pp.348–362.

Helgaker, T. et al., 1997. Basis-set convergence of correlated calculations on water. The Journal of Chemical Physics, 106(23), pp.9639–9646.

Hehre, W.J. et al., 1986. Ab initio molecular orbital theory, Wiley (New York).

Hehre, W.J., 1969. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. The Journal of Chemical Physics, 51(6), pp.2657–2664.

Hess, B., 2002. Convergence of sampling in protein simulations. Physical review. E, Statistical, nonlinear, and soft matter physics, 65(3 Pt 1), p.031910.

Hess, B., 2000. Similarities between principal components of protein dynamics and random diffusion. Physical Review E.

Hockney, R.W. & Eastwood, J.W., 1988. Computer Simulation Using Particles, Taylor & Francis Group.

Hohenberg, P., 1964. Inhomogeneous Electron Gas. Physical Review, 136(3B), pp.B864-B871.

Hoover, W., 1985. Canonical dynamics: Equilibrium phase-space distributions. Physical review. A, 31(3), pp.1695–1697.

Hospital, A. et al., 2012. MDWeb and MDMoby: An integrated web-based platform for molecular dynamics simulations. Bioinformatics, 28(9), pp.1278–1279.

Hospital, A. et al., 2013. NAFlex: a web server for the study of nucleic acid flexibility. Nucleic Acids Research, 41(W1), pp.W47–W55.

Jenkins, H.D.B. et al., 2000. Basis set and correlation effects in the calculation of accurate gas phase dimerization energies of two M+ 2 to give M 2+ 4 (M= S, Se). Journal of Computational Chemistry, 21(3), pp.218–226.

Jeziorski, B., Moszynski, R. & Szalewicz, K., 1994. Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes. Chemical Reviews, 94(7), pp.1887–1930.

Jorge, M. et al., 2010. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration. Journal of Chemical Theory and Computation, 6(4), pp.1018–1027.

Jorgensen, W.L. & Ravimohan, C., 1985. Monte Carlo simulation of differences in free energies of hydration. The Journal of Chemical Physics, 83(6), p.3050.

Jorgensen, W.L., 1989. Free energy calculations: a breakthrough for modeling organic chemistry in solution. Accounts of Chemical Research, 22(5), pp.184–189.

Kirkwood, J.G., 1935. Statistical Mechanics of Fluid Mixtures. The Journal of Chemical Physics, 3(5), pp.300–313.

Kirkwood, J.G., 1968. Theory of liquids, Gordon & Breach Science Pub.

Kohn, W. & Sham, L.J., 1965. Self-Consistent Equations Including Exchange and Correlation Effects. Physical Review, 140(4A), pp.A1133–A1138.

Kollman, P., 1993. Free energy calculations: Applications to chemical and biochemical phenomena. Chemical Reviews, 93(7), pp.2395–2417.

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

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

Lavery, R. et al., 2009. Conformational analysis of nucleic acids revisited: Curves. Nucleic Acids Research, 37(17), pp.5917–5929.

Leach, A.R., 2001. Molecular modelling: principles and applications Addison-Wesley Longman Ltd, Addison-Wesley Longman Ltd.

Lewars, E.G., 2011. Computational Chemistry, Springer Verlag.

Lu, N., Singh, J.K. & Kofke, D.A., 2003. Appropriate methods to combine forward and reverse free-energy perturbation averages. The Journal of Chemical Physics.

Lu, X.-J. & Olson, W.K., 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Research, 31(17), pp.5108–5121.

Luque, F.J., Bachs, M., et al., 1996a. Extension of MST/SCRF method to organic solvents:Ab initio and semiempirical parametrization for neutral solutes in CCl4. Journal of Computational Chemistry, 17(7), pp.806–820.

Luque, F.J., Zhang, Y., et al., 1996b. Solvent Effects in Chloroform Solution: Parametrization of the MST/SCRF Continuum Model. The Journal of Physical Chemistry, 100(10), pp.4269–4276.

MacKerell, A.D., Banavali, N.K. & Foloppe, N., 2000. Development and current status of the CHARMM force field for nucleic acids. Biopolymers, 56(4), pp.257–265.

MacKerell, A.D., Feig, M. & Brooks, C.L., 2004. Improved treatment of the protein backbone in empirical force fields. Journal of the American Chemical Society, 126(3), pp.698–699.

MacKerell, A.D., Feig, M. & Brooks, C.L., 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. Journal of Computational Chemistry, 25(11), pp.1400–1415.

Metropolis, N. et al., 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 21(6), p.1087.

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

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

Momany, F.A., 1978. Determination of partial atomic charges from ab initio molecular

electrostatic potentials. Application to formamide, methanol, and formic acid. The Journal of Physical Chemistry, 82(5), pp.592–601.

Mulliken, R.S., 1955. Electronic Population Analysis on LCAO[Single Bond]MO Molecular Wave Functions. I. The Journal of Chemical Physics, 23(10), pp.1833–1840.

Møller, C. & Plesset, M.S., 1934. Note on an Approximation Treatment for Many-Electron Systems. Physical Review, 46(7), pp.618–622.

Nosé, S., 1984. A molecular dynamics method for simulations in the canonical ensemble. Molecular Physics, 52(2), pp.255–268.

Olson, W.K. et al., 2001. A standard reference frame for the description of nucleic acid base-pair geometry. Journal of Molecular Biology, 313(1), pp.229–237.

Orozco, M. et al., 2003. Theoretical methods for the simulation of nucleic acids. Chemical Society Reviews, 32(6), p.350.

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

Perdew, J. & Wang, Y., 1992. Accurate and simple analytic representation of the electron-gas correlation energy. Physical review. B, Condensed matter, 45(23), pp.13244–13249.

Pérez, A. et al., 2005. Exploring the Essential Dynamics of B-DNA. Journal of Chemical Theory and Computation, 1(5), pp.790–800.

Pérez, A. et al., 2008. Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Research, 36(7), pp.2379–2394.

Pérez, A., Luque, F.J. & Orozco, M., 2007a. Dynamics of B-DNA on the Microsecond Time Scale. Journal of the American Chemical Society, 129(47), pp.14739–14745.

Pérez, A., Marchán, I., et al., 2007b. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. Biophysical Journal, 92(11), pp.3817–3829.

Pérez, A., Luque, F.J. & Orozco, M., 2012. Frontiers in Molecular Dynamics Simulations of DNA. Accounts of Chemical Research, 45(2), pp.196–205.

Pierotti, R.A., 1976. A scaled particle theory of aqueous and nonaqueous solutions. Chemical Reviews.

Roothaan, C., 1951. New Developments in Molecular Orbital Theory. Reviews of Modern Physics, 23(2), pp.69–89.

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

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

Schlitter, J. & Klähn, M., 2003. A new concise expression for the free energy of a reaction coordinate. The Journal of Chemical Physics, 118, p.2057.

Shavitt, I., 1998. The history and evolution of configuration interaction. Molecular Physics, 94(1), pp.3–17.

Shirts, M.R. & Pande, V.S., 2005. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. The Journal of Chemical Physics, 122(14), p.144107.

Slater, J.C., 1930. Atomic Shielding Constants. Physical Review, 36(1), pp.57–64.

Soliva, R., Orozco, M. & Luque, F.J., 1997. Suitability of density functional methods for calculation of electrostatic properties. Journal of Computational Chemistry, 18(8), pp.980–991.

Sponer, J., Leszczynski, J. & Hobza, P., 1996. Nature of Nucleic Acid-Base Stacking: Nonempirical ab Initio and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs. The Journal of Physical Chemistry, 100(13), pp.5590–5596.

Stewart, J.J.P., 1989. Optimization of parameters for semiempirical methods I. Method. Journal of Computational Chemistry, 10(2), pp.209–220.

Szalewicz, K., 2012. Symmetry-adapted perturbation theory of intermolecular forces. WIREs Computational Molecular Science, 2(2), pp.254–272.

Tembre, B.L. & Mc Cammon, J.A., 1984. Ligand-receptor interactions. Computers & Chemistry, 8(4), pp.281–283.

Torrie, G.M. & Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics, 23(2), pp.187–199.

Truhlar, D.G., 1998. Basis-set extrapolation. Chemical Physics Letters, 294(1), pp.45–48.

Tsuzuki, S., Uchimaru, T. & Tanabe, K., 1998. Intermolecular interaction potentials of methane and ethylene dimers calculated with the Møller–Plesset, coupled cluster and density functional methods. Chemical Physics Letters, 287(1-2), pp.202–208.

Tuckerman, M., Berne, B.J. & Martyna, G.J., 1992. Reversible multiple time scale molecular dynamics. The Journal of Chemical Physics, 97(3), p.1990.

Vanquelef, E. et al., 2011. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucleic Acids Research, 39(Web Server issue), pp.W511–7.

Verlet, L., 1967. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review, 159(1), pp.98–103.

Vosko, S.H., Wilk, L. & Nusair, M., 1980. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. Canadian Journal of Physics, 58(8), pp.1200–1211.

Wang, B. et al., 2004. Fast semiempirical calculations for nuclear magnetic resonance chemical shifts: a divide-and-conquer approach. The Journal of Chemical Physics, 120(24), pp.11392–11400.

Watts, J.D., Gauss, J. & Bartlett, R.J., 1992. Open-shell analytical energy gradients for triple excitation many-body, coupled-cluster methods: MBPT(4), CCSD+T(CCSD), CCSD(T),and QCISD(T). Chemical Physics Letters, 200(1-2), pp.1–7.

Wiberg, K.B. & Rablen, P.R., 1993. Comparison of atomic charges derived via different procedures. Journal of Computational Chemistry, 14(12), pp.1504–1518.

Woodcock, L.V., 1971. Isothermal molecular dynamics calculations for liquid salts. Chemical Physics Letters, 10(3), pp.257–261.

Young, D.C., 2001. Computational chemistry: a practical guide for applying techniques to real world problems Wiley-Interscience, Wiley-Interscience.

Zgarbová, M. et al., 2011. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. Journal of Chemical Theory and Computation, 7(9), pp.2886–2902.

Zhao, Y. & Truhlar, D.G., 2007. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-

class functionals and 12 other functionals. Theoretical Chemistry Accounts, 120(1-3), pp.215–241.

Zwanzig, R.W., 1954. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. The Journal of Chemical Physics.

3 The structure of nucleic acids

There is no need
to know everything
but to know where to find it.
My father (1946-)

Genomic DNA is around 990 mm long in a single human cell containing more than 3900 Mbp (mega base pairs). This thread-like polymer is made up of monomers called nucleotides which have mostly evolved to a common set of *letters*. In the 1950s, DNA was unexpectedly found to show a simple secondary structure and in the 1960s, it became clear that RNA and DNA helices were able to coexist in different alloforms, the most important called A and B respectively. In fact, we know nowadays that all important polynucleotide secondary structures are all helical (single, double, triple and quadruple) and they have seen as 'boring' repeating units by many. However, differences in base sequence can have an impact in their characteristic signature becoming more or less rigid, being more or less prone to interact with small molecules interactions and proteins. But, what makes nucleic acids to have such different behaviors?

3.1 The building blocks: physical properties of nucleosides and nucleotides

All nucleotides are composed of three main components, i.e. the nitrogen base, the ribose sugar and the phosphate group. While DNA and RNA share some of the nitrogen bases (adenine (A), guanine (G) and cytosine (C)), DNA completes its four-letters-set with thymine (T) which becomes methylated from their RNA counterpart uracil (U).

The second main chemical difference of RNA is placed in the ribose sugar where a hydroxyl group is located in the C2' position which gives their name to RNA (ribonucleic acid) whereas the monomers of DNA are 2'-deoxyribonucleotides.

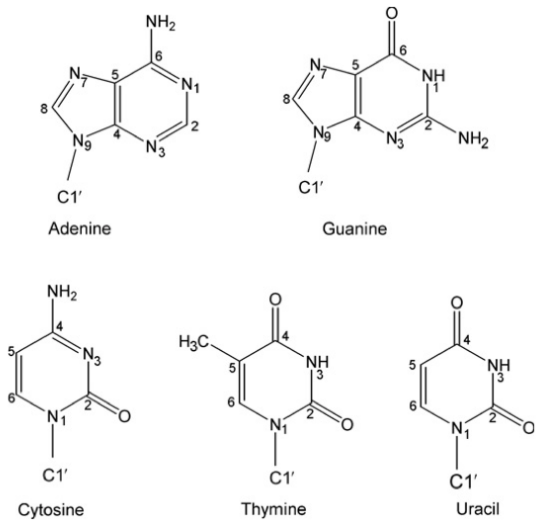


Figure 3.1: Structures of the major purine and pyrimidine bases in nucleic acids in their most abundant tautomeric conformation (Neidle 2007).

As their own name stands, nucleic acids present an acid-base behavior which determines their charge and reactivity. All pK_a value of the nucleobases are in the the physiological range ($5 \leq$ pH ≤ 9) from which A, C and G become preferentially protonated on one of their ring nitrogens rather than on their exocyclic amino group since the corresponding ionization maintains the aromaticity of the system.

Nucleobases can have alternative tautomers depending on the position of hydrogen atoms. Since the mid twentieth century, it is believed that the tautomeric equilibrium favors the keto-amino tautomeric forms for coding nucleobases. However, tautomeric preferences are highly dependent on the exact nature of the nucleobase and on environmental conditions, having in solvent interactions a major contributor. In fact, theoretical and experimental measurements show that polar solvents can change the *in vacuo* intrinsic tautomeric preferences of the nucleobases increasing the population of the most polar tautomer (Colominas et al. 1996; Zhanpeisov & Leszczynski 1998; Zhanpeisov et al. 1998; Reichardt & Welton 2011). Changes in the close nucleobase environment or introduction of substituents in the purine/pyrimidine ring can modify the tautomeric preference and therefore, induce changes in recognition properties.

3.2 Base-base interactions: hydrogen bonded and stacking interactions

One of the first features that were discovered from DNA sequence analysis was the identical quantities of A and T and G and C nucleobases, something that Chargaff's experimental data (Chargaff 1950) pointed to be the key for mutual recognition and on which DNA replication and translation processes are based.

All nitrogen bases have several hydrogen bond acceptors and donors which allow several edges for interactions. While the Watson-Crick base pairing (WC) is the main interaction pattern between neighboring strands, there have been identified other kinds of interaction like wobble pairs (w) that allowed Francis Crick to explain the degeneracy of the genetic code, and Hoogsteen pairs (H) that it is thought to play an important role in protein-DNA recognition (Nikolova et al. 2012).

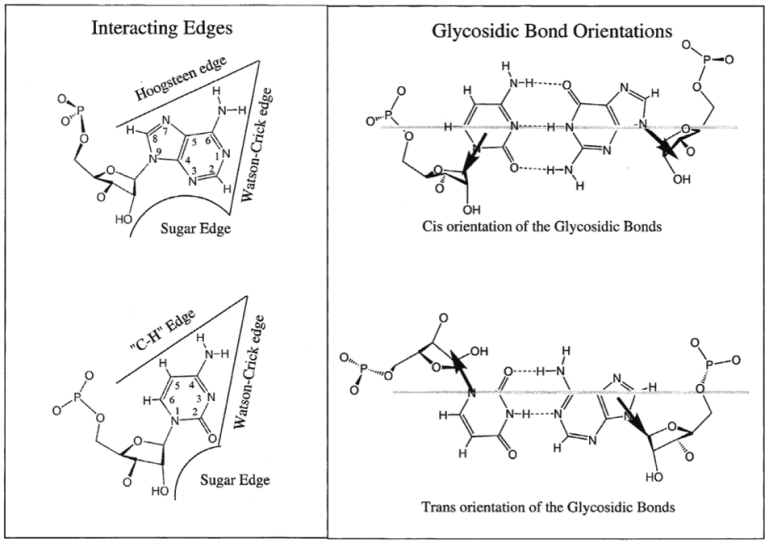


Figure 3.2: Watson-Crick base pairings for G·C and A·T base pairs and alternative Hoogsteen pairing for A·T (From Leontis 2001).

Double stranded helices are favored not only by hydrogen bonds but, even more important, by stacking interactions. While hydrogen bonds have an electrostatic nature, in general, stacking interactions are mainly due to the π orbital overlap between aromatic rings (i.e., dispersive term).Base stacking interactions have been shown to facilitate local conformational variability in DNA (Calladine 1982; Hunter 1993) probably through changes in the sugar-phosphate backbone which are known to be crucial in several protein-DNA interactions. Theoretical studies have shown the following order in stacking energies in gas phase for consecutive base pairs: R:R (purine:purine)

< Y:R (pyrimidine:purine) < Y:Y (pyrimidine:pyrimidine). Regarding hydrogen bond energetics, hydrogen bonds range from the most stable G·C base pair (-25.8 kcal/mol) to the least stable T·T (-10.6 kcal/mol) while the most stable stacked pair is the G//G dimer (-11.3 kcal/mol), and the least stable is the U//U dimer (-6.5 kcal/mol) (Sponer et al. 1996).

3.3 Conformational variability: sugar puckering, backbone torsions and helical parameters

The current view that sugar-phosphate backbone plays an active role in the sequence-dependence flexibility is strongly supported by correlation analysis of base pair dinucleotide step helical parameters and the corresponding backbone torsions. Up to seven different torsions can be identified in natural nucleic acids: α (P-O5'), β (O5'-C5'), γ (C5'-C4'), δ (C4'-C3'), ϵ (C3'-O3'), ζ (O3'-P), and the glycosidic bond χ (C1'-N9 for purines and C1'-N1 for pyrimidines). Some of them present strong self correlation like α/γ and ϵ/ζ torsions.

3.3.1 Sugar pucker

The internal concerted movements within the five-member ring result in interconvertible conformations which are separated by energy barriers. The ring conformation is typically defined by two parameters: τ_m, the amplitude of puckering, and P, the phase angle of pseudorotation (Altona & Sundaralingam 1972). The value of P is determined by the different internal sugar torsions about the pentose bonds τ_j where τ_1 = O4'-C1', τ_2 = C1'-C2', τ_3 = C2'-C3', τ_4 = C3'-C4', and τ_5 = C4'-O4', respectively:

$$tanP = \frac{(\tau_4 + \tau_1) - (\tau_3 + \tau_0)}{2\tau_2(sin36^\circ + sin72^\circ)} \tag{3.1}$$

and the amplitude of puckering:

$$\tau_m = \frac{\tau_2}{cosP} \tag{3.2}$$

Since most conformational changes do not alter the amplitude of puckering (around 38°), P usually drives the puckering location in the pseudorotation cycle (Figure 3.3). According to experimental data, both ribose and deoxyribose rings are confined in quite narrow regions in the pseudorotation cycle: 15-20°for north region, N, and 160-164°for south region, S, respectively (Figure 3.3).

A wide variety of ring pucker geometries have been observed in experimental structures. It is usual to see one of the ring atoms out of the plane of the other four in the so-called envelope pucker type. The most commonly observed envelope conformations in crystal structures are either close to C2'-endo or C3'-endo. The C2'-endo puckering, related to 2'-deoxyribonucleotides and south region in the pseudorotation cycle, corresponds to a C2' displacement out of the C1'-O4'-C4' plane on the same side as the base and C4'-C5' bond, while the C3'-endo puckering, characteristic of ribonucleotides and north region in the pseudorotation cycle, face the C3' out of the same plane on the same side as the base and C4'-C5' bond. In practice, pure envelope coexist with twist conformations, where the major displacement is on the endo side and a minor displacement is observed in the exo (opposite) side.

The simple substitution of the 2'-hydroxyl group in ribose by hydrogen in deoxyribose alters both the conformational preferences and the rate of pseudorotation between different pucker states although according to coupling constants measurements (Altona & Sundaralingam 1973), the transition between different puckerings is more difficult in ribose than in deoxyribose. Besides this, north $\rightarrow$ south transitions through the east region are preferred while the west region intermediates are barely observed. NMR measurements of the ratio of H1'-H2' and H3'-H4' coupling constants suggest as well a rapid interconversion between major conformations through a potential barrier around 3-5 kcal/mol (through east intermediates).

3.3.2 Glycosidic torsion χ

The bond linking the sugar-phosphate backbone to the nucleobase has two major conformations. The *anti* orientation has the H6(Y) or H8(R) atom pointing to the backbone atoms (-120 $> \chi >$ 180°) while the *syn* orientation keeps that hydrogen away from the phosphate group ($0 < \chi <$ 90°) therefore, avoiding the WC base pairing. Thus, the *anti* conformation is the general preferred, and even, there are experimental measurements which show that some bases can display *syn* conformation such as guanines in alternating oligomers of the form d(CpGpCpG).

The nucleobase involved in the glycosyl linkage influences the conformational variability. Theoretical (Foloppe & MacKerell 1999; Hocquet et al. 2000) and experimental measurements have shown that the anti purine distribution is shifted to the high anti region (237°$\pm$ 24°) and broader than for pyrimidines (230°$\pm$ 18°). MP2 and DFT calculations have shown that the anti energy well is shallower for purines than for pyrimidines with the sharpest slope for cytidine in the anti region (150°$< \chi <$ 240°). These differences have been explained by means of intramolecular interactions between the base

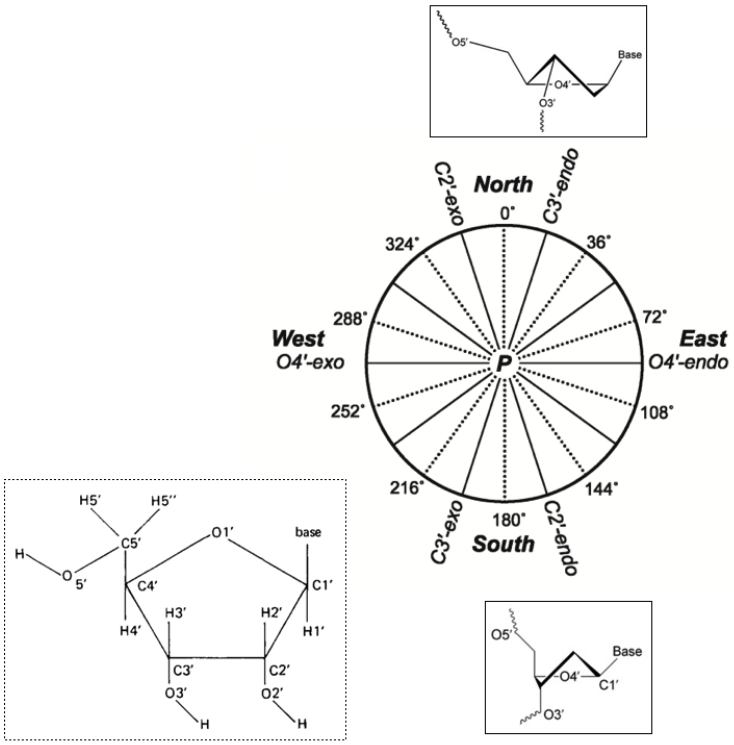
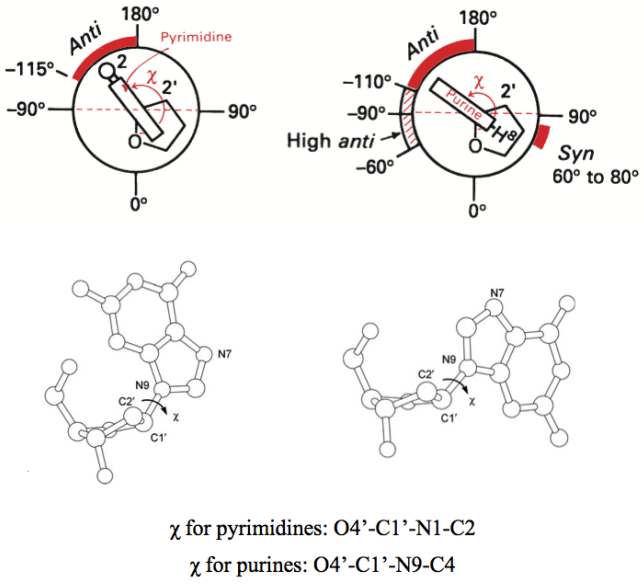


Figure 3.3: Pseudorotation phase angle (P) cycle with the range of angles for selected puckering types.



χ for pyrimidines: O4'-C1'-N1-C2

χ for purines: O4'-C1'-N9-C4

Figure 3.4: Anti and syn ranges for the glycosyl linkage in pyrimidine (top left) and purine (top right) nucleosides. Anti (bottom right) and syn (bottom left) conformations for guanosine (From Blackburn 2006).

and the O5' atom of the sugar moiety coupled with sugar puckering changes (Hocquet et al. 2000).

3.3.3 Concerted α/γ and ϵ/ζ torsions

As it was said at the beginning of this section, there are some torsions which are highly correlated. In fact, they are not only correlated but they are restricted to relatively narrow regions (Rao & Sundaralingam 1969; Olson & Sussman 1982). As a convention, backbone torsions are classified in three ranges with means around +60°(g^+, *gauche*$^+$), -60°(g^-, *gauche*$^-$) and 180°(t, *trans*). Thus, in the case of the α/γ torsions (O3'-P-O5'-C5' and O5'-C5'-C4'-C3' respectively) the canonical conformation corresponds to g^-/g^+ in B-DNA, but both the canonical α/γ:g^-/g^+ and the alternative α/γ:t/t conformations have been observed in A-DNA and RNA (Srinivasan & Olson 1987; Hartmann et al. 2003).

In the case of ϵ and ζ conformations (also highly correlated) are typically expressed in terms of BI (t/g^-) and BII (g^-/t) conformations which are the most abundant in DNA. It has been proposed that these states may play a role in its sequence-specific recognition by proteins (Gorenstein 1994; Pichler et al. 1999) through apparently non-specific contacts.

In addition, some dihedral angles are tightly constrained to the sugar pucker conformation such as dihedral angle related to pucker, δ, and ζ (Neidle & Balasubramanian 2006), and interdependence between different conformational angles (χ, α, γ, ...) is well established.

3.3.4 Helical parameters

Nucleic acids structures are typically described in terms of local and global structural parameters. These parameters define the relative position of the two bases in a base pair or the two base pairs in a base pair step (two consecutive base pairs) based on geometric definitions from the 1999 Tsukuba Accord (Olson et al. 2001). For the calculation of base pair parameters, each base is defined by an origin and a vector triad where the origin corresponds to the middle point between N1$\cdots$C4 distance for purines or N3$\cdots$C6 distance for pyrimidines. In addition, in the case of base pair step parameters, the base pair coordinate frame is defined by an origin, which is placed in the intersection of the perpendicular bisector of the C1'$\cdots$C1' vector and the vector connecting the pyrimidine Y(C6) and purine R(C8) atoms, and a vector triad (Figure

3.5). Therefore, base pair and base pair step helical parameters typically describe the relative rotations and translations of bases and base pairs respectively (Figure 3.6).

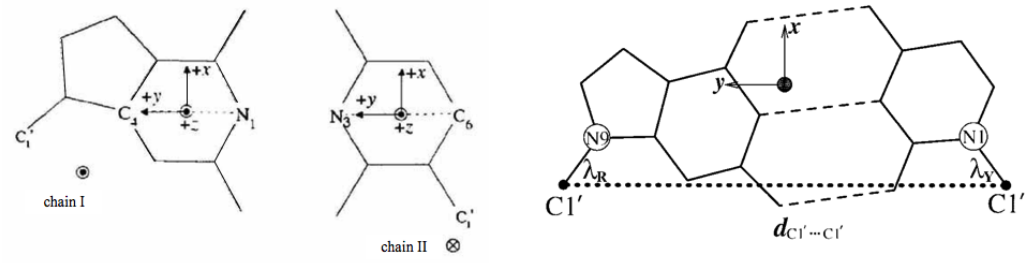


Figure 3.5: Coordinate frames for bases (on the left) and for a base pair (right) within idealized double stranded base pair. The (•) shows the axis origin and corresponding vector triads with x component pointing in the direction of the major groove, y follows the long axis of the base pair, and z is defined by x and y ($z = x \times y$) (From Olson et al. 2001).

- Base pair parameters include:
  - **buckle** (κ), **propeller twist** (ω), and **opening** (σ) are defined as the angle formed due to rotations about the x, y and z axis of the bases, respectively.
  - **shear** (S_x), **stretch** (S_y), and **stagger** (S_z) which define the relative translation between bases within base pair in the corresponding direction.
  - **inclination** (η) is defined as the angle between the long axis of the base pair and the plane perpendicular to the helix axis.
  - **X- and Y-displacements** define translations along the x and y axis of the base pair.
- Base pair step parameters:
  - **tilt** (τ) is defined as the rotation between successive base pairs along the x axis of the base pair.
  - **roll** (ρ) is defined as the rotation between base pairs along the y axis of the base pair.
  - **twist** (Ω) is defined as the rotation between consecutive base pairs.
  - and three translations: **shift** (D_x), **slide** (D_y), and **rise** (D_z) that correspond to the relative translational movement between consecutive base pairs along the corresponding x, y and z axis of the base pair.

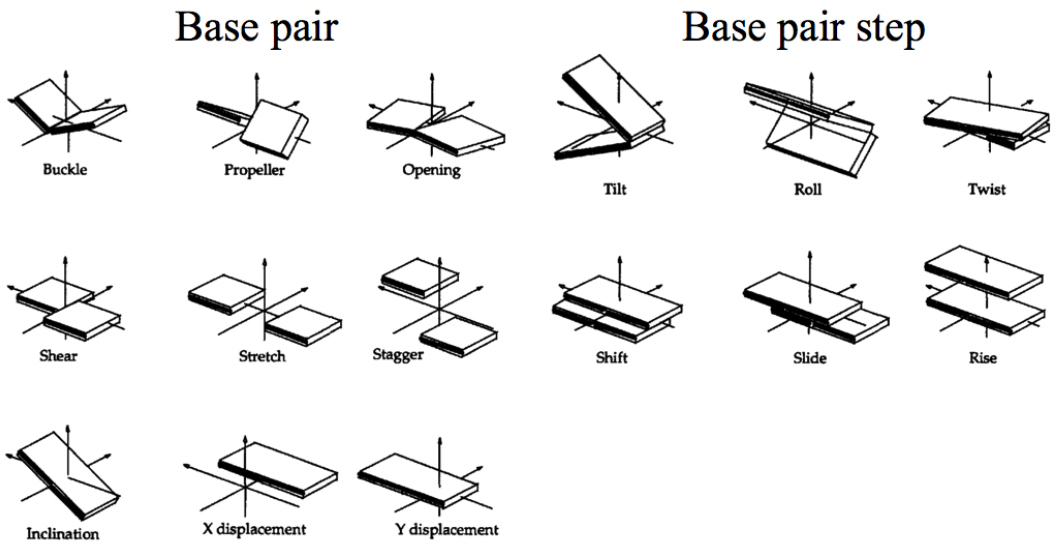


Figure 3.6: Helical parameters for base pairs (left) and base pair steps (right). Rotations are shown in the upper part and translations in the lower part.

The major and minor grooves emerge from base pair asymmetry. The glycosidic bonds (C1'-N9 in purines and C1'-N1 in pyrimidines) are by convention on the **minor groove** side, so the opposite side is called **major groove**.

3.4 Secondary structures in nucleic acids

Early diffraction studies on DNA fibres (Watson & Crick 1953), identified DNA duplex as highly polymorphic since this can form different conformations depending on sequence or environmental conditions. At low humidity and high salt (or water-ethanol mixtures), a more compact A-DNA conformation is favored whereas at high humidity and low salt concentration B-DNA conformation is dominant. There are several minor conformations (denoted as A', α-B', β-B', C, C', D, E, T and Z) which has been obtained in rather different conditions or have much less significance and there are different conditions, or have much less significance compared to those predominant species, A and B (Travers 1989).

The first X-ray studies on the so-called Dickerson-Drew dodecamer d(CGCGAATTCGCG) (Wing et al. 1980) at high humidity conditions revealed the main **B-DNA conformation** attributes, the dominant conformation for DNA and considered to be the most important in genomic DNA. B-DNA base pairs are placed in the center of the helix in perpendicular orientation to its axis which makes both minor and major grooves of similar depth. However, minor and major groove widths are clearly different with

major groove 6 Å wider than minor groove. Every base pair twist around 36°yielding an helical pitch of ∼10 bp per turn. Concerning the helical descriptors, B-DNA has little roll and small positive slide. The sugar pucker in the B-DNA duplex is typically restricted to the south/south-east region in the pseudorotation cycle related to C2'-endo sugar conformation.

	Unit repeat	Rise (Å)	Helical twist (°)	Base pair propeller twist (°)	Base step roll (°)	Base pair inclination (°)
A	11	2.54	32.7	-10.5	0.0	22.6
B	10	3.38	36.0	-15.1	0.0	2.8
Z(C)	6	7.25	-49.3	8.3	5.6	0.1
Z(G)	6	7.25	-10.3	8.3	-5.6	0.1

Table 3.1: Average helix parameters for the major DNA conformations (From Neidle 2007).

The **A-DNA conformation** is a right-handed and antiparallel double helix with 4.5 Å base displacement from the helical axis which creates a central hollow with a diameter around 3 Å. As mentioned before, A-DNA conformation is not only favored by low humidity or salt concentration but also, reversible B → A transition occur when deoxyribose is substituted by ribose or when N puckering is forced (Soliva et al. 1999). Every base pair has an average twist around ∼32°which gives a helical pitch of 11 residues per turn. In terms of helical parameters, A-conformation has high roll and negative slide. Since the average base-base distance between covalently connected residues is around 2.6 Å, base pairs must tilt in order to maintain the van der Waals separation with ∼20°inclination. Consequently, the A-DNA grooves present a deep major groove and a shallow minor groove, and even, the minor groove becomes wider than the major groove.

The **Z-DNA conformation** (Wang et al. 1979) is characterized by a levo- rather than dextro-chirality of A and B conformations. It mainly exists in regions with GpC alternance $d(GC)_n$, although it can occur in other base sequences at a higher cost (Haniford & Pulleyblank 1983), is presumed to be favored by high ionic concentrations (> 2.5 M NaCl) (Drew et al. 1980), negative supercoiling (Singleton et al. 1982) and protein binding (Ha et al. 2005; Li et al. 2006). It has been shown that substitution of bulky groups in C5 position of cytosine facilitate the transition from B-DNA conformation (Behe & Felsenfeld 1981) but the underlying mechanism remains unknown. Guanine residues typically adopt the *syn* conformation for their glycosidic bonds whereas *anti* conformation is typically found in cytosines. This alternating glycosidic bond conformation generates the typical zig-zag backbone (Figure 3.7). The cytosine sugar pucker remains in the C2'-endo conformation while the guanine's adopts the C3'-endo conformation (Rich 1995). In Z-DNA, twist values show very different tendencies since GpC steps present a very negative value (-50.6°) while CpG steps are characterized

by higher values (-9°). These differences do influence in the topology of the grooves generating a very deep minor groove while the major groove becomes extremely shallow. The Z-DNA conformation presents a longer structure with 12 base pairs per turn and a narrower central hole compared to B-DNA. Experimental data suggest that it might be involved in transcription (Oh et al. 2002; Rothenburg et al. 2001), chromatin remodeling (Liu et al. 2006) and recombination (Wang et al. 2006).

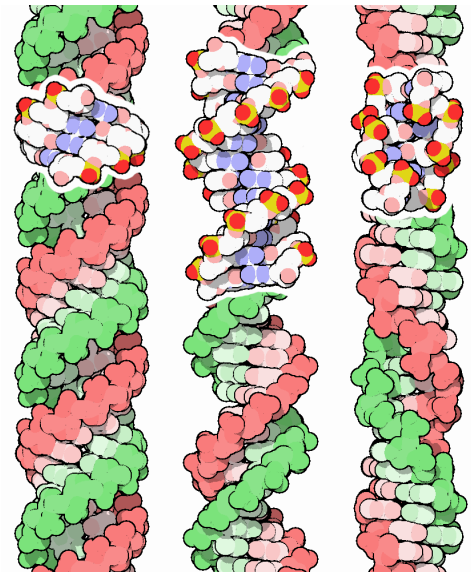


Figure 3.7: Double stranded DNA conformations. DNA can adopt three major conformations: A-DNA (left) with deep major groove and tipped bases and, B-DNA (center) with similar groove depths. However, the Z-DNA (right) adopts a left-handed structure with the deepest minor groove and almost invisible major groove. (From Protein Data Bank)

Sequence-dependence in DNA Analysis of helical parameters derived from an increasing number of x-ray and NMR structures and from molecular dynamics simulations of B-DNA has revealed on one hand, a number of correlations between helical parameters (Suzuki et al. 1997; Subirana & Faria 1997), and sequence-dependent features (Olson et al. 1998; Lankas et al. 2003; El Hassan & Calladine 1997; Dans et al. 2012; Maehigashi et al. 2012). According to the nearest-neighbor model, pyrimidine:purine steps (YpR) are typically more flexible (defined as low stiffness constants associated to base pair step helical deformations) since they oscillate between two major conformations defined by *high twist-positive slide-low roll* and *low twist-negative slide-high roll* parameters. By contrary, purine:pyrimidine steps (RpY) behave very rigid and RpR steps (or YpY) present an intermediate behavior (Figure 3.8). In fact, extended correlation analysis to backbone torsions show that conformational variability of YpR steps is strongly correlated with helical parameters (Figure 3.9) within the same step or in

the neighboring step, e.g., high twist values are coupled to higher backbone B_{II} populations (Djuranovic & Hartmann 2004; Dršata et al. 2012). These sequence-specific backbone variability points to a key role of certain backbone torsions in local flexibility and prove that extension of the nearest-neighbor model must be considered at least for some cases.

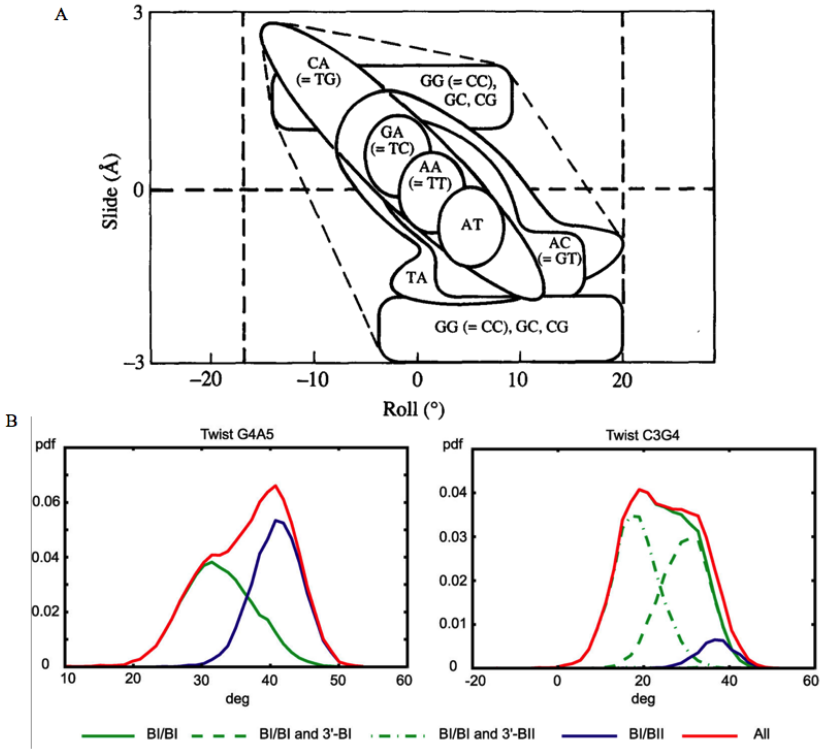


Figure 3.8: On top, correlation between slide and roll helical parameters for the ten representative steps according to the nearest-neighbor model (From El Hassan & Calladine 1996). Bottom, probability density plots of twist for G4A5 and C3G4 steps within Dickerson dodecamer from MD simulations (Dršata et al. 2012). Decomposition of twist values in backbone states within corresponding step and with 3' neighboring step showing strong correlation between twist and backbone conformations.

3.5 Non-canonical structures

Double-stranded DNA is the principal genetic molecule in most biological systems. However, non-B DNA structures are facilitated, perhaps transiently, at specific sequence motifs (e.g., expandable repeats) by negative supercoiling, which are generated in part by transcription, protein binding and other factors (Mirkin 2006; Wells 2007). Several non-B DNA conformations are known to exist in DNA repetitive sequences such as

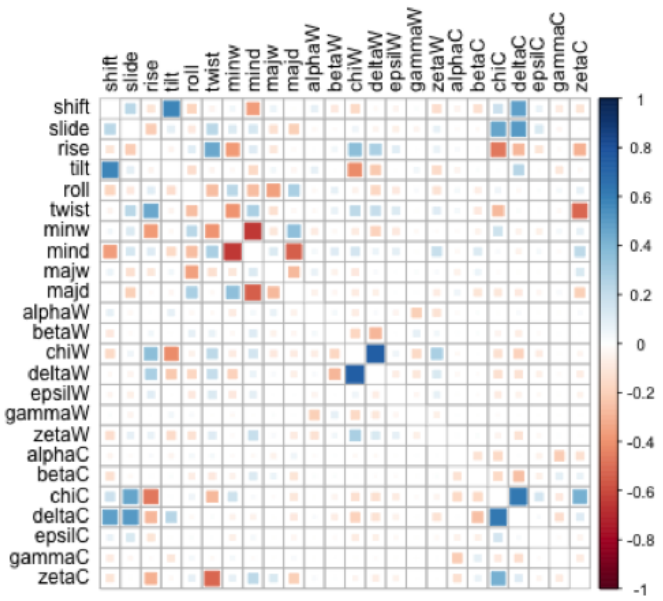


Figure 3.9: Correlation matrix for CG step between helical parameters and backbone torsions from both Watson (W) and Crick (C) strands. Color code denotes to correlation (positive value or blue) or anticorrelation (negative value or red). Several strong correlations are shown like twist and ζ torsion or glycosidic bond χ and sugar pucker related δ torsion.

triplexes, cruciforms, slipped structures, left-handed Z-DNA and quadruplexes (Figure 3.10). In this section I will focused on DNA triplex and G-quadruplex structures.

3.5.1 Triple-stranded DNA

Single-stranded triplex-forming oligonucleotide (TFO) can bind to the major groove of a double stranded DNA, generally to homopurine strands (Strobel & Dervan 1990). Their high affinity and specificity lead to the formation of DNA triplexes being involved in genomic events including inhibition of transcription and DNA replication, promotion of site-specific DNA damage, and enhancement of recombination (Davis & Kayser 2010; McNeer et al. 2011). This unique ability of TFOs to recognize duplexes forming triplexes has been exploited in several biotechnological applications such as antigene strategies or for site-specific delivery of mutagenic agents (Chin et al. 2007; Vasquez 2010).

Triplexes can be formed by inter- or intramolecular interaction with short oligonucleotides. Intramolecular triple helix requires a single strand with appropriate sequence to fold back on itself driven by supercoiling (Htun & Dahlberg 1989). In general,

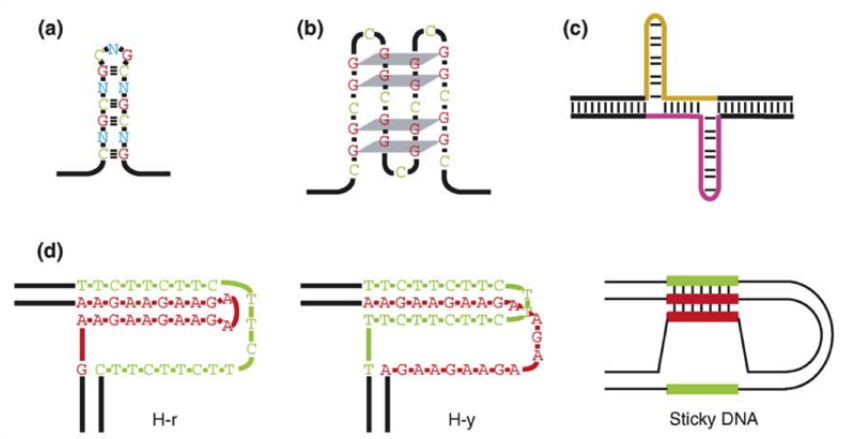


Figure 3.10: Unusual DNA structures formed by expandable repeats. (a) Imperfect hairpins composed of $(CNG)_n$ repeats. (b) G-quartets composed of $(CGG)_n$ repeats. (c) Slip-stranded DNA. (d) Various triplexes formed by $(GAA)_n$ repeats (only one possible conformation of sticky DNA is shown). In the repeats, purines are red and pyrimidines are green; flanking DNA is shown in black. (From Mirkin 2006).

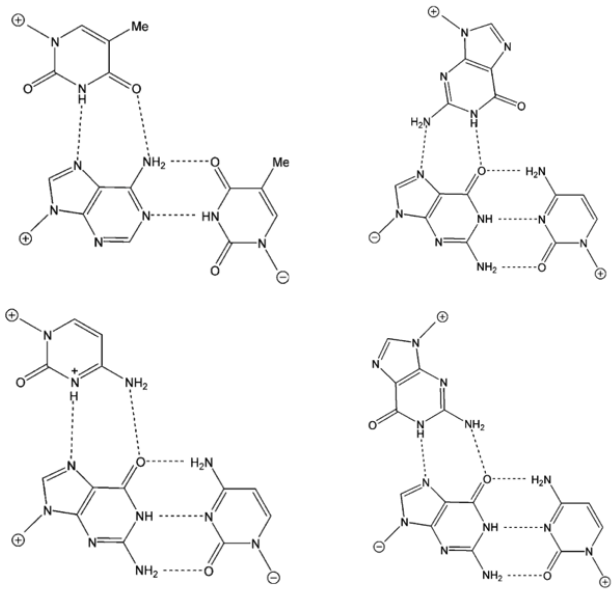


Figure 3.11: Hydrogen bonding within parallel and antiparallel base triplets. The TFO nucleobase within the triad is always showed in the upper part. Strand orientation is indicated by the $\oplus$ and $\ominus$ symbols (From Neidle 2007).

triplexes are classified as either parallel or antiparallel depending on the orientation of the TFO relative to the interacting homopurine strand of the double helix. Parallel triplexes present a homopyrimidine strand with either thymines or N3-protonated cytosines forming Hoogsteen base pairings (H) to bind to the dsDNA adopting $C^{+}*G{\cdot}C$ or $T*A{\cdot}T$ pairings (where '·' denotes WC pairing and '∗' represents Hoogsteen base pairing) (Figure 3.11). This requirement makes parallel triplexes extremely pH-dependent and rapidly unstable when pH rises above 5.0. In addition, antiparallel triplexes place a homopurine strand as TFO by reverse Hoogsteen (rH) base pairing between neutral nucleotides, avoiding the pH dependence from parallel triplexes and forming G#G·C or A#A·T pairings (where '#' stands for reverse Hoogsteen pairing) (Figure 3.11). Not only homopurine or homopyrimidine TFO sequences do not exclusively form triple helices but alternative GT-rich TFOs can also undergo both parallel and antiparallel orientations depending on their sequence, although these have been less studied and seems less stable (Thuong & Hélène 1993).

From the thermodynamical point of view, triple helices are in general less stable than corresponding duplexes. However, triplex stability can be enhanced by using modified nucleotides like 5-methylcytosine instead of cytosine or 5-bromouracil instead of thymine and even more stabilizing is the use of ribonucleotides in the TFO (Roberts & Crothers 1992). Peptide nucleic acids (PNA) sequences have been also shown to stabilize the complex with double stranded DNA molecules (Nielsen et al. 1991; Betts et al. 1995). An important drawback of G-rich TFOs is their tendency to form highly stable aggregates such as G-quadruplexes instead of triplexes which it is thought to occur through the formation of a G-triplex structural motif (G#G·G triad) (Limongelli et al. 2013). In fact, the substitution of heavy atoms in standard residues in the TFO has been shown to reduce the competition between triplex and tetraplex formation favoring triplex formation. In this regard, the theoretical study on seleno-modified guanines undertook in this PhD (publication in preparation) opens a potential solution for the study of biological processes which are triplex-forming mediated.

3.5.2 G-quadruplex DNA

G-rich DNA can form a highly compacted structure named G-quadruplex, which is defined as a right handed oligo with G-tetrads formed by four guanines forming eight Hoogsteen hydrogen bonds (Gellert et al. 1962). G-rich DNA sequences have attracted great attention not only due to its extreme stability, but because they have been proved to form *in vivo* during the cell-cycle progression (Granotier et al. 2005; Biffi et al. 2013). Genomic G-rich sequences mainly include telomeric regions located at chromosomic ends where repetitive G sequences end in a single stranded 3' overhang. Other repetitive

G-rich sequences have been found within rDNA (Hanakahi et al. 1999), immunoglobulin heavy chain switch regions (Dempsey et al. 1999) and in some hypervariable regions (Buroker et al. 1987; Wong et al. 1987) which enhance the biological relevance of these structures with a protective role against genomic instability (Eddy & Maizels 2006).

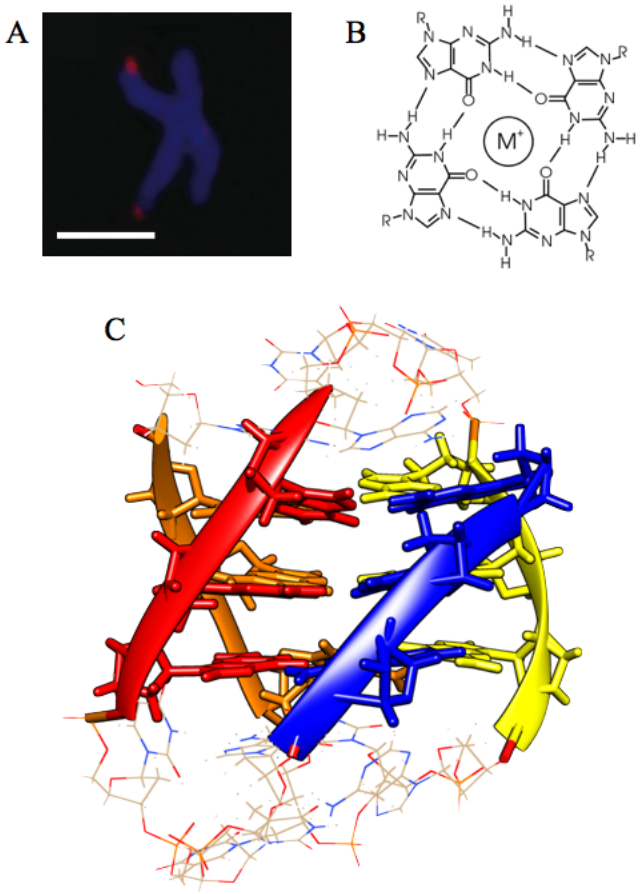


Figure 3.12: Localization of G-quadruplexes in chromosomic telomeres (A) (Biffi et al. 2013). Four guanines forming a G-tetrad structure with cation in the center (B). First NMR structure from (Wang & Patel 1993) (C).

Each G-quartet is formed by four co-planar guanines with each guanine pairing with two neighbors by Hoogsteen bonding. DFT calculations have shown that the strong preference of guanine to form quartets arises from a cooperative effect that strengthens the hydrogen bonds within the G-quartet network relative to those isolated G dimers (Otero et al. 2005). To this hydrogen bonding network, stacking interactions must be added between parallel G-tetrads with a twist of 30°and 3.3 Å separation between consecutive tetrads. This geometric arrangement assures that every O6 guanine atom points to the central core of the tetrad which is big enough to accommodate a stabilizing cation which neutralize the electrostatic repulsion (Arnott et al. 1974). In

fact, G4 DNA forms spontaneously in physiological salt levels, being stabilized by K^+ concentrations at around 10mM, lower than typical levels of mammal cells (120mM) (Maizels 2006).

To date, quadruplexes studied by crystallography and NMR has revealed a variability of topologies and structures which is not present in other type of DNA (Neidle & Balasubramanian 2006):

- The simplest topology, termed *intermolecular or tetramolecular quadruplexes*, are formed by self-association of four DNA strands such as $d(X_nG_pX_n)$ (where X is any nucleotide and G stands for guanine) which, although four different ways of association are possible, only a parallel arrangement has been observed in solution (Figure 3.12e) (Laughlan et al. 1994; Phillips et al. 1997).

- *Bimolecular quadruplexes* are formed by two strands of the general form $d(X_nG_o$-$X_pG_oX_n)$ (where G_o represents any number of guanines forming a tetrad and X_p is any nucleotide involved in loop formation) which associate to form a four-stranded G-quadruplex. Unlike intermolecular quadruplexes, there is no preference on how to associate and can present different types of loops (diagonal, lateral or external). The nature of nucleotides forming the loops have been shown to be crucial for the stability of the quadruplex (Smirnov & Shafer 2000; Risitano & Fox 2003).

- *Intramolecular quadruplexes* are formed by single oligonucleotides with general sequence $d(X_nG_oX_pG_oX_pG_oX_pG_oX_n)$ and can fold in a variety of structures according to several X-ray (Parkinson et al. 2002; Padmanabhan et al. 1993) and NMR examples (Wang & Patel 2004; Phan et al. 2004). Several loop combinations are possible without any preference.

C-rich strands containing a number of protonated cytosines can also form quadruplex structures by creating *i-motifs*. However, acidic conditions (pH < 4.5) or higher temperatures are required to first dissociate from Watson-Crick G·C pairing in DNA duplex interactions.

3.6 Effect of water and ions

Water, usually the solvent for any relevant nucleic acid structure, has been considered as the 'fourth component' of DNA structure after bases, sugars and phosphates. However, the polymorphic equilibrium depends greatly on the concentration of water molecules and salt concentration. The characteristic hydrogen bonding features and the high dipolar moment of water favor the interaction with functional groups within the

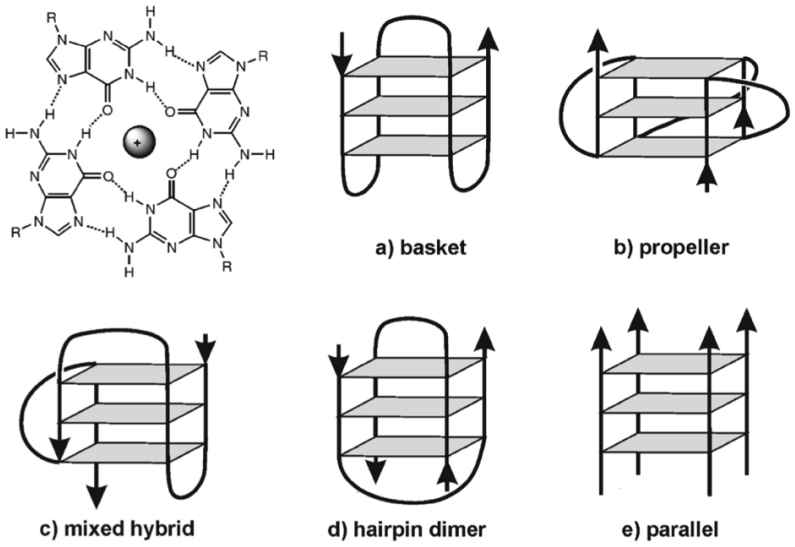


Figure 3.13: Schematic representation of complexity loops for intramolecular (a, b, c), bimolecular (d) and tetramolecular (e) G-quadruplexes (From Neidle &Balasubramanian 2006).

nucleobases. Up to 14 unique waters per base pair can be observed in high resolution X-ray structures.

Since the B-DNA conformation is favored under high water activity conditions, highly ordered water molecules can be seen in both major and minor grooves. The wider major groove is typically covered by a molecular layer of water molecules that directly interact with exposed hydrogen bonding acceptors and donors and also with phosphate groups. Experimental measurements suggest that DNA phosphates bind waters very strongly as expected due to their negative charge. The narrow minor groove usually contains the so-called spine of hydration which, specially in AT base pairs, consists in two water layers: first, a group of alternating water molecules buried in the floor of the groove which directly interact with the nucleobases, and the second is located around the first group of waters (Tereshko et al. 1999). Highly conserved clusters of water molecules are usually found around minor-groove ligand binders, clusters that have been suggested to have a stabilizing role. 2'-Hydroxyl groups of RNA not only attract water molecules into the shallow minor groove, but these interactions have been proved to be important in the thermodynamic stabilization of RNA molecules (Egli et al. 1996).

In triple stranded structures, any of the three grooves represent potential sites of

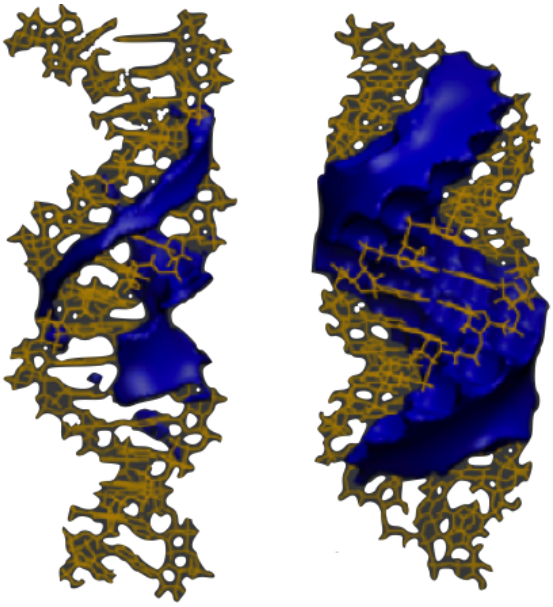


Figure 3.14: Water molecular interaction potential maps for one of the four sequences studied in Faustino et al. 2010, being DNA on the left and RNA on the right computed for the corresponding time-averaged structure. Contour plots showed here correspond to -5 kcal/mol in both cases.

interaction with ligands (Shields et al. 1997). However, among the three grooves, the minor-major groove is quite narrow (∼2-3 Å) and long-lived water molecules can be detected in this groove. Such water molecules can contribute to the stability of the molecule by screening the electrostatic repulsion of close phosphate groups. The minor-minor groove shows instead very similar features with the minor groove of B-DNA duplexes although the spine of hydration is not so clear. The third and widest groove, the major-major groove, shows instead a less clear hydration network due to the presence of methyl groups of thymine residues in parallel triplexes. However, long-lived water molecules have been observed along the major-major groove of antiparallel triplexes suggesting that they play an important role in stability of this type of triplexes (Radhakrishnan & Patel 1994).

The G-quadruplex grooves contain also an ordered network of water molecules specifically about the exocyclic amino groups N2, and the heterocyclic N3 atoms. While the geometric symmetry of parallel stranded G-quadruplexes bring a symmetric hydration in the four grooves, a less ordered hydration network is observed in antiparallel G-quadruplexes. The minor groove sugar O4' oxygens are also associated to well-ordered water molecules.

Metal ions can interact with nucleic acids in two different binding modes modulating their structure and function: diffuse binding and site binding (Hud 2009). In

the diffuse binding, the interaction between nucleic acids and metal ions is through the conserved first hydration layer. For example, the delocalized counterion atmosphere interacts with negatively charged phosphate groups (Subirana & Soler-López 2003). However, in the site-binding mode the metal interacts either directly with the nucleic acid (upon partial dehydration) or, more frequently, through a water molecule showing important sequence-dependent features.

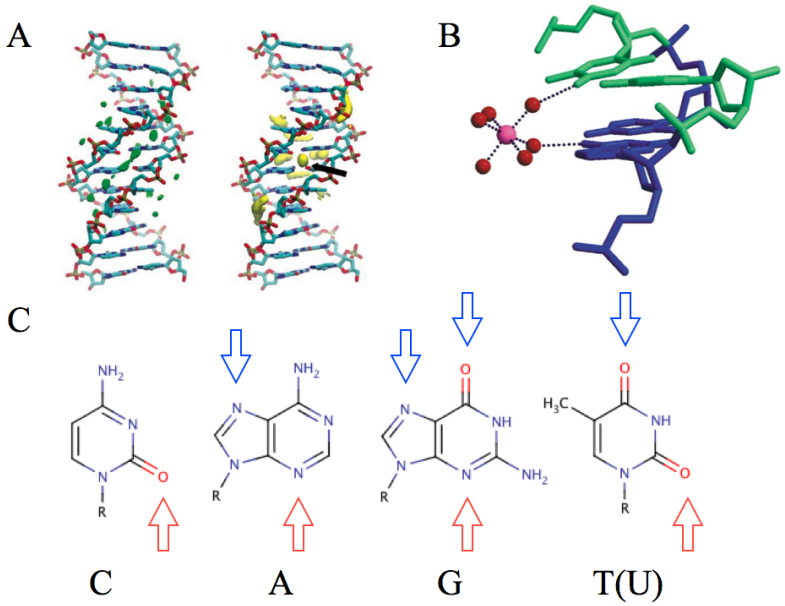


Figure 3.15: Localization of water and Na^+ molecules around the DNA Dickerson dodecamer over the course of a 1 ms MD simulation (Pérez et al. 2007) (A). Example of Mg^{2+}-induced local bending at a GpA step (NDB code: bd0037 in Guéroult et al. 2012) (B). Potential cation binding sites for the 4(5) neutral tautomeric forms both in the major groove side (blue) and in the minor groove (red) (C).

Without considering the polymorphic nature of a double stranded nucleic acid, the preferred cation binding sites on nucleobases at physiological conditions are: in the minor groove, adenine N3 and thymine O2 of A·T base pairs and, in the major groove guanine N7 and O6 and cytosine N3 of C·G base pairs and adenine N7 and thymine O4 of A·T base pairs. It has been shown that the cation interaction with the major groove is more favorable at G·C base pairs due to the N7 and O6 of guanine presenting a better binding site that the adenine N7, which has the close N6 amino group. The G·C amino group in the minor groove presents a positive electrostatic potential that interferes in cation interactions. Therefore, in general, cation interactions may take place not only with phosphate groups but also with the minor groove in A·T base pairs and with the major groove in C·G base pairs. However, local flexibility and the nature

of the cation could facilitate the interaction with C·G base pairs in the minor groove. Interestingly, a single metal ion can interact with two different cation binding sites or may be chelated by a nucleobase. X-ray structural databases have several examples on how binding sites can be combined. For example, Mg^{2+} ion, which has a coordination sphere of six water molecules, can interact specifically with the major groove side of GpN steps (N can be any nucleotide) -or even further- generating a local bending in the structure (Guéroult et al. 2012).

For RNA molecules, ions play an essential role in folding and catalysis (Hud 2009; Siegel et al. 2011), and some bound cations should be considered as a intrinsic part of the RNA structure.

Triple stranded nucleic acid structures have been shown to interact with cations at TFO purine N7 atoms, a coordination that could screen phosphate repulsions. Theoretical calculations have suggested that antiparallel triplexes may have stronger G#G reverse-Hoogsteen interactions due to a polarization mechanism induced by the divalent cation binding in the minor-major groove (Sponer et al. 1998).

3.7 Therapeutic applications of nucleoside analogues

Development of synthetic oligonucleotides (ONs) has received much attention in the past years due to their ability to silence the expression of undesired overexpressed genes. Among the current strategies, siRNAs have shown to be highly sequence-specific. However, their biomedical application is limited due to several *in vivo* drawbacks. Synthetic ONs must cross several biological membranes, resist the action of cellular nucleases, and finally, result stable enough when binding to the target sequence. One of the current strategies to increase stability involves the modification of olignucleotides. These modifications can be placed both in the backbone or nucleobase moieties (Robles et al. 2002; Blackburn et al. 2006), and try to improve specificity and affinity while creating nuclease-resistant oligos.

Modifications in the backbone are probably the most popular (Sproat 1995; Leumann 2002). **Backbone modifications** can be classified into: sugar modifications, phosphate modifications, and full backbone replacement. Among sugar modifications, locked nucleic acids (LNA) (Obika et al. 1997; Singh et al. 1998) and the O2'-modifications (Frier & Altmann 1997) are the most popular. LNAs has a methylene bridge between the O2' and the C4' which restricts the sugar conformation to C3'-endo resulting in a stronger binding of LNA to complementary DNA or RNA strands (Figure 3.15). On the other hand, O2'-modifications like 2'-O-methyloligoribonucleotides or

2'-deoxy-2'-fluoro-β-D- arabinonucleoside oligomers (2'-F-ANA) (Damha et al. 1998), have been shown to increase the stability of the oligonucleotide towards exonuclease degradation.

Different strategies for the modification of the phosphate moiety have increased the resistance to nucleases. A very popular approach was first introduced by Fritz Eckstein by replacing a non-bridging oxygen atom of the phosphodiester group by sulfur (Eckstein 1985). The phosphorothioate modification provides stable RNA·DNA duplexes while maintaining similar charge distributions. However, it leads to unstable triplexes avoiding their application in antigene strategies (Lacoste et al. 1997; Latimer et al. 1989). Further sulfur substitution of the second non-bridging oxygen atom of the phosphorothioate group leads to phosphorodithioates linkages which have been also proved to be resistant to cleavage by nucleases. However, their usage is reduced since they have a decrease sequence-specificity and bind to some proteins (Blackburn et al. 2006). Neutral phosphate linkages like methylphosphonates have shown moderate success although they present enhanced stability to nucleases and increase the corresponding melting temperature values when incorporated to duplexes (Robles et al. 2002).

The complete replacement of backbone ONs has been widely studied. Morpholino linkages (Summerton & Weller 1997) and, specially, peptide nucleic acids (PNA) (Nielsen et al. 1991; Egholm et al. 1993) have shown excellent nuclease and protease resistance. PNAs result from substitution of the phosphoribose backbone by a peptide chain. Molecular dynamics simulations of PNA·DNA hybrid molecules (Shields et al. 1998; Soliva et al. 2000) have shown that these molecules present an A-type conformation with flexible backbone features.

It is worth to mention that single terminal modifications at 3' residues of ONs also enhance exonuclease resistance. During this PhD, a nucleobase modification, N-ethyl-N-coupled nucleosides, were probed to avoid 3'-exonucleases hydrolysis (Figure 3.16). Moreover, *in vitro* experiments showed that modified siRNAs with these modifications decrease mRNA levels of the corresponding target gene, while keeping RISC susceptibility.

Modified nucleosides have been used as antimetabolites, compounds that prevent the DNA synthesis, and therefore, used as anti-cancer drugs (Christopherson et al. 2002; Galmarini et al. 2002). Their inhibition mechanisms include competition for enzymatic binding sites and their incorporation into forming nucleic acids. Several analogues have been successfully tested and, in general, they can be classified in antifolates, and purine and pyrimidine analogues.

Among purine-like antimetabolites, thiopurines have found a useful application in

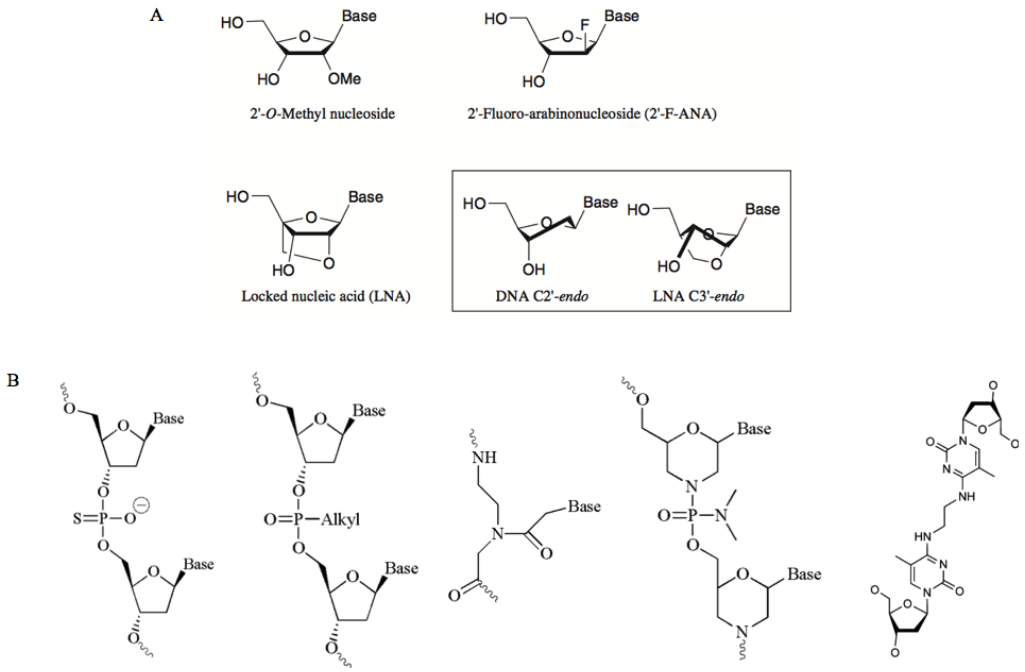


Figure 3.16: Sugar modifications: On top, 2'-O-methyl, 2'-F-ANA and LNA nucleosides. Bottom, from left to right, phosphorothioate, alkylphosphonate, PNA, morpholino derivative, and N-ethyl-N-coupled cytosines.

cancer chemotherapy against childhood acute lymphoblastic leukemia (Erb et al. 1998). Both 6-mercaptopurine and 6-thioguanine (figure 3.17) have been proved to mimic hypoxanthine and guanine respectively and can be incorporated into *de novo* nucleic acid strands (Bohon & de los Santos 2005). Different studies suggest that the thiopurines may work as a result of a combined effect, first by feedback inhibiting the synthesis of 5-phosphoribosylamine from PRPP (α-D-5-phosphoribosyl 1-pyrophosphate), and second, by avoiding the conversion of IMP (inosine monophosphate) into XMP (xanthine monophosphate) and adenylosuccinate. The combined action of these mechanisms eventually results in the incorporation of the corresponding triphosphate compounds in DNA and RNA. Moreover, the incorporation of these guanine analogs in single stranded guanine-rich sequences prevent the formation of G-quadruplex structures and favor the formation of triplex structures (Marathias et al. 1999).

During this PhD other possible antimetabolites have been considered, among them, thioketothymines resembles natural thymine but these sulfur-derivatives have been successfully applied to treat photosensitive cancer cells (Faustino et al. 2009). We also consider 6-selenoguanine, a purine derivative that has been incorporated into different nucleic acid structures showing that this molecule, although has a destabilizing effect in the DNA environment (Salon et al. 2008), it can be well fitted in the DNA duplex

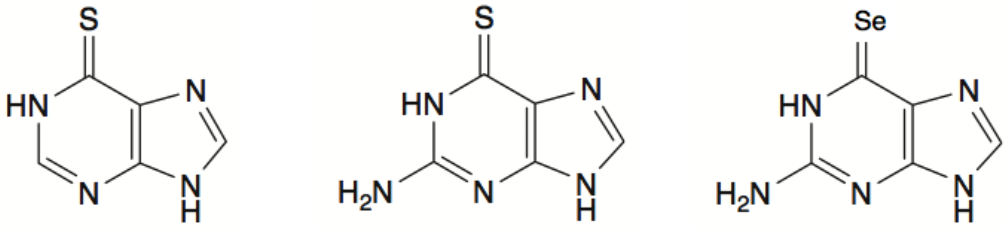


Figure 3.17: Nucleobase modifications. From left to right, 6-thiopurine, 6-mercaptopurine, and 6-selenoguanine.

structure. Besides this, antiparallel triple stranded DNA structures have been studied with and without 6-selenoguanine and interestingly, the modification in the TFO strand (rH) has less detrimental effect than the WC modification. Finally, a G-quadruplex structure (thrombin binding aptamer, TBA) has been studied when incorporating one single seleno-modification. The thermodynamic analysis suggest that the formation of an antiparallel triplex with 6-selenoguanine-rich TFO will be favored instead of the G-quadruplex formation.

3.8 References

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

Altona, C. & Sundaralingam, M., 1973. Conformational analysis of the sugar ring in nucleosides and nucleotides. Improved method for the interpretation of proton magnetic resonance coupling constants. Journal of the American Chemical Society, 95(7), pp.2333–2344.

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

Behe, M. & Felsenfeld, G., 1981. Effects of methylation on a synthetic polynucleotide: the B→Z transition in poly(dG-m5dC).poly(dG-m5dC). Proceedings of the National Academy of Sciences of the United States of America, 78(3), pp.1619–1623.

Betts, L. et al., 1995. A nucleic acid triple helix formed by a peptide nucleic acid-DNA complex. Science, 270(5243), pp.1838–1841.

Blackburn, G.M. et al., 2006. Nucleic acids in chemistry and biology Royal Society of Chemistry, Royal Society of Chemistry.

Biffi, G. et al., 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. Nature Chemistry.

Bohon, J. & de los Santos, C.R., 2005. Effect of 6-thioguanine on the stability of duplex DNA. Nucleic Acids Research, 33(9), pp.2880–2886.

Buroker, N. et al., 1987. A hypervariable repeated sequence on human chromosome 1p36. Human genetics, 77(2), pp.175–181.

Calladine, C.R., 1982. Mechanics of sequence-dependent stacking of bases in B-DNA. Journal of Molecular Biology, 161(2), pp.343–352.

Chargaff, E., 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia, 6(6), pp.201–209.

Cheong, C. & Moore, P.B., 1992. Solution structure of an unusually stable RNA tetraplex containing G- and U-quartet structures. Biochemistry, 31(36), pp.8406–8414.

Chin, J.Y., Schleifman, E.B. & Glazer, P.M., 2007. Repair and recombination induced by triple helix DNA. Front Biosci, 12, pp.4288–4297.

Christopherson, R.I., Lyons, S.D. & Wilson, P.K., 2002. Inhibitors of de novo nucleotide biosynthesis as drugs. Accounts of Chemical Research, 35(11), pp.961–971.

Colominas, C., Luque, F.J. & Orozco, M., 1996. Tautomerism and protonation of guanine and cytosine. Implications in the formation of hydrogen-bonded complexes. Journal of the American Chemical Society, 118(29), pp.6811–6821.

Damha, M.J. et al., 1998. Hybrids of RNA and Arabinonucleic Acids (ANA and 2'F-ANA) Are Substrates of Ribonuclease H J. Am. Chem. Soc. 1998, 120, 12976-12977.

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

Davis, G. & Kayser, K.J., 2010. Chromosomal Mutagenesis, Humana Press.

Dempsey, L.A., 1999. G4 DNA Binding by LR1 and Its Subunits, Nucleolin and hnRNP D, A Role for G-G pairing in Immunoglobulin Switch Recombination. Journal of Biological Chemistry, 274(2), pp.1066–1071.

Djuranovic, D. & Hartmann, B., 2004. DNA fine structure and dynamics in crystals

and in solution: the impact of BI/BII backbone conformations. Biopolymers, 73(3), pp.356–368.

Doluca, O., Withers, J.M. & Filichev, V.V., 2013. Molecular Engineering of Guanine-Rich Sequences: Z-DNA, DNA Triplexes, and G-Quadruplexes. Chemical Reviews, 113(5), pp.3044–3083.

Drew, H. et al., 1980. High-salt d(CpGpCpG), a left-handed Z' DNA double helix. Nature, 286(5773), pp.567–573.

Dršata, T. et al., 2012. Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. Journal of Chemical Theory and Computation, 9, pp.707–721.

Eckstein, F., 1985. Nucleoside Phosphorothioates. Annual review of biochemistry, 54(1), pp.367–402.

Eddy, J. & Maizels, N., 2006. Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Research, 34(14), pp.3887–3896.

Egholm, M. et al., 1993. PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. Nature, 365(6446), pp.566–568.

Egli, M., Portmann, S. & Usman, N., 1996. RNA hydration: a detailed look. Biochemistry, 35(26), pp.8489–8494.

Erb, N., Harms, D.X.R.O. & Janka-Schaub, G., 1998. Pharmacokinetics and metabolism of thiopurines in children with acute lymphoblastic leukemia receiving 6-thioguanine versus 6-mercaptopurine. Cancer Chemotherapy and Pharmacology, 42(4), pp.266–272.

Eritja, R. et al., 1999. Modified Oligonucleotides with Triple-Helix Stabilization Properties. Nucleosides and Nucleotides, 18(6-7), pp.1619–1621.

Faustino, I. et al., 2009. Unique tautomeric and recognition properties of thioketothymines? Journal of the American Chemical Society, 131(35), pp.12845–12853.

Faustino, I., Pérez, A. & Orozco, M., 2010. Toward a consensus view of duplex RNA flexibility. Biophysical Journal, 99(6), pp.1876–1885.

Foloppe, N. & MacKerell, A.D., 1999. Contribution of the phosphodiester backbone and glycosyl linkage intrinsic torsional energetics to DNA structure and dynamics. The Journal of Physical Chemistry B, 103(49), pp.10955–10964.

Freier, S.M. & Altmann, K.H., 1997. The ups and downs of nucleic acid duplex stability: structure-stability studies on chemically-modified DNA:RNA duplexes. Nucleic Acids Research, 25(22), pp.4429–4443.

Galmarini, C.M., Mackey, J.R. & Dumontet, C., 2002. Nucleoside analogues and nucleobases in cancer treatment. The lancet oncology, 3(7), pp.415–424.

Garcia, R.G. et al., 1999. Theoretical calculations, synthesis and base pairing properties of oligonucleotides containing 8-amino-2'-deoxyadenosine. Nucleic Acids Research, 27(9), pp.1991–1998.

Gellert, M., Lipsett, M.N. & Davies, D.R., 1962. Helix formation by guanylic acid. Proceedings of the National Academy of Sciences of the United States of America, 48, pp.2013–2018.

Gorenstein, D.G., 1994. Conformation and Dynamics of DNA and Protein-DNA Complexes by 31P NMR. Chemical Reviews, 94(5), pp.1315–1338.

Guéroult, M. et al., 2012. Mg(2+) in the Major Groove Modulates B-DNA Structure and Dynamics. PLoS ONE, 7(7), p.e41704.

Granotier, C. et al., 2005. Preferential binding of a G-quadruplex ligand to human chromosome ends. Nucleic Acids Research, 33(13), pp.4182–4190.

Ha, S.C. et al., 2005. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. Nature, 437(7062), pp.1183–1186.

Hanakahi, L.A., Sun, H. & Maizels, N., 1999. High affinity interactions of nucleolin with G-G-paired rDNA. Journal of Biological Chemistry, 274(22), pp.15908–15912.

Haniford, D.B. & Pulleyblank, D.E., 1983. Facile transition of poly[d(TG) x d(CA)] into a left-handed helix in physiological conditions. Nature, 302(5909), pp.632–634.

Hartmann, B., Sullivan, M.R. & Harris, L.F., 2003. Operator recognition by the phage 434 cI repressor: MD simulations of free and bound 50-bp DNA reveal important differences between the OR1 and OR2 sites. Biopolymers, 68(2), pp.250–264.

Hocquet, A., Leulliot, N. & Ghomi, M., 2000. Ground-State Properties of Nucleic Acid Constituents Studied by Density Functional Calculations. 3. Role of Sugar Puckering and Base Orientation on the Energetics and Geometry of 2'-Deoxyribonucleosides and Ribonucleosides. The Journal of Physical Chemistry B, 104(18), pp.4560–4568.

Htun, H. & Dahlberg, J.E., 1989. Topology and formation of triple-stranded H-DNA. Science, 243(4898), pp.1571–1576.

Hud, N.V., 2009. Nucleic Acid-Metal Ion Interactions The Royal Society of Chemistry, The Royal Society of Chemistry.

Hunter, C.A., 1993. Sequence-dependent DNA Structure. Journal of Molecular Biology, 230(3), pp.1025–1054.

Kang, C. et al., 1992. Crystal structure of four-stranded *Oxytricha* telomeric DNA. Nature, 356(6365), pp.126–131.

Lacoste, J., François, J.C. & Hélène, C., 1997. Triple helix formation with purine-rich phosphorothioate-containing oligonucleotides covalently linked to an acridine derivative. Nucleic Acids Research, 25(10), pp.1991–1998.

Latimer, L.J., Hampel, K. & Lee, J.S., 1989. Synthetic repeating sequence DNAs containing phosphorothioates: nuclease sensitivity and triplex formation. Nucleic Acids Research, 17(4), pp.1549–1561.

Laughlan, G. et al., 1994. The high-resolution crystal structure of a parallel-stranded guanine tetraplex. Science, 265(5171), pp.520–524.

Leontis, N.B. & Westhof, E., 2001. Geometric nomenclature and classification of RNA base pairs. RNA, 7(4), pp.499–512.

Leumann, C., 2002. DNA Analogues: From Supramolecular Principles to Biological Properties. Bioorganic & Medicinal Chemistry, 10(4), pp.841–854.

Limongelli, V. et al., 2013. The G-triplex DNA. Angewandte Chemie International Edition, 52(8), pp.2269–2273.

Liu, H. et al., 2006. Cooperative activity of BRG1 and Z-DNA formation in chromatin remodeling. Molecular and cellular biology, 26(7), pp.2550–2559.

Maehigashi, T. et al., 2012. B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. Nucleic Acids Research, 40(8), pp.3714-3722.

Marathias, V.M., Sawicki, M.J. & Bolton, P.H., 1999. 6-Thioguanine alters the structure and stability of duplex DNA and inhibits quadruplex DNA formation. Nucleic Acids Research, 27(14), pp.2860–2867.

McNeer, N.A. et al., 2011. Polymer delivery systems for site-specific genome editing. Journal of controlled release : official journal of the Controlled Release Society, 155(2), pp.312–316.

Mirkin, S.M., 2006. DNA structures, repeat expansions and human hereditary disorders. Current Opinion in Structural Biology, 16(3), pp.351–358.

Neidle, S., 2007. Principles of nucleic acid structure Academic Press, Academic Press.

Neidle, S. & Balasubramanian, S., 2006. Quadruplex nucleic acids Royal Society of Chemistry, Royal Society of Chemistry.

Nielsen, P.E. et al., 1991. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. Science, 254(5037), pp.1497–1500.

Nikolova, E.N., Gottardo, F.L. & Al-Hashimi, H.M., 2012. Probing transient Hoogsteen hydrogen bonds in canonical duplex DNA using NMR relaxation dispersion and single-atom substitution. Journal of the American Chemical Society, 134(8), pp.3667–3670.

Obika, S. et al., 1997. Synthesis of 2'-O,4'-C-methyleneuridine and -cytidine. Novel bicyclic nucleosides having a fixed C3'-endo sugar puckering. Tetrahedron Letters, 38(50), pp.8735–8738.

Oh, D.-B., Kim, Y.-G. & Rich, A., 2002. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. Proceedings of the National Academy of Sciences of the United States of America, 99(26), pp.16666–16671.

Olson, W.K. & Sussman, J.L., 1982. How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. Journal of the American Chemical Society, 104(1), pp.270–278.

Olson, W.K. et al., 2001. A standard reference frame for the description of nucleic acid base-pair geometry. Journal of Molecular Biology, 313(1), pp.229–237.

Otero, R. et al., 2005. Guanine Quartet Networks Stabilized by Cooperative Hydrogen Bonds. Angewandte Chemie International Edition, 44(15), pp.2270–2275.

Padmanabhan, K. et al., 1993. The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. Journal of Biological Chemistry, 268(24), pp.17651–17654.

Parkinson, G.N., Lee, M.P.H. & Neidle, S., 2002. Crystal structure of parallel quadruplexes from human telomeric DNA. Nature, 417(6891), pp.876–880.

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

Phan, A.T., Modi, Y.S. & Patel, D.J., 2004. Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter. JOURNAL-AMERICAN CHEMICAL SOCIETY, 126(28), pp.8710–8716.

Phillips, K. et al., 1997. The crystal structure of a parallel-stranded guanine tetraplex at 0.95Å resolution. Journal of Molecular Biology, 273(1), pp.171–182.

Pichler, A. et al., 1999. Unexpected BII Conformer Substate Population in Unoriented Hydrated Films of the d(CGCGAATTCGCG)$_2$ Dodecamer and of Native B-DNA from Salmon Testes. Biophysical Journal, 77(1), pp.398–409.

Qin, Y. & Hurley, L.H., 2008. Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. Biochimie, 90(8), pp.1149–1171.

Radhakrishnan, I. & Patel, D.J., 1994. DNA triplexes: solution structures, hydration sites, energetics, interactions, and function. Biochemistry, 33(38), pp.11405–11416.

Rao, S.T. & Sundaralingam, M., 1969. Stereochemistry of nucleic acids and their constituents. V. The crystal and molecular structure of a hydrated monosodium inosine 5'-phosphate. A commonly occurring unusual nucleotide in the anticodons of tRNA. Journal of the American Chemical Society, 91(5), pp.1210–1217.

Reichardt, C. & Welton, T., 2011. Solvents and Solvent Effects in Organic Chemistry, Wiley-VCH.

Rich, A., 1995. The Nucleic Acids A Backward Glance. Annals of the New York Academy of Sciences, 758(1 DNA), pp.97–142.

Risitano, A. & Fox, K.R., 2004. Influence of loop size on the stability of intramolecular DNA quadruplexes. Nucleic Acids Research, 32(8), pp.2598–2606.

Roberts, R. & Crothers, D., 1992. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. Science, 258(5087), pp.1463–1466.

Robles, J. et al., 2002. Nucleic acid triple helices: stability effects of nucleobase modifications. Current Organic Chemistry, 6(14), pp.1333–1368.

Rothenburg, S. et al., 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proceedings of the National Academy of Sciences of the United States of America, 98(16), pp.8985–8990.

Salon, J. et al., 2008. Derivatization of DNAs with selenium at 6-position of guanine for function and crystal structure studies. Nucleic Acids Research, 36(22), pp.7009–7018.

Shields, G.C., Laughton, C.A. & Orozco, M., 1997. Molecular Dynamics Simulations of the d (T·A·T) Triple Helix. Journal of the American Chemical Society, 119(32), pp.7463–7469.

Shields, G.C., Laughton, C.A. & Orozco, M., 1998. Molecular Dynamics Simulation of a PNA·DNA·PNA Triple Helix in Aqueous Solution. Journal of the American Chemical Society, 120(24), pp.5895–5904.

Siegel, A., Sigel, H. & Sigel, R.K.O., 2011. Structural and Catalytic Roles of Metal Ions in RNA.

Singh, S.K. et al., 1998. LNA (locked nucleic acids): synthesis and high-affinity nucleic acid recognition. Chemical Communications, (4), pp.455–456.

Singleton, C.K. et al., 1982. Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. Nature, 299(5881), pp.312–316.

Smirnov, I. & Shafer, R.H., 2000. Effect of Loop Sequence and Size on DNA Aptamer Stability. Biochemistry, 39(6), pp.1462–1468.

Soliva, R. et al., 1999. Role of Sugar Re-Puckering in the Transition of A and B Forms of DNA in Solution. A Molecular Dynamics Study. Journal of Biomolecular Structure & Dynamics, 17(1), pp.89–99.

Soliva, R. et al., 2000. Molecular Dynamics Simulations of PNA·DNA and PNA·RNA Duplexes in Aqueous Solution. Journal of the American Chemical Society, 122(25), pp.5997–6008.

Sponer, J. et al., 1998. Stabilization of the purine.purine.pyrimidine DNA base triplets by divalent metal cations. Journal of Biomolecular Structure & Dynamics, 16(1), pp.139–143.

Sponer, J., Leszczynski, J. & Hobza, P., 1996. Nature of Nucleic Acid-Base Stacking: Nonempirical ab Initio and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs. The Journal of Physical Chemistry, 100(13), pp.5590–5596.

Sproat, B.S., 1995. Chemistry and applications of oligonucleotide analogues. Journal of Biotechnology, 41(2-3), pp.221–238.

Srinivasan, A.R. & Olson, W.K., 1987. Nucleic Acid Model Building: The Multiple Backbone Solutions Associated with a Given Base Morphology. Journal of Biomolecular Structure & Dynamics, 4(6), pp.895–938.

Strobel, S.A. & Dervan, P.B., 1990. Site-specific cleavage of a yeast chromosome by oligonucleotide-directed triple-helix formation. Science, 249(4964), pp.73–75.

Subirana, J.A. & Faria, T., 1997. Influence of sequence on the conformation of the

B-DNA helix. Biophysical Journal, 73(1), pp.333–338.

Subirana, J.A. & Soler-López, M., 2003. Cations as hydrogen bond donors: a view of electrostatic interactions in DNA. Annual review of biophysics and biomolecular structure, 32, pp.27–45.

Summerton, J. & Weller, D., 1997. Morpholino antisense oligomers: design, preparation, and properties. Antisense and Nucleic Acid Drug Development, 7(3), pp.187–195.

Suzuki, M. et al., 1997. Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. Journal of Molecular Biology, 274(3), pp.421–435.

Temiz, N.A. et al., 2012. The Role of Methylation in the Intrinsic Dynamics of B- and Z-DNA. PLoS ONE, 7(4), p.e35558.

Tereshko, V., Minasov, G. & Egli, M., 1999. A "hydrat-ion" spine in a B-DNA minor groove. Journal of the American Chemical Society, 121(15), pp.3590–3595.

Thuong, N.T. & Hélène, C., 1993. Sequence-Specific Recognition and Modification of Double-Helical DNA by Oligonucleotides. Angewandte Chemie (International ed. in English), 32(5), pp.666–690.

Travers, A.A., 1989. DNA Conformation and Protein Binding. Annual review of biochemistry, 58(1), pp.427–452.

Vasquez, K.M., 2010. Targeting and processing of site-specific DNA interstrand crosslinks. Environmental and molecular mutagenesis, 51(6), pp.527–539.

Wang, A.H. et al., 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature, 282(5740), pp.680–686.

Wang, G., Christensen, L.A. & Vasquez, K.M., 2006. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. Proceedings of the National Academy of Sciences of the United States of America, 103(8), pp.2677–2682.

Wang, Y. & Patel, D.J., 1993. Solution structure of the human telomeric repeat d[$AG_3(T_2AG_3)_3$] G-tetraplex. Structure, 1(4), pp.263–282.

Watson, J.D. & Crick, F.H., 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171(4356), pp.737–738.

Wing, R. et al., 1980. Crystal structure analysis of a complete turn of B-DNA. Nature, 287(5784), pp.755–758.

Wong, Z. et al., 1987. Characterization of a panel of highly variable minisatellites cloned from human DNA. Annals of human genetics, 51(Pt 4), pp.269–288.

Zhanpeisov, N.U. & Leszczynski, J., 1998. The specific solvation effects on the structures and properties of adenine-uracil complexes: A theoretical ab initio study. The Journal of Physical Chemistry A, 102(30), pp.6167–6172.

Zhanpeisov, N.U., Sponer, J. & Leszczynski, J., 1998. Reverse Watson-Crick Isocytosine-Cytosine and Guanine-Cytosine Base Pairs Stabilized by the Formation of the Minor Tautomers of Bases. An ab Initio Study in the Gas Phase and in a Water Cluster. The Journal of Physical Chemistry A, 102(50), pp.10374–10379.

4

Results, discussion and conclusions

4.1 Study of modified DNA and RNA nucleobases and their thermodynamical stability

4.1.1 Unique tautomeric and recognition properties of thioketothymines?

Ignacio Faustino, Anna Aviñó, Iván Marchán, F. Javier Luque, Ramón Eritja and Modesto Orozco.

Journal of the American Chemical Society. 2009, 131(35), pp.12845–12853.

J|A|C|S ARTICLES

Published on Web 08/14/2009

Unique Tautomeric and Recognition Properties of Thioketothymines?

Ignacio Faustino,[†] Anna Aviño,[‡] Ivan Marchán,[†] F. Javier Luque,[§] Ramon Eritja,[‡] and Modesto Orozco*[,†,||,⊥]

Joint IRB-BSC Program on Computational Biology, Institute of Research in Biomedicine, Parc Científic de Barcelona, Josep Samitier 1−5, Barcelona 08028, Spain and Barcelona Supercomputing Centre, Jordi Girona 31, Edifici Torre Girona, Barcelona 08034, Spain, Institute of Research in Biomedicine, IQAC-CSIC, CIBER-BBN Networking Centre on Bioengineering, Biomaterials and Nanomedicine, Baldiri Reixac 15, Barcelona 08028, Spain, Departament de Fisicoquímica and Institut de Biomedicina (IBUB), Facultat de Farmàcia, Universitat de Barcelona, Avgda Diagonal 643, Barcelona 08028, Spain, National Institute of Bioinformatics, Parc Científic de Barcelona, Josep Samitier 1−5, Barcelona 08028, Spain, and Departament de Bioquímica, Facultat de Biología, Universitat de Barcelona, Avgda Diagonal 647, Barcelona 08028, Spain

Received June 15, 2009; E-mail: modesto@mmb.pcb.ub.es

Abstract: The tautomeric and recognition properties of thymine, 2- and 4-thioketothymines have been studied by means of accurate ab initio methods combined with molecular dynamics simulations and free energy calculations. In contrast to previous suggestions in the literature, the replacement of carbonyl oxygens by sulfur atoms does not lead to dramatic changes in tautomeric properties of the pyrimidine derivatives neither in vacuum nor in aqueous solution. Moreover, the presence of thioketothymines induces only mild changes in DNA structure, stability and fidelity. Despite the fact that mismatching can largely stabilize minor tautomeric forms, thioketothymines are found in the canonical thioketo-form irrespective of the paired base. Our theoretical results, confirmed by new experimental studies, describe the complete tautomeric and recognition characteristics of thioketothymines and demonstrate that both 2-thioketo and 4-thioketothymine are excellent molecules to introduce special chemical properties in modified DNA.

Introduction

The genetic code is a minimalist language made with only four letters (the coding nucleobases A, G, C, and T), which appear in the DNA as complementary pairs (A·T and G·C) involving the major keto/amino tautomeric forms.[1−5] Their structure reveals how exquisite was evolution in defining bases that confer stability, specificity, and flexibility to the DNA. However, it is also clear that other chemical entities are compatible with the DNA structure.[6,7] Some of these nonstandard bases are in fact spontaneously found in DNA as a result of chemical/physical stress (i.e., radiation or oxidation) leading to DNA damage.[8] Additionally, synthetic bases can be introduced in DNA either chemically or enzymatically to induce conformational changes in nucleic acids, to modulate their intrinsic stability, to alter processing or reading of DNA, the functioning of reparatory machinery, to induce mutations or cross-linking, or to modify the recognition properties of the strands.[9] As a result, nonstandard nucleobases can change the functionality of DNA, opening a wide range of biotechnological and biomedical applications.[10]

Nonstandard nucleobases are typically designed assuming the preponderance of keto/amino tautomers in DNA. The validity of this assumption is not always guaranteed, as the tautomeric preferences of nucleobases depends not only on the intrinsic (gas-phase) stability between tautomers but also on the differential stabilization exerted by the DNA environment.[11,13,19] Thus, previous studies have shown that *N*-methyl-derivatives of cytosines[12] or isoguanine[13] have a significant population of

[†] Joint IRB-BSC Program on Computational Biology.
[‡] Institute of Research in Biomedicine.
[§] Facultat de Farmàcia, Universitat de Barcelona.
[||] National Institute of Bioinformatics.
[⊥] Facultat de Biología, Universitat de Barcelona.

(1) Watson, J. D.; Crick, F. H. *Nature* **1953**, *171*, 737–738.
(2) Colominas, C.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1996**, *118*, 6811–6821.
(3) Topal, M. D.; Fresco, J. R. *Nature* **1976**, *263*, 289–293.
(4) Sanger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.
(5) Neidle, S. *Nucleic Acid Structure and Recognition*; Oxford University Press, New York, 2002.
(6) Swann, P. F.; Waters, T. R.; Moulton, D. C.; Xu, Y. Z.; Zheng, Q.; Edwards, M.; Mace, R. *Science* **1996**, *273*, 1109–1111.
(7) Maki, H.; Kornberg, A. *J. Biol. Chem.* **1985**, *260*, 12987–12992.
(8) Kamiya, H. *Nucleic Acids Res.* **2003**, *31*, 517–531.
(9) Morales, J. C.; Kool, E. T. *Nat. Struct. Biol.* **1998**, *5*, 950–954.
(10) Herdewijn, P. *Antisense Nucleic Acid Drug Dev.* **2000**, *10*, 297–310.
(11) Hernández, B.; Soliva, R.; Luque, F. J.; Orozco, M. *Nucleic Acids Res.* **2000**, *28*, 4873–4883.
(12) (a) Anand, N. N.; Brown, D. M.; Salisbury, S. A. *Nucleic Acids Res.* **1987**, *15*, 8167–8176. (b) Fazakerley, G. V.; Gdaniec, Z.; Sowers, L. C. *J. Mol. Biol.* **1993**, *230*, 6–10. (c) Schuerman, G. S.; Van Meervelt, L.; Loakes, D.; Brown, D. M.; Kong Thoo Lin, P.; Moore, M. H.; Salisbury, S. A. *J. Mol. Biol.* **1998**, *282*, 1005–1011.
(13) (a) Robinson, H.; Gao, Y. G.; Bauer, C.; Roberts, C.; Switzer, C.; Wang, A. H. *Biochemistry* **1998**, *37*, 10897–10905. (b) Blas, J. R.; Luque, F.; Orozco, M. *J. Am. Chem. Soc.* **2004**, *126*, 154–164.

 © 2009 American Chemical Society

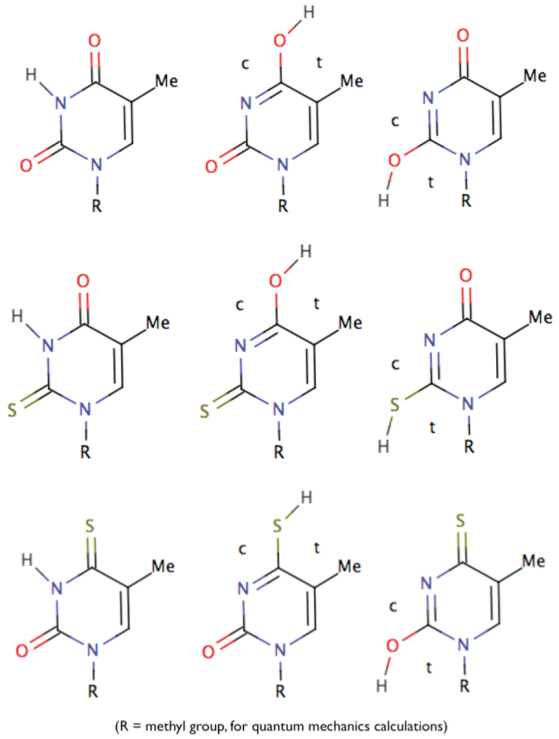


Figure 1. Schematic representation of thymine (T, upper), 2-thioketothymine (^{2}S), 4-thioketothymine (^{4}S), and their tautomeric species.

imino or enol tautomers in the DNA duplex, and more modified nucleobases are expected to display recognition modes modulated by tautomeric equilibria.

Among the thymine (T) derivatives designed under the assumption that keto tautomers are prevalent,[14] 2- and 4-thioketothymines (^{2}S and ^{4}S; Figure 1) are relevant as the reactivity of the thioketocarbonyl group can be exploited in nucleophilic and photocross-linking reactions in DNA.[15] Moreover, the possibility to photoactivate thioketothymines in conditions where coding nucleobases are unaltered confer potential interest as prodrugs in the phototherapy of psoriasis and skin cancer.[16] Early hybridization experiments suggested that insertion of thioketothymidine in front of either G or A showed a selectivity pattern similar to that of T, while leading to a slight destabilization of the duplex.[17] However, this behavior has been challenged by more recent analysis, which suggests that one of the thioketo-derivatives, ^{4}S, shows mutagenic properties due to stabilization of G-mismatching.[18] These findings raise doubts about the prevalence of keto/thioketo tautomers in thioketopyrimidines, suggesting a partner-dependent tautomeric preference (see below), as noted previously by other modified nucleobases.[13,19] This behavior, if confirmed, will open interesting possibilities for the design of promiscuous nucleobases, though it will seriously challenge the real impact of thioketothymines in the chemistry and phototherapy of nucleic acids.

A detailed comparison of the tautomeric preferences of thymine and thioketothymines in different environments, including DNA, is presented. Attention is paid to the impact of thioketothymines in the structure, recognition properties, and stability of DNA upon insertion in both canonical and mismatched positions. Our results strongly suggest that, contrary to previous suggestions, the standard Watson−Crick tautomeric rules are applicable to thioketothymines. Thus, the presence of thymine surrogates does not significantly affect the DNA structure and only introduces marginal destabilization in the duplex, without altering the DNA pairing fidelity and therefore lacking mismatch-related mutagenic properties. Overall, our results demonstrate that thioketothymines can be safely used as surrogates of T with improved reactive properties and pharmacological activities.

Methods

Gas-Phase Calculations. High level ab initio theory was used to investigate the intrinsic (gas-phase) tautomeric preferences of T, ^{2}S, and ^{4}S. For this purpose all possible tautomers of the *N*1-methyl derivatives were generated (Figure 1) and fully optimized at the MP2/6-311G(d,p) level. Single-point calculations at the MP2/cc-pVDZ, MP2/cc-pVTZ, and MP2/cc-pVQZ levels were carried out to estimate the energy differences between tautomers using Truhlar's extrapolation scheme[20] to complete basis set (almost identical results were obtained using Erlangen's extrapolation scheme[21]). Higher-order electron correlation effects were accounted for by adding the difference between MP2 and CCSD(T) energies using the 6-31G(d) basis set. Zero-point, thermal, and entropic terms needed to compute tautomerization free energies were added by using the harmonic oscillator model implemented in Gaussian03[22] at the MP2/6-311G(d,p) level. Combining all of these corrections, we obtained our *best estimate* that are expected to be within a few tenths of the real value in kcal/mol.

To check the goodness of classical force fields to represent canonical and noncanonical pairings of thymine and thioketothymines with A and G, the corresponding dimers (Figure 2) were optimized at the B3LYP/6-31G(d) level. The final structures were subjected to single-point force field calculations, as well as to additional ab initio computations at the HF/6-311+G(d,p) and MP2/6-31G(d) levels, which were combined using standard protocols to derive MP2/6-311+G(d,p) estimates of the interaction energy (basis set superposition error was corrected using the counterpoise method.[23] Comparison of the values obtained here with state of the art (CCSD(T)/CBS-quality) results reported by Sponer and Hobza[24] suggests that our best (MP2/6-311+G(d,p)-quality) estimates correlate well with the "gold-standard" reference values, except for a constant scaling factor of 1.17, which was introduced as an empirical correction to our reference BSSE-free MP2/6-311+G(d,p) values.

Solvation Calculations. To examine the impact of solvation in the tautomeric population of T and its thioketo-derivatives, the relative hydration free energies of tautomers was determined using (i) quantum mechanical self-consistent reaction field (SCRF) calculations and (ii) thermodynamic integration coupled to molecular dynamics simulations (MD/TI).

Within the SCRF framework the differential hydration between tautomers A and B ($\Delta\Delta G_{hyd}^{B-A}$) was determined from the hydration

(14) Orozco, M.; Hernandez, B.; Luque, F. J. *J. Phys. Chem. B* **1998**, *102*, 5228–5233.
(15) (a) Coleman, R. S.; Siedlecki, J. *J. Am. Chem. Soc.* **1992**, *114*, 9229–9230. (b) Xu, Y. Z. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *10*, 401. (c) Jing, Y.; Kao, J. F.; Taylor, J. S. *Nucleic Acids Res.* **1998**, *26*, 3845–3853. (d) Massey, A.; Xu, Y. Z.; Karran, P. *Curr. Biol.* **2001**, *11*, 1142–1146.
(16) Massey, A.; Xu, Y. Z.; Karran, P. *Curr. Biol.* **2001**, *11*, 1142–1146.
(17) (a) Connolly, B. A.; Newman, P. C. *Nucleic Acids Res.* **1989**, *17*, 4957–4974. (b) Massey, A.; Xu, Y. Z.; Karran, P. *DNA Repair* **2002**, *1*, 275–286. (c) Rao, T. V. S.; Haber, M. T.; Sayer, J. M.; Jerina, D. M. *Bioorg. Med. Chem. Lett.* **2000**, 907–910.
(18) Sintim, H. O.; Kool, E. T. *J. Am. Chem. Soc.* **2006**, *128*, 396–397.
(19) Spacková, N.; Cubero, E.; Sponer, J.; Orozco, M. *J. Am. Chem. Soc.* **2004**, *126*, 14642–14650.
(20) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45–48.
(21) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W. *Chem. Phys. Lett.* **1999**, *302*, 437–446.
(22) Frisch, M. J. et al. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.
(23) Boys, S.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
(24) Sponer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.

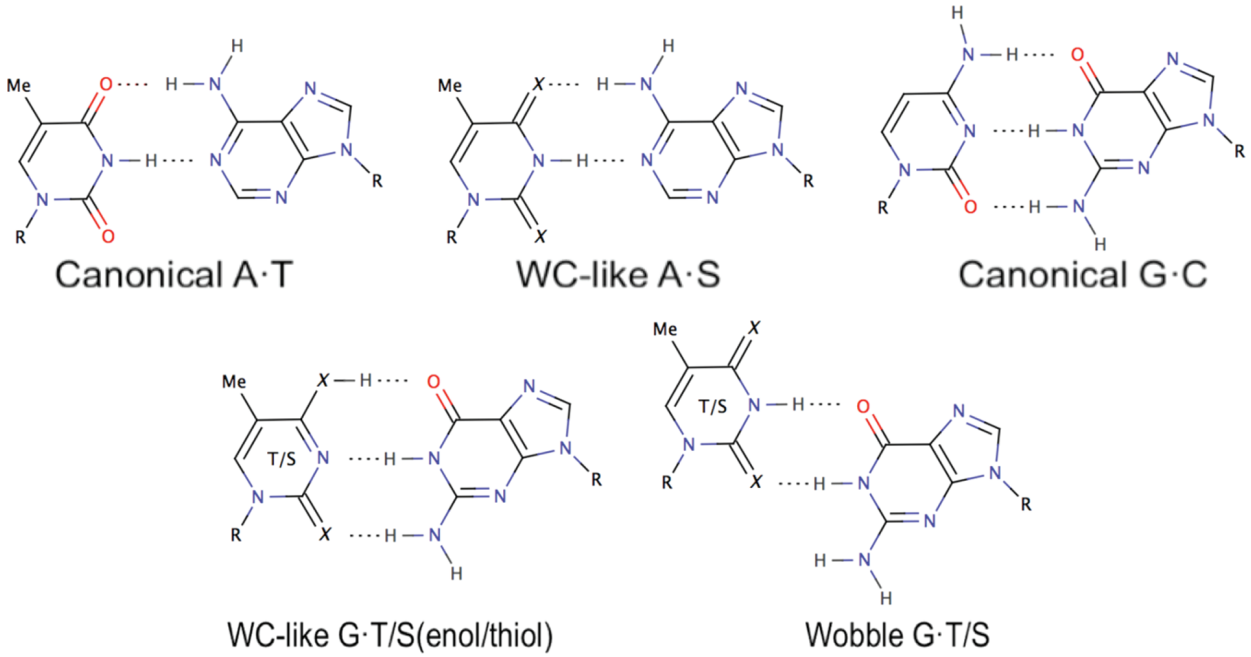


Figure 2. Different pairing schemes considered in this study for recognition of adenine and guanine. X = O or S.

free energies of those tautomers (ΔG^{A}_{hyd} and ΔG^{B}_{hyd}) computed using our HF/6-31G(d)-optimized version[25] of the IEF/MST method.[26] In contrast, MD/TI calculations yield directly the relative hydration free energy between tautomers ($\Delta\Delta G^{B-A}_{hyd}$) from the reversible work required to mutate tautomer A to tautomer B in aqueous solution. Every mutation was performed in both forward (A→B) and reverse (B→A) directions and using in each case either 21 or 41 windows. Each of these windows (40 ps) was divided in two halves, which means that 8 independent estimates were derived for each $\Delta\Delta G^{B-A}_{hyd}$ value, thus allowing us to estimate the statistical confidence of the mean values. Additionally, mutations defining futile cycles (A→B→C→A) were performed to further check the statistical error in our estimates. Noteworthy, all the futile cycles considered here were closed with an error lower than 0.4 kcal/mol.

Finally, the tautomerization free energy in aqueous solution ($\Delta G^{A\rightarrow B}_{taut}$(sol); eq 1) was derived by adding the best estimate of the gas-phase tautomerization free energy ($\Delta G^{A\rightarrow B}_{taut}$(gas)) to the differential hydration free energy between tautomers ($\Delta\Delta G^{B-A}_{hyd}$).

$$\Delta G^{A\rightarrow B}_{taut}(sol) = \Delta G^{A\rightarrow B}_{taut}(gas) + \Delta\Delta G^{A\rightarrow B}_{hyd} \qquad (1)$$

DNA Simulations. To examine the impact of thioketothymines in the DNA, two models systems, d(CGCGA**X**GACGCG)·d(CGCGTC**Y**TCGCG), and d(CGCGA**X**TACGCG)·d(CGCGTA-**Y**TCGCG) (X = T, ^{2}S or ^{4}S; Y = G or A), were considered. They are close to Dickerson's dodecamer, which has been largely studied by us and other groups,[27,28a] but display a central triad equal to that used in experimental studies on the stability of thioketothymine-containing DNA duplexes.[17b,18] For each sequence structural models of A·T, A·^{2}S, A·^{4}S, G·T(keto), G·^{2}S(thioketo), G·^{4}S(thioketo), G·T(enol), and G·^{4}S(thiol) were built up using as reference the structure equilibrated after 1 μs MD simulation of Dickerson's dodecamer.[28] The 16 systems were neutralized by a suitable number of Na^{+} ions and hydrated with around 4000 water molecules. Each system was optimized, thermalized, and equilibrated using our standard protocol with extended equilibration periods.[29] Production runs were extended for 50 ns, using the last 10 ns to characterize the structural aspects of the duplexes.

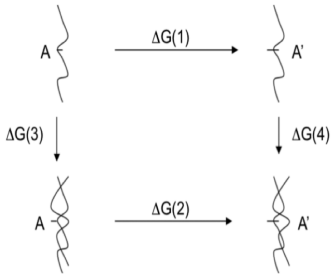


Figure 3. Example of thermodynamic cycle used to determine the contribution of mutation A→A′ to the stability of the DNA duplex.

MD/TI calculations were performed to explore the impact of different tautomers in the A·X and G·X recognition. To this end, different mutations were studied starting from the structure collected at the end of the MD simulations (see above). First, we analyzed the difference in reversible work associated to mutations between canonical keto(thioketo)/amino tautomers of T and thioketothymines in both duplexes and isolated single strands (Figure 3), as it provides the difference in stability in X·Y pairs induced by changing X from T to ^{2}S and ^{4}S (Y = G or A). Following our standard protocols aimed at reducing noise,[13b,19] mutations in the single strand were performed for 5-mers of sequences d(GA**X**TA) and d(GA**X**GA), which were equilibrated for 5 ns prior to the mutation. MD/TI calculations were also used to investigate the possibility that enol-imino or thiol-imino forms played a role in mismatched G·X pairs (it does not make chemical sense to study the same for A·X pairs).

(25) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M. *J. Mol. Struct.: THEOCHEM* **2005**, *727*, 29–40.
(26) Curutchet, C.; Orozco, M.; Luque, F. J. *J. Comput. Chem.* **2001**, *22*, 1180–1193.
(27) (a) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 2179–2183. (b) Noy, A.; Pérez, A.; Lankas, F.; Javier Luque, F.; Orozco, M. *J. Mol. Biol.* **2004**, *343*, 627–638.
(28) (a) Pérez, A.; Luque, F.; Orozco, M. *J. Am* **2007**, *129*, 14739–14745. (b) Orozco, M.; Pérez, A.; Noy, A.; Luque, F. J. *Chem. Soc. Rev.* **2003**, *32*, 350–364. (c) Cheatham, T. E.; Kollman, P. A. *Ann. Rev. Phys. Chem.* **2000**, *51*, 435–471. (d) Cheatham, T. E.; Kollman, P. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435–471. (e) Beveridge, D.; McConnell, K. J. *Curr. Opin. Struct. Biol.* **2000**, *10*, 182–196.
(29) (a) Shields, G. C.; Laughton, C. A.; Orozco, M. *J. Am. Chem. Soc.* **1997**, *119*, 7463–7469. (b) Soliva, R.; Laughton, C. A.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1998**, *120*, 11226–11233.

For this purpose mutations of keto/thioketo-amino to enol/thiol-imino forms were performed in both duplexes in the central T or S. The reversible work associated with these processes can be assimilated to a "specific-DNA solvation" effect ($\Delta\Delta G^{A\rightarrow B}_{solvDNA}$), which can be combined with the intrinsic gas-phase tautomerization ($\Delta G^{A\rightarrow B}_{taut}$(gas)) to obtain the tautomerization free energy in the duplex ($\Delta G^{A\rightarrow B}_{taut}$(DNA); eq 2).

$$\Delta G^{A\rightarrow B}_{taut}(DNA) = \Delta G^{A\rightarrow B}_{taut}(gas) + \Delta\Delta G^{A\rightarrow B}_{solvDNA} \quad (2)$$

where A and B stands for two tautomeric forms of thymine or thioketothymine.

Note that in the case where eq 2 reveals that for one of the pairings the canonical keto/thioketo tautomer is not the most stable one, the difference in stability due to a T→S mutation derived from thermodynamic cycles in Figure 3 has to be corrected to account for the presence of an alternative tautomer (see eq 3a).

$$\Delta\Delta G^{stab}_{T\rightarrow S} = \Delta\Delta G^{stab}_{T\rightarrow S}(A,A') \text{ if } \Delta G^{A\rightarrow B}_{taut}(DNA,T) > 0 \text{ and } \Delta G^{A'\rightarrow B'}_{taut}(DNA,S) > 0 \quad (3a)$$

$$\Delta\Delta G^{stab}_{T\rightarrow S} = \Delta\Delta G^{stab}_{T\rightarrow S}(A,A') - \Delta G^{A\rightarrow B}_{taut}(DNA,T) \text{ if } \Delta G^{A\rightarrow B}_{taut}(DNA,T) < 0 \text{ and } \Delta G^{A'\rightarrow B'}_{taut}(DNA,S) > 0 \quad (3b)$$

$$\Delta\Delta G^{stab}_{T\rightarrow S} = \Delta\Delta G^{stab}_{T\rightarrow S}(A,A') + \Delta G^{A'\rightarrow B'}_{taut}(DNA,S) \text{ if } \Delta G^{A\rightarrow B}_{taut}(DNA,T) > 0 \text{ and } \Delta G^{A'\rightarrow B'}_{taut}(DNA,S) < 0 \quad (3c)$$

$$\Delta\Delta G^{stab}_{T\rightarrow S} = \Delta\Delta G^{stab}_{T\rightarrow S}(A,A') + \Delta G^{A'\rightarrow B'}_{taut}(DNA,S) - \Delta G^{A\rightarrow B}_{taut}(DNA,T) \text{ if } \Delta G^{A\rightarrow B}_{taut}(DNA,T) < 0 \text{ and } \Delta G^{A'\rightarrow B'}_{taut}(DNA,S) < 0 \quad (3d)$$

where A and A′ stand for the canonical tautomer (keto/thioketo) of thymine (T) and thioketothymine (S), respectively, and B and B′ stands for the corresponding enol/thiol tautomer. Note that the correction is zero if the canonical tautomers are the dominant species in the duplex.

Finally, MD/TI calculations were used to investigate the differential stability of A·X pairings with respect to G·X mismatches. To this end, A→G (and G→A) mutations in the duplex and in a single stranded oligonucleotide (see above) were performed to determine the "stabilization" free energy of the mismatch using standard thermodynamic cycles. Note that these values can be combined (using equations above) with the "stabilization" energies for T→S mutation (paired to A or G) and with the tautomerization free energy to obtain a full thermodynamic description of the tautomeric/binding scenario of T and S in both normal and mismatched DNAs.

All MD/TI mutations in DNA were performed following the same protocol described above for pure solvent calculations, except for G→A (and A→G) mutations, where the simulation times were 3-fold larger to ensure convergence. In all cases mutations were carefully monitored to guarantee the lack of hysteresis effects and the goodness of original and final points. As in pure solvent calculations, 8 independent estimates were determined, and futile cycles were designed to check the statistical quality of our estimates.

Technical Details of Molecular Dynamics Simulations. All MD simulations were carried out in the isothermal−isobaric ensemble (NPT; 1 atm, 298 K) using periodic boundary conditions and the Particle Mesh Ewald.[30] An integration step of 2 fs was used in conjunction with SHAKE,[31] which guarantees that all chemical bonds are kept at equilibrium distances. The latest version of the AMBER force-field, including parmBSC0 corrections[32] was used to describe standard nucleotides and nucleic acids. Parameters for thioketothymines (Supporting Information, Table S1) were obtained from different sources: (i) equilibrium parameters for bonded terms were taken from B3LYP optimized geometries, (ii) van der Waals parameters for sulfur were taken from a previous work,[19] and (iii) atomic charges for every tautomer were fitted using the standard RESP procedure using HF-31G(d) wave functions.[33] Since the accuracy of the force-field parameters to represent A·X and G·X (X = T, ^{2}S and ^{4}S) interactions is a crucial requisite we made an effort in analyzing the quality and compatibility of the derived force-field parameters. Results demonstrate that AMBER interaction energies correlate extremely well ($c = 1.00$; $r^2 = 0.99$) with scaled-MP2/6-311+G(d) estimates for the 9 systems considered here (see Supporting Information, Figure S1). Accordingly, the largest source of errors is expected to be statistical noise or uncertainties related to incomplete samplings in the simulations, not force-field artifacts.

All quantum mechanical calculations were done with a local version of Gaussian03 that incorporates the MST method. MD simulations were carried out using different modules of the AMBER9.0 suite of programs[34] and the GIBBS module in AMBER5.0. Analysis was mainly done using the PCAZIP suite of programs (http://mmb.pcb.ub.es/pcazip) and standard tools in AMBER. All calculations were carried out in the *MareNostrum* supercomputer at the Barcelona Supercomputing Centre and in local computers in our laboratory.

Experimental Studies. Because of the controversy in the experimental data available in the literature, specific experiments were designed to check the validity of the results derived from our theoretical models. To this end, we created DNAs containing either thymines or thioketothymines. Several sequences (including those studied in previous works) were considered, but detailed results are reported only for 5′-GCAATGGAXCCTCTA-3′/3′-CGTTAC-CTYGGAGAT-5′ (X = T, ^{2}S and ^{4}S; Y = A, G, C, T). Oligodeoxynucleotides were prepared using an Applied Biosystems 3400 DNA synthesizer. The corresponding 2-thioketothymidine and 4-thioketothymidine 2-cyanoethyl phosphoramidites were from commercial sources. 2-Thioketothymidine (Glen Research, USA) was protected with the toluoyl group. 4-Thioketothymidine (Link Tecknologies. Scotland) was protected with the 2-cyanoethyl group. The protecting groups of the natural bases were benzoyl for A and C and dimethylformamidine for G. The standard LV200 synthesis cycle and 0.02 M iodine solution was used. Syntheses were performed with the removal of the last DMT group (DMT off). Oligonucleotides carrying ^{2}S were deprotected in concentrated ammonia (50 °C, 1 h). Three different protocols were considered for deprotection of oligonucleotides carrying ^{4}S: (i) concentrated ammonia (50 °C, 1 h), (ii) 1 M 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) in acetonitrile (room temperature, 3 h) followed by 50 mM NaSH in concentrated ammonia (room temperature, 24 h; supplier's recommendation), and (iii) 50 mM NaSH in concentrated ammonia (room temperature, 24 h). All three deprotection protocols gave a similar HPLC profile. Finally, synthesized oligonucleotides were purified using reversed-phase HPLC. Solutions were as follows: Solvent A, 5% acetonitrile in 100 mM triethylammonium

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

(31) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(32) (a) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T.; Laughton, C.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829. (b) Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197. (c) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.

(33) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(34) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz JR, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

Table 1. Tautomerization Free Energy (kcal/mol) for Tautomers of *N*1-Methylthymine (T), 2-Thioketothymine (^{2}S), and 4-Thioketothymine (^{4}S) in the Gas Phase Determined at Different Levels of Theory[a]

tautomer	MP2/6-3111G(d,p)	MP2/cc-pVDZ	MP2/cc-pVTZ	MP2/cc-pVQZ	ΔCCSD(T)/MP2	best estimate CCSD(T)/CBS
T_H2c	18.3	18.4	17.4	17.2	−0.2	*16.6*
T_H2t	29.5	29.6	27.9	27.7	−0.6	*26.6*
T_H4c	12.4	12.4	11.4	11.2	0.1	*10.8*
T_H4t	19.8	19.6	17.7	17.4	−0.1	*16.7*
2S_H2c	15.6	15.6	15.7	15.6	−0.9	*15.0*
2S_H2t	20.3	21.3	20.8	20.6	−1.2	*19.7*
2S_H4c	12.0	12.2	11.3	11.2	0.3	*11.0*
2S_H4t	19.8	19.7	17.9	17.6	0.0	*17.1*
4S_H2c	17.6	17.9	17.0	16.9	−0.1	*16.4*
4S_H2t	29.3	29.7	28.1	28.0	−0.4	*27.0*
4S_H4c	10.9	10.4	10.7	10.7	−0.7	*10.3*
4S_H4t	13.2	13.5	13.0	12.8	−0.9	*12.2*

[a] The best estimate (quality CBS/CCSD(T)) is displayed in italics. All values are referred to the canonical keto/amino or thioketo/amino tautomers (see Figure 1 for nomenclature).

acetate (pH 6.5); solvent B, 70% acetonitrile in 100 mM triethylammonium acetate pH 6.5. Columns: Nucleosil 120C18 (10 μm), 200 mm × 10 mm. Flow rate: 3 mL/min. Conditions: 20 min linear gradient from 0−50% B. Mass spectrometry (MALDI-TOF): 15mer with 4ST: found 4569.4, expected 4568; 15mer with 2ST: found 4568.5, expected 4568. UV: 15mer with 4ST max 259.3 and 337.8; 15mer with 2ST max 260.4.

Melting experiments were performed to determine the relative stability of the duplexes. For this purpose solutions of equimolar amounts of oligonucleotides (5′-GCAATGGAXCCTCTA-3′/3′-CGTTACCTYGGAGAT-5′) were mixed in 50 mM NaCl, 10 mM sodium phosphate buffer pH 7.0. The solutions were heated to 90 °C, allowed to cool slowly to room temperature, and stored at 4 °C. UV absorption spectra and melting experiments (absorbance vs temperature) were recorded in 1-cm path-length cells using a spectrophotometer, with a temperature controller and a programmed temperature increase rate of 1 °C/min. Melts were run by duplicate using a strand concentration of 7−8 μM at 260 nm. Melting curves were analyzed by computer-fitting the denaturation data using Meltwin 3.5 software. On the basis of multiple experiments, the uncertainty in T_m values was estimated at ± 0.7 °C.

Results and Discussion

Gas-Phase Studies. High level ab initio calculations demonstrate that the tautomeric preference of T, ^{2}S, and ^{4}S in the gas phase is vastly dominated by the canonical keto/thioketo forms (Table 1). The convergence in the results with respect to the level of calculation is almost perfect, suggesting that our best estimates (quality CBS/CCSD(T)) are very accurate. The canonical keto form is the most stable tautomer of T (by almost 11 kcal/mol), indicating that there are 10^8 thymines in the keto tautomer for each one in the enol (T_H4c) form. No relevant changes in the tautomerization scenario in vacuum are found when ^{2}S or ^{4}S are considered (Table 1). In summary, contrary to previous suggestions,[18] the intrinsic tautomeric preferences of T and its thioketothymine analogues are almost identical and strongly favor Watson−Crick pairing, which is in contrast with the behavior found for cytosine, which exists in the gas phase in an "unusual" imino tautomeric form.[13b,35]

Aqueous Simulations. Water typically stabilizes more polar tautomers with respect to the canonical forms (see Table 2). However, even though the quantitative effect of hydration in the tautomeric stability can be very large (up to 10 kcal/mol for some unusual tautomers), no significant changes are expected in the population of the most stable tautomers, and the keto/thioketo forms are predicted to be the dominant species for T, ^{2}S, and ^{4}S in water. In particular, the keto/thioketo→enol/thiol tautomerism involving O/S at position 4 is not significantly modified by the effect of water, which argues against the presence of these enol forms in physiological conditions. It has to be recognized that MD/TI and MST estimates of solvation effects are more prone to numerical uncertainties than the gas-phase CCSD(T)/CBS tautomerization free energies and that the free energy estimates in Table 2 might be not as accurate as the gas-phase values. However, the excellent agreement found between MD/TI and MST/SCRF calculations (also found in previous studies on similar systems[13b]) gives confidence to the estimated differences in hydration free energies (see Table 2 and Figure S2, Supporting Information), and to the tautomerization free energy in aqueous solution, for which errors are not expected to be larger than 1 kcal/mol.

Table 2. Hydration Free Energy (Relative to the Respective Canonical Tautomers) of the Different Tautomers of *N*1-Methylthymine (T), 2-Thioketothymine (^{2}S), and 4-Thioketothymine (^{4}S) Determined from MST-SCRF and MD/TI Simulations[a]

tautomer	ΔG_{taut} gas phase	ΔG_{sol} (MST)	ΔG_{sol} (MD/TI)	ΔG_{taut} water (MST)	ΔG_{taut} water (MD/TI)
T_H2c	16.6	−3.6	−2.8	13.0	14.8
T_H2t	26.6	−9.6	−7.6	17.0	19.0
T_H4c	10.8	−1.7	−1.5	9.1	9.3
T_H4t	16.7	−6.5	−5.4	10.2	11.3
^{2}S_H2c	15.0	−0.7	−1.9	14.3	13.1
^{2}S_H2t	19.7	−2.9	−3.2	16.8	16.5
^{2}S_H4c	11.0	−3.2	−1.3	7.8	9.7
^{2}S_H4t	17.1	−8.2	−6.2	8.9	10.9
^{4}S_H2c	16.4	−5.1	−5.5	11.3	10.9
^{4}S_H2t	27.0	−11.0	−9.7	16.0	17.3
^{4}S_H4c	10.3	1.5	−0.3	11.8	10.0
^{4}S_H4t	12.2	−0.1	−1.9	12.1	10.3

[a] Solvation free energies are added to the best estimates of the gas-phase tautomerization free energy (first column; taken from Table 1) to obtain the tautomerization free energy in aqueous solution. All values are in kcal/mol.

DNA Simulations. Equilibrium MD simulations on d(CGCGA**X**GACGCG)·d(CGCGTC**Y**TCGCG) and d(CGCGA**X**-TACGCG)·d(CGCGTA**Y**TCGCG), where X = T, ^{2}S, and ^{4}S in their standard keto/thioketo tautomeric form, were performed to study the structural impact of introducing (i) G·T mismatches and (ii) thioketothymines into DNA duplexes. All trajectories were stable, and the helical B-like structures were always well preserved during the entire simulation (Figure S3, Supporting Information). Thus, the DNA appeared as a very robust structure, able to accommodate local distortions and keeping the general conformation unaltered. The structural impact of having ^{4}S or ^{2}S instead of T paired to A is negligible, and the small variations detected in helical and groove parameters (Table S2, Supporting Information) are within the thermal noise of the reference simulation. G·T and G·S mismatches (for standard keto/thioketo tautomers) lead to wobble pairings associated with local distortions clearly visible in roll and twist parameters (see examples in Figure S4, Supporting Information). As expected, wobble interactions define a very flexible pattern of contacts in G·T and G·S mismatches, where breathing is much more common and severe than in canonical A·T and noncanonical A·^{2}S steps, which conserve the Watson−Crick pairing since sulfur atom in 2-thioketothymine is not involved in hydrogen bonding (see Figure 2), while G·T and G·S mismatches hardly

(35) Kobayashi, R. *J. Phys. Chem. A* **1998**, *102*, 10813–10817.

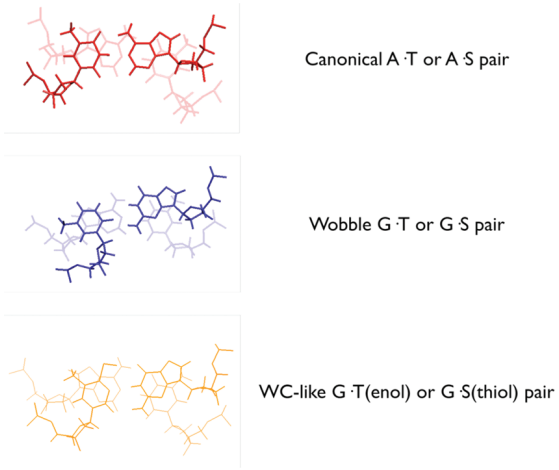


Figure 4. Representation of hydrogen bonding and stacking interactions found in simulations of canonical A·X, wobble G·X, and Watson−Crick-like G·X (enol/thiol) pairings.

maintain two hydrogen bonds along the trajectory (see Figure S5, Supporting Information). No major structural differences are found in d(G·S) pairs compared to the d(G·T) ones, though the presence of sulfur atom at position 2 decreases the percentage of time in which bases are hydrogen-bonded (see Figure S5, Supporting Information and below) but keeping canonical structure without inducing large distortions during simulation time since stacking interactions are well conserved (see Table 4 and Figure S3, Supporting Information).

Finally, we performed simulations for duplexes d(CGCGA**X**GACGCG)·d(CGCGTC**Y**TCGCG) and d(CGCGA**X**TACGCG)·d(CGCGTA**Y**TCGCG) (X being the unusual enol/thiol tautomeric form of T/^{4}S and Y = G). All trajectories were stable, sampling B-DNA conformers close to the canonical helix (Figure S3, Supporting Information) and displaying the expected triple Watson−Crick-like hydrogen-bond scheme at the substitution site (see Figure 4 and Figure S5, Supporting Information) with very little distortions from ideal values. Disruption movements in helical space, which were common in duplexes containing G·X mismatches, (where X was in the canonical keto/thioketo form) are drastically reduced when X is in the enol/thiol form (Figures S4 and S5, Supporting Information). As previously noted by others,[18] these findings suggest that the DNA environment favors enol/thiol tautomers when T/S are paired with G (but obviously not to A). However, it is unclear whether those pairings with G are strong enough to justify the presence in the DNA of tautomers with low intrinsic stability (10−11 kcal/mol lower than the canonical tautomer; see Table 1).

To examine the importance of enol/thiol tautomers of T/S in the DNA duplex and to analyze the preference between G·X and A·X pairings, MD/TI simulations were run, from which we estimate (i) the change in stability induced by the A→G mutation paired to T, ^{2}S, or ^{4}S in their standard keto/thioketo forms, (ii) the change in stability due to the T→^{2}S and T→^{4}S mutations (in their standard keto/thioketo forms) paired to A or G, and (iii) the change in stability of G·T or G·^{4}S pairs due to the tautomeric change of T or ^{4}S from the keto/thioketo species to the 4-enol/thiol one. This vast set of calculations (performed for two different sequences and replicated to obtain 8 individual replicas of each value) allowed us to examine the stability of pairing/mismatching/ tautomerism for thymine and thioketothymines.

The first set of MD/TI calculations was designed to determine the change in stability induced by the A→G mutation paired to T, ^{2}S, or ^{4}S in their keto/thioketo form. Despite the complexity of the A→G mutation in the duplex (it implies a change of hydrogen-bond pattern from Watson−Crick to wobble pairings), all mutations happened smoothly, without any apparent discontinuity indicative of hysteresis effects (see Figure S6, Supporting Information for examples). The independent estimates of the reversible work associated with each mutation also agreed well (see Figure S7, Supporting Information for examples), allowing us to determine the free energy associated to the transduction change for an adenine to a guanine with very small statistical uncertainties. Results in Table 3 demonstrate that the A→G mutation is associated with a significant destabilization of the duplex (between 1.5 and 2.1 kcal/mol), which seems to be quite independent of the sequence at the mutation site (AXG or AXT) and of the nature of the pyrimidine (T or S). Analysis of the energetic contributions due to hydrogen-bond and stacking interactions in the central trimer indicates that the poor stability of G·X wobble pairs compared to canonical Watson−Crick A·X ones is not related to either poor stacking or weak hydrogen bonds (see Table 4) but probably to the mechanical distortion in the helix related to the wobble pairing geometry (see Figure S6, Supporting Information) and to the large cost of dehydrating a guanine (9.1 kcal/mol larger than that of an adenine according to our MST calculations).

It is worth noting that our theoretical simulations agree with the available data in the literature in all cases but for the A→G mutation in presence of ^{4}S, where our results are close to the

Table 3. Change in Free Energy (kcal/mol) Associated with the A→G Mutation in the Two Sequences Considered Here (Identified by the Central Triad)[a]

mutation	central triad (**X**)	comp base	ΔΔ*G*(stab)	exptl data (lit.)[b]
A→G	A**X**G	T	1.5 ± 0.2	1.7[c]/1.7
	A**X**T	T	1.6 ± 0.2	1.7[d]/2.4
A→G	A**X**G	^{2}S	2.2 ± 0.2	1.7[c]/3.4
	A**X**T	^{2}S	2.1 ± 0.2	
A→G	*A**X**G*	*^{4}S*	*1.7 ± 0.1*	*−0.5[c]/−0.7*
	A**X**T	^{4}S	1.5 ± 0.2	1.9[d]/2.6

[a] Values were computed/measured with the complementary pyrimidine equal to thymine (T), 2-thioketothymine (^{2}S), and 4-thioketothymine (^{4}S). Standard errors are also shown. *The case of large discrepancy with the literature is in italics.* [b] Values after the slash refers to estimates obtained from a linear regression (see ref 40) with melting temperatures reported in the corresponding paper. [c] Data from ref 18. [d] Data from ref 17b.

Table 4. Hydrogen-Bond Interaction (in the Central Pair) and Stacking (Both Intra- and Interstrand for the Central Triad) Energies for the Substitution Site in the Two Different Sequences Used in Our Simulations[a]

pair	central triad (**X**)	*E*(hbond)	*E*(stack)	*E*(tot)
T·A	A**X**G	−11.2	−30.2	−41.4
	A**X**T	−11.1	−29.4	−40.5
T·G	A**X**G	−13.2	−33.0	−46.2
	A**X**T	−13.9	−28.7	−42.6
^{2}S·A	A**X**G	−10.1	−32.4	−42.5
	A**X**T	−9.9	−31.4	−41.3
^{2}S·G	A**X**G	−10.1	−32.6	−42.7
	A**X**T	−10.5	−31.3	−41.8
^{4}S·A	A**X**G	−9.6	−31.1	−40.8
	A**X**T	−9.6	−31.0	−40.6
^{4}Sthiol·G	A**X**G	−19.7	−35.2	−54.9
	A**X**T	−20.0	−30.7	−50.6
^{4}S·G	A**X**G	−11.8	−35.1	−46.9
	A**X**T	−12.4	−30.3	−42.7
Tenol·G	A**X**G	−25.4	−34.6	−59.9
	A**X**T	−25.3	−30.7	−56.0

[a] All values are in kcal/mol.

Table 5. Free Energy Change Associated with the Substitution of Thymine by Thioketothymine (Keto and Thioketo Tautomers) in the Two Sequences Considered Here (Identified by the Central Triad)[a]

mutation	central triad (**X**)	comp base	$\Delta\Delta G$(stab)	exptl data (lit.)[b]
T→^{4}S	A**X**G	A	-0.3 ± 0.2	0.4[c]/0.8
	A**X**T	A	0.4 ± 0.1	0.4[d]/0.8
T→^{4}S	A**X**G	G	0.3 ± 0.2	*−1.8[c]/−2.2*
	A**X**T	G	-0.3 ± 0.1	0.4[d]/1.1
T→^{2}S	A**X**G	A	-0.5 ± 0.2	0.1[c]/0.0
	A**X**T	A	-0.7 ± 0.2	−0.9[e]/−0.9
T→^{2}S	A**X**G	G	0.5 ± 0.2	0.1[c]/1.1
	A**X**T	G	0.4 ± 0.2	0.5[e]/0.5

[a] Values were computed/measured with the complementary purine equal to G or A. Standard errors in the theoretical estimates are displayed. All values are in kcal/mol. *The case of large discrepancy with the literature is in italics.* [b] Values after the slash refers to estimates obtained from a linear regression (see ref 40) with melting temperatures reported in the corresponding paper. [c] Data from ref 18. [d] Data from ref 17b. [e] Data from Haynes's group on LNA, ref 39.

Table 6. Change in Stability of the Duplex Due to the Tautomeric Change from Keto/Thioketo Forms to the Enol/Thiol Species for T and ^{4}S Paired to G ($\Delta\Delta G$(stab))[a]

mutation	central triad (**X**)	ΔG(solv)	ΔG(int)	$\Delta\Delta G$(stab)
^{4}S→^{4}S_H4C	A**X**G	-9.2 ± 0.3	10.3	1.1 ± 0.3
	A**X**T	-7.9 ± 0.2	10.3	2.4 ± 0.2
T→T_H4C	A**X**G	-8.1 ± 0.2	10.8	2.7 ± 0.2
	A**X**T	-7.7 ± 0.2	10.8	3.1 ± 0.2

[a] The solvation term (ΔG(solv)) accounts for the effect of DNA, counterions, and water on the equilibrium computed from MD/TI calculations. The intramolecular term (ΔG(int)) represents the intrinsic free energy of tautomerization and is computed at the QM level (see Table 1). All values are in kcal/mol.

values reported by Karran[17b] but deviate more than 2 kcal/mol from more recent estimates.[18] To determine whether this discrepancy might be attributed to errors in the simulation conditions, T (in the keto form and paired to A or G) was mutated to ^{2}S or ^{4}S (in the thioketo species and paired to the same purine) and the theoretical values were compared with experimental data in the literature. These mutations are technically simpler and easier to obtain with small noise, and hence they are ideal to detect inconsistencies in the previous sets of calculations. The results (Table 5) confirm those shown in Table 3 (in fact all futile cycles can be closed with errors close to zero; see Figure S7, Supporting Information). There is a good agreement with all experimental data with the only exception of the T→^{4}S mutation in the presence of G, where our simulations predict small changes in stability differing by around 2 kcal/mol from the estimates reported in ref 18. Inspection of the different interaction terms (see Table 4) suggests that the T→S mutation (for keto/thioketo tautomers) is small due to the balance between the lost of hydrogen-bond energy (sulfur is a poorer H-bond acceptor than oxygen) and the gain in stacking energy related to the stronger dispersive interactions of the sulfur.

The preceding results demonstrate that statistical errors in the simulations protocols cannot be responsible of the sizable discrepancies between theoretical results and experimental data given in reference.[18] Nevertheless, such a difference could be originated from the presence of a thiol tautomer of ^{4}S able to form a Watson−Crick-like pairing with G. To check this possibility (see Methods), keto (T) and thioketo (^{4}S) tautomers were mutated into the corresponding 4-enol and 4-thiol species in the presence of G, and the associated free energy was determined using MD/TI simulations. The results (Table 6) demonstrate the dramatic effect of DNA stabilizing enol/thiol tautomers when T (around 8 kcal/mol) or ^{4}S (up to 9 kcal/mol) are paired to G. However, such a large stabilization, which reflects the formation of the third G·X hydrogen bond (see Table 4), does not suffice to invert the intrinsic tautomeric preferences of neither T nor ^{4}S (see Tables 1 and 6). Thus, our calculations suggest that the keto→enol mutation of T in DNA (paired to G) is disfavored by 3 kcal/mol, and such a difference amounts to 1.1−2.4 kcal/mol for the thioketo→thiol mutation of ^{4}S. This means that the enol form of T, which populates only $1/10^8$ in the gas phase ($1/10^7$ in aqueous solution), has a population of $1/10^2$ in the mismatched (G·T) DNA. For the thiol form of ^{4}S, which was in the range of $1/10^7$ in the gas or aqueous phase, the population increases to a sizable 1/5−1/50 when paired in front of G in a DNA duplex. It is then clear that mismatched pairings in duplex DNA has a dramatic effect in the tautomeric scenario of thymine and thioketothymines, but even in the case of A→G transductions keto/thioketo tautomers are the dominant species in DNAs. Accordingly, our data strongly support the wobble scheme as the prevalent pairing for both G·T and G·S pairings, even though a small fraction of thiol tautomer might be expected for G·^{4}S dimers. Note that these results agree with the known tautomeric preferences of T in DNA, in particular in G·T mismatches,[17b] but not with recent suggestions derived from experimental data and semiempirical calculations for ^{4}S,[18] which suggested the thiol tautomer (^{4}S_H4c) as the active species in G·^{4}S recognition in duplex DNA.

Overall, present theoretical calculations support Karran's experimental results and suggest that (i) the presence of ^{2}S or ^{4}S in the DNA does not dramatically alter the structure or stability of the duplex when paired to A and (ii) thioketothymines, including ^{4}S, respond as T to the presence of a X·G mismatch. Moreover, the increase in the population of thiol tautomer in ^{4}S when paired to G does not justify a change in the pairing scheme, which would explain a significant alteration in the relative stability of G·^{4}S vs A·^{4}S pairings. On the basis of these theoretical findings, we can suggest that thioketothymines can be safely used as mimics of T in DNA, thus leading to nucleic acids with improved chemical possibilities and without altering the DNA fidelity and integrity.

Experimental Validation. Theoretical calculations described above are expected to be accurate enough as to support our claims, but to further check the goodness of our theoretically derived conclusions a series of additional experiments were conducted. As noted in Methods, several oligos containing thymines and thioketothymines paired with both G and A were synthesized using solid-phase 2-cyanoethylphosphoramidite chemistry. The required synthons to incorporate 2-thioketothymidine and 4-thioketothymidine into oligodeoxynucleotides were obtained from commercial sources.[17a,36] Since 4-thioketothymidine is labile to ammonia,[17a] the dimethylformamidine group was selected for the protection of the 2′-deoxyguanosine. Ammonia treatment was performed either at room temperature for 24 h or at 55 °C for 1 h to minimize decomposition of the 4-thioketothymidine residue. Also we studied the use of 50 mM NaSH or the removal of the 2-cyanoethyl group with 1 M 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) solution in acetonitrile at room temperature for 3 h. In all cases the desired oligonucleotide was obtained as the major compound. The resulting oligonucleotides were purified by HPLC and gave the expected

(36) Kuimelis, R. G.; Nambiar, K. P. *Nucleic Acids Res.* **1994**, *22*, 1429–1436.

Table 7. Thermodynamic Parameters of Double-Stranded DNA to Single-Stranded DNA Transition[a]

basepair	ΔH (kcal/mol)	ΔS (cal/K·mol)	ΔG(kcal/mol)	T_m (°C)
T·A	−120.5	−344.4	−13.7	52.8
T·C	−109	−321.5	−9.3	40.5
T·G	−107.8	−311.4	−11.2	45.9
T·T	−99.7	−291.8	−9.1	39.5
^{2}S·A	−113.8	−323.7	−13.4	51.8
^{2}S·C	−80.0	−221.7	−8.2	37.0
^{2}S·G	−107.8	−311.4	−11.2	45.9
^{2}S·T	−105.6	−306.8	−10.4	44.1
^{4}S·A	−110.7	−313.7	−12.5	50.9
^{4}S·C	−95.1	−297.2	−8.5	38.1
^{4}S·G	−104.0	−301.9	−10.4	43.5
^{4}S·T	−95.6	−286.8	−8.6	38.2

[a] Duplex sequence: 5′-GCAATGGAXCCTCTA-3′/3′-CGTTACCTYGGAGAT-5′, X = T, ^{2}S, ^{4}S; Y = A, G, C, T). Conditions: 50 mM NaCl, 10 mM sodium phosphate buffer, pH 7.0.

molecular weight by mass spectrometry, confirming that the desired nucleobases were successfully introduced into the oligos.

Though different oligos were synthesized and tested, only results obtained for the pentadecamer sequence (5′-GCAATGGAXCCTCTA-3′/3′-CGTTACCTYGGAGAT-5′, X = T, ^{2}S, and ^{4}S; Y = A, G, C, T) are displayed here (results for the other sequences are available upon request to authors). Melting temperatures (T_m) and thermodynamic parameters of the different duplex are shown in Table 7. It is clear that the most stable duplex was that containing at d(X·Y) position an A·T base pair ($\Delta G = -13.7$ kcal/mol), and the T·G pair ($\Delta G = -11.2$ kcal/mol) is the most stable of the T mismatches ($\Delta G = -9.1$ (T·T) and -9.3 (T·C) kcal/mol). The difference in stability between A·T and G·T pairs found experimentally agree reasonably well with previous experimental and current theoretical estimates (see Table 3 and refs 37 and 38). The most stable pairing for ^{2}S also involves A ($\Delta G = -13.4$ kcal/mol), followed by the mismatch with G ($\Delta G = -11.2$ kcal/mol). Thus, the presence of ^{2}S has small impact in the stability of either canonical pairing with A or mismatch with G, in agreement with theoretical (and previous experimental) data (see Table 5 and ref 18). The most stable pair for ^{4}S is also formed with A ($\Delta G = -12.5$ kcal/mol), followed by the mismatch with G ($\Delta G = -10.4$ kcal/mol). As suggested by theoretical calculations, ^{4}S does not change the pairing scheme of DNA, and 4-thioketothymine when introduced into the DNA largely prefers to bind A, the mismatch with G being clearly less stable. Present experimental measures cannot directly rule out the involvement of thiol tautomers in the duplex formation in presence of thioketothymines, but the agreement between theoretical and experimental data makes very difficult to believe that thiol tautomers might play something else than a residual role in ^{4}S·G mismatches.

Conclusions

Combination of high-level quantum mechanical calculations with "state of the art" molecular dynamics and free energy calculations, complemented with experimental measures, allowed us to draw a complete picture of the tautomeric and binding properties of thioketothymines. It is found that keto/thioketo tautomers are the prevalent ones for both thymine and thioketothymine in the gas phase, the situation being mostly unaltered in water. When inserted into a DNA duplex, thioketothymines induces small changes in structure and stability. A guanine paired to thymine or thioketothymine help to stabilize minor enol/thiol forms, but this effect is not large enough to change the tautomeric preferences of the pyrimidines considered here. Overall, our theoretical results, confirmed by experimental measures, support aquite classical behavior for thioketothymines, which are incorporated in

(37) Eritja, R.; Horowit, D. M.; Walker, P. A.; Ziehler-Martin, J. P.; Boosalis, M. S.; Goodman, M. F.; Itakura, K.; Kaplan, B. E. *Nucleic Acids Res.* **1986**, *14*, 8135–8153.
(38) Aboul-ela, F.; Koh, D.; Tinoco Jr, I.; Martin, F. H. *Nucleic Acids Res.* **1985**, *13*, 4811–4824.
(39) Hughesman, C. B.; Turner, R. F.; Haynes, C. *Nucleic Acids Symp. Ser.* **2008**, 245–246.
(40) Kool, E. T.; Morales, J. C.; Guckian, K. M. *Angew. Chem., Int. Ed.* **2000**, *39*, 990–1009.
(41) Piccirilli, J. A.; Krauch, T.; Moroney, S.; Benner, S. *Nature* **1990**, *343*, 33–37.
(42) Henry, A. A.; Olsen, A. G.; Matsuda, S.; Yu, C.; Geierstanger, B. H.; Romesberg, F. E. *J. Am. Chem. Soc.* **2004**, *126*, 6923–6931.
(43) Strobel, H.; Dugue, L.; Marliere, P.; Pochet, S. *Nucleic Acids Res.* **2002**, *30*, 1869–1878.
(44) Seela, F.; Thomas, H. *Helv. Chim. Acta* **1995**, *78*, 94–108.
(45) Froehler, B. C.; Wadwani, S.; Terhorst, T. J.; Gerrard, S. R. *Tetrahedron Lett.* **1992**, *33*, 5307–5310.
(46) Povsic, T. J.; Dervan, P. B. *J. Am. Chem. Soc.* **1989**, *111*, 3059–3061.
(47) Heystek, L. E.; Zhou, H.; Dande, P.; Gold, B. *J. Am. Chem. Soc.* **1998**, *120*, 12165–12166.
(48) Altmann, K. H.; Freier, S. M.; Pieles, U.; Winkler, T. *Angew. Chem., Int. Ed.* **1994**, *33*, 1654–1657.
(49) Epple, C.; Leumann, C. *Chem. Biol.* **1998**, *5*, 209–216.
(50) Matteucci, M. D.; von Krosigk, U. *Tetrahedron Lett.* **1996**, *37*, 5057–5060.
(51) Gryaznov, S.; Schultz, R. G. *Tetrahedron Lett.* **1994**, *35*, 2489–2492.
(52) Lin, K. Y.; Matteucci, M. D. *J. Am. Chem. Soc.* **1998**, *120*, 8531–8532.
(53) Gutierrez, Terhorst, T. J.; Matteucci, M. D.; Froehler, B. C. *J. Am. Chem. Soc.* **1994**, *116*, 5540–5544.
(54) Gutierrez, A. J.; Froehler, B. C. *Tetrahedron Lett.* **1996**, *37*, 3959–3962.
(55) Seela, F.; Debelak, H. *Nucleic Acids Res.* **2000**, *28*, 3224–3232.
(56) Seela, F.; Becher, G. *Nucleic Acids Res.* **2001**, *29*, 2069–2078.
(57) Mikhailov, S. N.; Rozenski, J.; Efimtseva, E. V.; Busson, R. *Nucleic Acids Res.* **2002**, *30*, 1124–1131.
(58) Reddy, P. M.; Bruice, T. C. *J. Am. Chem. Soc.* **2004**, *126*, 3736–3747.
(59) Tor, Y.; Dervan, P. B. *J. Am. Chem. Soc.* **1993**, *115*, 4461–4467.
(60) Moran, S.; Rex, X. F. R.; Rumney IV, S.; Kool, E. T. *J. Am. Chem. Soc.* **1997**, *119*, 2056–2057.
(61) Moran, S.; Ren, R. X. F.; Kool, E. T. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10506–10511.
(62) Schweitzer, B. A.; Kool, E. T. *J. Am. Chem. Soc.* **1995**, *117*, 1863–1872.
(63) Yu, C.; Henry, A. A.; Romesberg, F. E.; Schultz, P. G. *Angew. Chem., Int. Ed.* **2002**, *41*, 3841–3844.
(64) Ogawa, A. K.; Wu, Y.; McMinn, D. L.; Liu, J.; Schultz, P. G. *J. Am. Chem. Soc.* **2000**, *122*, 3274–3287.
(65) McMinn, D. L.; Ogawa, A. K.; Wu, Y.; Liu, J.; Schultz, P. G. *J. Am. Chem. Soc.* **1999**, *121*, 11585–11586.
(66) Kong, P.; Lin, T.; Brown, D. M. *Nucleic Acids Res.* **1992**, *20*, 5149.
(67) Vallone, P. M.; Benight, A. S. *Nucleic Acids Res.* **1999**, *27*, 3589–3596.
(68) Anand, N. N.; Brown, D. M.; Salisbury, S. A. *Nucleic Acids Res.* **1987**, *15*, 8167–8176.
(69) Goodman, M. F. *Nature* **1995**, *378*, 260–263.
(70) Goodman, M. F. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10493–10495.
(71) Nishio, H.; Ono, A.; Matsuda, A.; Ueda, T. *Nucleic Acids Res.* **1992**, *20*, 777–782.
(72) Hazra, T. K.; Izumi, T.; Boldogh, I.; Imhoff, B.; Kow, Y. W. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3523–3528.
(73) Bruner, S. D.; Norman, D. P. G.; Verdine, G. L. *Nature* **2000**, *403*, 859–866.
(74) Xu, Y.; Zhang, X.; Wu, H. C.; Massey, A.; Karran, P. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 995–997.
(75) Kutyavin, I. V.; Rhinehart, R. L.; Lukhtanov, E. A.; Gorn, V. V. *Biochemistry* **1996**, *35*, 11170–11176.
(76) Diop-Frimpong, B.; Prakash, T. P.; Rajeev, K. G. *Nucleic Acids Res.* **2005**, *33*, 5297–5307.
(77) Appel, C. D.; Maxwell, E. S. *RNA* **2007**, *13*, 899–911.

canonical forms into DNA, introducing small changes in structure, stability, and fidelity properties of the duplex, which make them excellent surrogates of thymines for deriving nucleic acids with improved chemical properties and with interesting pharmacological profiles.

(78) Connolly, B. A. *Methods Enzymol.* **1992**, *211*, 36–53.
(79) Leumann, C. *Bioorg. Med. Chem.* **2002**, *10*, 841–854.
(80) Kool, E. T. *Curr. Opin. Chem. Biol.* **2000**, *4*, 602–608.
(81) Kim, T. W.; Kool, E. T. *J. Org. Chem.* **2005**, *70*, 2048–2053.
(82) Wiseman, H.; Halliwell, B. *Biochem. J.* **1996**, *313*, 17–29.
(83) Krueger, A. T.; Kool, E. T. *Curr. Opin. Chem. Biol.* **2007**, *11*, 588–594.
(84) Rae, P. M.; Steele, R. E. *Bio. Systems* **1978**, *10*, 37–53.
(85) Petruska, J.; Goodman, M. F.; Boosalis, M. S.; Sowers, L. C. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 6252–6256.
(86) Luque, F. J.; Bachs, M.; Aleman, C.; Orozco, M. *J. Comput. Chem.* **1996**, *17*, 806–820.
(87) Luque, F. J.; Zhang, Y.; Aleman, C.; Bachs, M.; Gao, J. *J. Phys. Chem.* **1996**, *100*, 4269–4276.

Acknowledgment. This work has been supported by the Spanish Ministry of Education and Science (BIO2006-01602, CONSOLIDER Project in Supercomputation, BFU2007-63287), the Spanish Ministry of Health (COMBIOMED network), the Fundación Marcelino Botín, and the National Institute of Bioinformatics.

Supporting Information Available: Correlations between quantum mehanics and molecular dynamics calculations, variation of helical parameters, and rms deviations profiles along trajectories and futile cycles derived from averaging free energy estimates. This material is available free of charge via the Internet at http://pubs.acs.org.

JA904880Y

4.1.1.1 Results and discussion

The tautomeric properties of thymine and two thioketoderivatives, 2- and 4-thioketothymines, have been studied by means of accurate *ab initio* calculations and molecular dynamics simulations coupled with free energy calculations. Previous results suggested that these modifications could stabilize DNA structure at least as good as the natural thymine leading to the fidelity consequences during the replication process (Sintim & Kool 2006). Indeed, those results suggested that some minor tautomeric forms could be present in the DNA double stranded environment which would explain the higher stability of certain mismatches. In the light of these results, we explored the impact of thioketothymines by means of both theoretical and experimental studies.

Tautomerization energies Both gas phase and solvated systems were considered in the calculation of the relative tautomerization energies. High level *ab initio* calculations were performed for different tautomeric forms of thymine (T), 2-thioketothymine (^{2}S) and 4-thioketothymine (^{4}S). As expected, the most stable tautomers were the canonical keto (or thioketo) form both for T and for the thioderivatives. MST/SCRF hydration calculations showed the stabilizing effect of water on the more dipolar tautomeric forms e.g., enol/thiol forms. However, no changes in the order of stability were observed confirming that keto/thioketo tautomeric forms are the predominant species for T, ^{2}S, and ^{4}S. Additional MD/TI technique was used to independently calculate the relative hydration free energies between tautomers from the reversible work required to mutate tautomer A to tautomer B in aqueous solution. MD/TI results confirm the validity of SCRF estimates of the effect of hydration on the tautomerism preferences.

Free energy calculations Molecular dynamics simulations were run with two different DNA sequences d(CGCGA**X**GACGCG)·d(CGCGTCYTCGCG) and d(CGCGA**X**TACGCG)·d(CGCGTA**Y**TCGCG), where X = T, ^{2}S, and ^{4}S in their most stable keto/thioketo tautomeric form to evaluate the influence of neighboring bases around the mutation site. These designs allowed us also to evaluate the effect of the G·T mismatch and the influence of thioketothymines in the DNA environment. The analysis of the MD trajectories shows, not surprisingly, that the wobble pairing in G·T and G·S exhibits frequent opening events. However, both A·T and A·^{2}S base pairs conserved canonical hydrogen bonding interactions since 2-thioketothymine S2 atom is not involved in direct contacts with their complementary base.

The last coordinates in the MD trajectories of the two DNA sequences considered in this study were used as starting coordinates in thermodynamic integration

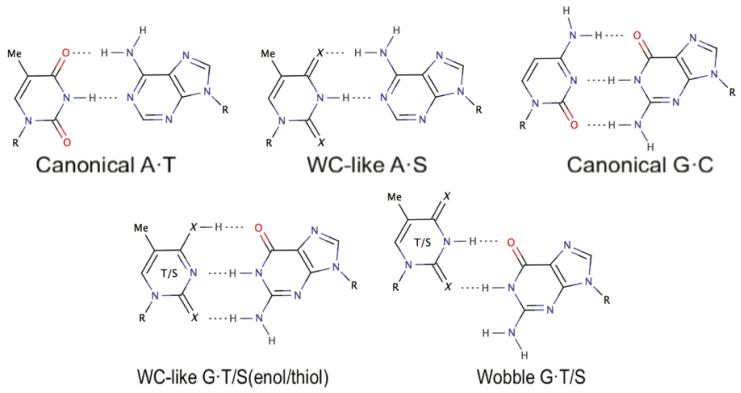


Figure 4.1: Different pairing schemes considered in this study for recognition of adenine and guanine. X = O or S.

calculations (MD/TI) to estimate the impact on DNA stability of the A→G mutation when paired either with T, ^{2}S, or ^{4}S (see Figure 5.1). Besides this, stability estimates were also computed for both T→^{2}S and T→^{4}S mutations when placing in front of adenine as complementary base. Finally, the free energy change associated to the keto/thioketo→enol/thiol chemical change was also calculated for the G·T and G·^{4}S base pairs.

mutation	central triad (X)	comp base	ΔΔG	exp data (lit)[b]
A→G	A**X**G	T	1.5 ± 0.2	1.7[c]/1.7
	A**X**T	T	1.6 ± 0.2	1.7[d]/2.4
A→G	A**X**G	^{2}S	2.2 ± 0.2	1.7[c]/3.4
	A**X**T	^{2}S	2.1 ± 0.1	
A→G	A**X**G	^{4}S	*1.7 ± 0.1*	-0.5[c]/-0.7
	A**X**T	^{4}S	1.5 ± 0.2	1.9[d]/2.6

Table 4.1: Change in free energy (kcal/mol) associated with the A→G mutation in the two sequences considered here (Identified by the central triad). [b]Values after the slash refers to estimates obtained from a linear regression (Kool et al. 2000) derived from the corresponding melting temperatures. [c]Data from (Sintim & Kool 2006). [d]Data from (Massey et al. 2002).

The results in tables 5.1 and 5.2 show our theoretical estimates of stability changes compared with two independent experimental measurements. There are two main conclusions that can be derived from these tables. These free energy values suggest on one side that, the destabilization associated to the A→G mutation is mainly due to the less stable wobble interaction pattern both for G·T and for G·S, and second, that the increase in population of the thiol ^{4}S tautomer when paired with G is not enough to reverse the intrinsic tautomeric preference of the molecule. This second conclusion disagrees with previous experimental results (Sintim & Kool 2006) that

suggested that thiol tautomer could change the relative stability of G·^{4}S and A·^{4}S base pairings. Note that agreement between experiment and theory is perfect in all other cases. Overall, our results suggest that thioketothymines mimic thymine interactions with adenine without disrupting the global structure of DNA duplex, without any change in tautomeric preferences and local destabilization of the duplex.

mutation	central triad (X)	comp base	ΔΔG(stab)	exptl data (lit.)[b]
T→^{4}S	AXG	A	-0.3 ± 0.2	0.4[c]/0.8
	AXT	A	0.4 ± 0.1	0.4[d]/0.8
T→^{4}S	AXG	G	0.3 ± 0.2	-1.8[c]/-2.2
	AXT	G	-0.3 ± 0.1	0.4[d]/1.1
T→^{2}S	AXG	A	-0.5 ± 0.2	0.1[c]/0.0
	AXT	A	-0.7 ± 0.2	-0.9[e]/-0.9
T→^{2}S	AXG	G	0.5 ± 0.2	0.1[c]/1.1
	AXT	G	0.4 ± 0.2	0.5[e]/0.5

Table 4.2: Free energy change associated with the substitution of thymine by thioketothymine (keto and thioketo tautomers) in the two sequences considered here (identified by the central triad). [b]Values after the slash refers to estimates obtained from linear regression derived from Kool et al. 2000. [c]Data from Sintim & Kool 2006. [d]Data from Massey et al. 2002. Data from Hughesman et al. 2008.

Experimental validation In the light of discrepancies between experimental results, an additional experimental validation was needed. For this purpose, several oligos containing thymines and thioketothymines paired either with G or with A were synthesized using standard solid phase procedures. Melting temperatures and the corresponding derived thermodynamic parameters were calculated from the melting curves. In the case of thymine, the most stable duplex contains the A·T base pair while the other mismatches were clearly destabilizing. These experimental measurements confirm our theoretical estimates and strongly suggest potential error in previous Kool's estimates.

4.1.1.2 Conclusions

The tautomeric preferences both in gas phase and solvation and in the DNA environment have been explored by means of high-level quantum mechanical calculations and molecular dynamics in combination with thermodynamic integration methods. Accurate QM calculations allowed us to determine the most populated species determined by their intrinsic preferences. The results have shown that thioketo tautomeric forms are the most stable as in the case of their natural counterpart thymine. The substitution of the latter by thioketothymines is well tolerated by the DNA duplex structure

basepair	ΔH (kcal/mol)	ΔS (cal/K· mol)	ΔG(kcal/mol)	Tm (°C)
T · A	-120.5	-344.4	-13.7	52.8
T · C	-109	-321.5	-9.3	40.5
T · G	-107.8	-311.4	-11.2	45.9
T · T	-99.7	-291.8	-9.1	39.5
^{2}S · A	-113.8	-323.7	-13.4	51.8
^{2}S · C	-80.0	-221.7	-8.2	37.0
^{2}S · G	-107.8	-311.4	-11.2	45.9
^{2}S · T	-105.6	-306.8	-10.4	44.1
^{4}S · A	-110.7	-313.7	-12.5	50.9
^{4}S · C	-95.1	-297.2	-8.5	38.1
^{4}S · G	-104.0	-301.9	-10.4	43.5
^{4}S · T	-95.6	-286.8	-8.6	38.2

Table 4.3: Experimental thermodynamic parameters of double-stranded DNA to single-stranded DNA transition.

although it generates small local distortions that are reflected in the destabilization of the duplex. Even when guanine is paired with thymine or thioketothymine its unlikely that the enol/thiol tautomeric form can be stable enough to change the intrinsic tautomeric preferences. We then support the idea that sulfur derivatives mimic thymidine which do not favor A→G transitions in contrast to previous suggestions. Is is then suggested that insertion of sulfur in DNA would lead to new reactivity possibilities without leading to a change in stability or fidelity of duplex.

4.1.1.3 References

Hughesman, C.B., Turner, R.F.B. & Haynes, C., 2008. Stability and mismatch discrimination of DNA duplexes containing 2,6-diaminopurine and 2-thiothymidine locked nucleic acid bases. Nucleic Acids Symposium Series, (52), pp.245–246.

Kool, E.T., Morales, J.C. & Guckian, K.M., 2000. Mimicking the Structure and Function of DNA: Insights into DNA Stability and Replication. Angewandte Chemie (International ed. in English), 39(6), pp.990–1009.

Massey, A., Xu, Y.Z. & Karran, P., 2002. Ambiguous coding is required for the lethal interaction between methylated DNA bases and DNA mismatch repair. DNA repair, 1(4), pp.275–286.

Sintim, H.O. & Kool, E.T., 2006. Enhanced Base Pairing and Replication Efficiency of Thiothymidines, Expanded-size Variants of Thymidine. Journal of the American Chemical Society, 128(2), pp.396–397.

4.1.2 The DNA-forming properties of 6-selenoguanine.

Ignacio Faustino, Carles Curutchet, F. Javier Luque and Modesto Orozco.
Publication submitted.

THE DNA-FORMING PROPERTIES OF 6-SELENOGUANINE

Ignacio Faustino[1,2], Carles Curutchet[3], F. Javier Luque[3] and Modesto Orozco[1,2,4]*

We present here an exhaustive characterization of the structure and properties of 6-selenoguanine, and isoster of guanine that can be introduced in DNA affecting in a very different way different canonical structures of DNA. By using a combination of state-of-the-art quantum mechanical calculations and atomistic MD simulations we determine that guanine → 6-selenoguanine (G→6SeG) substitution leads to stable DNA duplex, antiparallel triplex and G-quadruplex structures with local distortions. These non-negligible changes affect the thermodynamic stability of the mutation leading to clear destabilization of the G→6SeG mutation for all studied systems. Interestingly, the lowest effect has been found when the mutation was placed in the TFO strand in a reverse Hoogsteen orientation which favors the antiparallel triplex formation regarding the G-tetraplex formation.

1. Institute for Research in Biomedicine (IRB Barcelona). Baldiri Reixac, 10. Barcelona 08028, Spain.

2. Joint IRB-BSC program on Computational Biology, Barcelona, Spain.

3. Departament de Fisicoquímica. Facultat de Farmàcia. Universitat de Barcelona. Avgda Diagonal sn. Barcelona 08028, Spain.

4. Department of Biochemistry and Molecular Biology, University of Barcelona. Avgda Diagonal 647. Barcelona 08028, Spain.

* Correspondence to Prof. M. Orozco modesto.orozco@irbbarcelona.org

INTRODUCTION

The chemical language of nature has been limited by the availability of chemical elements in Earth. For example, nucleic acids are constituted exclusively by hydrogen, three first row (C,N,O) and one second row (P) elements. In fact, the chemical diversity of nucleobases is extremely limited, with has obvious advantages like the simplicity and fidelity in replication and transcription, and some disadvantages, like the need for very large genomes (genes are at least 3 time larger than the corresponding proteins), problems for specific sequence recognition, and limited possibilities for reactivity and recognition. Such limitations have generated the interest in many groups for expanding the genetic alphabet by introducing modified nucleobases with properties different to those of the coding ones [1-3]. One family of such modified nucleobases is originated by the substitution of carbonyl oxygens of coding nucleobases by sulfur or selenium [4-6].

Different thio-derivatives of nucleobases have been synthesized showing interesting properties. For example, thio-thymine can be incorporated into DNA in the place of a thymine, showing similar sequence specificity, and leading to only a modest lost of stability in the duplex [7-10]. Thio-thymine has been explored as a photo-activated crosslinker of DNA [11,12], with potential impact in the treatment of psoriasis and skin cancer [13]. 6-Thio-guanine can be also incorporated into different forms of DNA [4,14-18], leading generally to stable structures. Incorporation of 6-thio-guanine to DNA has been related to its wide repertoire of pharmaceutical activities in the treatment of pathologies such as leukemia, inflammatory diseases, AIDS and others [19-23]. Thio-cytosine has been also synthesized and incorporated into DNA, showing interesting properties in hybridization assays [15].

Seleno-derivatives of nucleobases have been less investigated than the corresponding thio-derivatives, but available data suggest that they can also display very interesting properties. For example, seleno-thymine derivatives have been incorporated in DNA as substitutes of thymine, showing good stability, excellent specificity in recognizing adenine and giving excellent conductor properties to the resulting DNA [24-26]. Very interesting, the high specificity of 2-seleno-uridine for adenine has been used in RNA

engineering as a method to reinforce Watson-Crick A·U (A·SeU) pairing against the common (in RNA) G·U (G·SeU) wobble interaction, leading then to changes in the secondary structure of target RNAs [27]. 6-Selenoguanine (6SeG) is known since the sixties, when it was described as an active drug against different types of lymphomes, and more recently, against hepatomes {Lin:2009br}-CITATION_IS_EMPTY. It has been described that 6SeG can be efficiently incorporated into DNA using standard phosphoramidite chemistry, leading to stable duplexes that exhibit anomalous dispersion, which has been widely used to solve phase problem in X-ray studies of nucleic acids [5,28-30].

From a chemical point of view, the substitution of carbonyl oxygen by sulphur or a selenium seems quite conservative, since all are elements with similar chemical properties. However, in the context of packed structures, such as substitution is expected to have a major effect, since both S and Se, can made stronger dispersion interactions, which can favor stacking contacts, but can also lead to steric clashes in the context of compact nucleic acid structures and are certainly poorer hydrogen bond acceptor than parent carbonyl group [16,17]. Obviously the, substitution of oxygen by bigger and less polar atoms, should also change the hydration free energy of the unpaired nucleobases impacting the equilibrium between unfolded and folded state of nucleic acids [16,17,28]. Finally, we cannot ignore that all the pattern of recognition of natural nucleic acids is based on the prevalence of keto/amino tautomeric forms in aqueous solution [31-34]. Whether or not such rules remain valid for derivatives containing –S or –Se is unclear [9,10,24,28].

We present here a comprehensive study of the properties of 6SeG, a quite unexplored derivative of guanine with many promising properties (see above). We studied its tautomeric preferences in the gas phase and solution and explored the impact of the incorporation of such molecule in the stability and properties of different nucleic acids: duplex, antiparallel triplexes and G-quadruplexes. Our results highlights 6SeG as a very interesting derivative with interesting nanotechnological properties and a quite unique capability to modulate the equilibrium between different forms of DNA, which opens interesting possibilities in anti-gene based therapies.

METHODS

Quantum mechanical calculations. Electronic structure calculations were carried out using the N9-methyl derivative of 6SeG as a reasonable model of the corresponding nucleoside. Geometry optimizations for all the potential tautomers of 6SeG (see Figure 1), were carried out at the MP2/6-31G(d,p) and B3LYP/cc-pVTZ levels of theory. Furthermore, single point energies were evaluated using basis sets of increasing complexity (up to cc-pVQZ) combined to get complete basis set (CBS) estimates by using Helgaker [35] and Truhlar [36] extrapolation schemes. The effect of higher order correlation effects on the tautomerization energy were obtained by comparing the MP2/6-31G(d,p) and CCSD(T)/6-31G(d,p) estimates, adding then the difference to the MP2/CBS estimates as discussed elsewhere [37,38]. Thermal and entropic correction to the tautomerization energies were computed at the MP2/6-31G(d,p) level of theory by using the corresponding vibrational harmonic analysis [10,39,40].

Solvation effects in the tautomerization process were introduced at the HF/6-31G(d,p) and B3LYP/cc-pVTZ levels within the IEF-MST continuum model [41-43]. Following our standard procedure [10,34] the free energy of tautomerization in water (ΔG_{aq}) is defined as:

$$\Delta G_{aq} = \Delta G_{gas} + \Delta\Delta G_{hyd} \qquad (1)$$

where ΔG_{gas} is the tautomerization free energy in the gas phase and $\Delta\Delta G_{hyd}$ is the difference in free energy of hydration of both tautomers computed from MST calculations (many of the values were also validated by MD/TI calculations; see below).

The H-bonding properties of the most interesting tautomeric forms of 6SeG were studied by analyzing their pairing with cytosine in the gas phase. Geometries were optimized at the BHandHLYP/cc-pVTZ level of theory, and the interaction energies were corrected by the basis set superposition errors (BSSE) using the counterpoise method [44]. Stacking interactions were calculated for the $(G{\cdot}C)_2$ dimer and the (G·C)(6SeG·C) dimer using average MD dimers as starting geometries and removing sugar and phosphate atoms. Stacking energies were computed as the difference between total interaction energy and hydrogen bond energy for a base pair step by fixing the relative orientation of nucleobases in the average geometries found in the MD trajectories and optimizing the intramolecular geometry at the M06-2X/6-31++G(d,p)

level of theory, a functional tuned to capture dispersion contributions [45]. The counterpoise method was used again to correct BSSE. Harmonic vibrational analysis was done to correct interaction energies by thermal, expansion and entropic corrections (using as reference state the ideal gas phase at 298 K and 1 atm).

Electronic properties of the optimized monomers, H-bonded dimers, triads and G-quartets were evaluated by calculating the HOMO, LUMO orbitals and the HOMO/LUMO gap using the M06-2X/6-31** level of theory. Estimations of the charge transfer rate were computed using the Marcus theory[46] from stacked systems based on our M06-2X/6-31++G(d,p)) calculations:

$$k_{ET} = \frac{2\pi}{\hbar} V_{da}^2 \frac{1}{\sqrt{4\pi\lambda\kappa_B T}} \exp\left(\frac{-(\Delta G + \lambda)^2}{4\lambda\kappa_B T}\right)$$

where $\hbar = h/2\pi$ (where h corresponds to the Planck's constant), V_{da} is the effective electronic coupling, λ represents the reorganization energy, ΔG is the driving force and $\kappa_B T$ is the Boltzmann factor at 298 K. Calculations were done assuming a reorganization energy value of 1 eV[47] and ΔG (G→G) = 0.

Molecular Dynamics simulations. Classical MD calculations were used to explore the impact of the G→6SeG modification in different canonical DNA structures: i) DNA duplex, ii) antiparallel triplex with the modified purine in either Watson Crick and Hoogsteen strands and iii) G-quadruplexes (with different number of 6SeG in the quartets). Model sequences used are shown in Table 1, where additional details on the simulated systems are also displayed.

System set-up: Original conformations for the duplex simulation were taken from Arnott's fiber data for B-DNA, following the *nab* module of the AmberTools 12 package {Case:2012vm}. Antiparallel triplex was created by modifying the nab script for parallel triplexes included in the same AmberTools 12 package. G-quadruplex starting structure was obtained from the NMR structure of the thrombin aptamer (PDB code 148D) previously pre-equilibrated in the group [48]. In all cases graphical programs were used to change the central guanine(s) by seleno-derivatives. All systems were neutralized by adding suitable amount of Na^+ and K^+ ions (to reduce equilibration problems care was taken to add K^+ in optimum positions of the central channel of G-quadruplex), and hydrated by adding from 3510 to 5467 TIP3P [49] water molecules defining simulation systems with at least 11 Å of solvent from the DNA to the nearest

face of the box. Solvated systems were then optimized, thermalized and pre-equilibrated using our standard procedure [50,51], and the final system were then equilibrated by 10 ns of unrestrained MD simulation, followed by 100 ns production runs.

Thermodynamic integration: Molecular dynamics thermodynamic integration (MD/TI) calculations were mainly performed to quantify the impact of G→SeG modification on the stability of different nucleic acids. As discussed elsewhere [17], this was done using a complex network of thermodynamic cycles, where the ΔG of interest (in horizontal axis in Figure 2) were obtained by combining the reversible work associated to the 6SeG→G mutation in the different nucleic acids (single stranded, duplex, triplex, quadruplex) structures. For example, the effect of 6SeG→G modification in duplex stability is determined as the difference in the reversible work associated to this mutation in a single stranded and a duplex DNA (see Figure 2). Starting structures were taken from the end of the unrestrained MD simulation (see above). Mutations were done three times using 21 and 42 double wide sampling windows for total simulation times of 420, 840 and 1680 ps each. Every window was divided in two halves, leading to a total of 12 independent estimates of the free energy change associated with the 6SeG→G change that where then combined to gain statistical quality in our results and to evaluate the potential impact of hysteresis effects [10,17].

Simulation details: All MD simulations were done in the isothermal (T=298K), isobaric (P= 1 atm) ensemble. Periodic boundary conditions and Particle Mesh Ewald were used to account for long-range electrostatic effects [52]. SHAKE [53] was used to maintain all the bonds at their equilibrium values, allowing the use of 2 fs time step for integration of Newton's equations of motion. The PARMBSC0 [54] modification of the parm99 force-field[55] was used to describe DNA interactions. Force-field parameters for 6SeG were obtained from different sources: i) SantaLucia's database [56] for bonded terms, ii) standard RESP procedure as implemented in RED [57] was used to fit point charges, and iii) van der Waals parameters were obtained by fitting classical and BHandHLYP dimerization geometries. To validate the quality of the derived force-field parameters we compared AMBER and BHandHLYP/ccpVTZ and M06-2X/6-31G++G(d,p) for H-bonding and stacking energies respectively energies finding excellent fitting (see Table 2) for all cases and almost perfect (r^2=0.98) correlations between QM and force-field

values. Note that total interaction energy data was not used at any point in refining the force-field parameters, which means that Table 2 represents an independent validation of the 6SeG force-field parameters.

Trajectory analysis: Helical analysis was carried out using Curves+ [58] program. Geometrical and energetic analysis was carried out using AmberTools 12 as well as a variety of in house programs including the NaFlex server (Hospital et al., Nucleic Acids Res. 2013 In Press). Classical molecular interaction potentials (CMIP [59]) were used to compute the differences in the interaction patterns associated to the different nucleic acids structures (normal or 6SeG containing) considered in this study.

RESULTS AND DISCUSSION

Tautomerism of 6-Selenoguanine. It has been suggested that the substitution of carbonyl oxygens by heavier atoms might introduce changes in the tautomeric preferences of the nucleobase, which might in turn modify its recognition properties [9]. Accordingly, we explored the intrinsic tautomeric preferences of 6-selenoguanine by means of very high-level of QM theory. Results (see Table 3) are quite converged with respect to the increase in the level of the calculation, and our best estimated (MP2/CBS corrected by higher order correlation at the CCSD(T) level) are expected to be exact within a few tenths of kcal/mol. The canonical "keto-like" tautomer (1H) is not the most dominant form in the gas phase, since the "enol-like" tautomers (6c and 6t) can be up to 6-7 kcal/mol more stable. The stability of the "enol-like" tautomer might appear counterintuitive, since it does not follow Watson-Crick recognition rules and disagree with previous results for guanine, which indicate that the keto tautomer was slightly preferred over the enol one [34,60]. However, present results for 6SeG agree with previous lower level calculations for this molecule [61], as well as with our previous analysis for 6-thioguanine. In fact, the relative stabilization of the "enol-like" form with respect to the canonical "keto-like" form (-7.2 kcal/mol, when the hydrogen at Se_6 is *trans* (see Figure 1) from our best estimates), agrees with the result found for the 6-thioguanine at the MP2/6-311++G(d,p) level of theory [62]. It is then clear that, the substitution of the oxygen carbonyl by a heavier atom, either sulphur or selenium, implies a significant stabilization of the "enol-like" tautomer.

The prevalence of the "enol-like" tautomer could suggest that the substitution of G by 6SeG might induce C→T transitions, but our previous experience with related systems [34,40,62] warns against this simplistic assumption, since water can change the *in vacuo* intrinsic tautomeric preferences of the nucleobases. SCRF (validated by MD/TI) results in Table 4 confirm our previous assumptions on the magnitude of the differential hydration free energy. Thus, the enol forms (6c and 6t) are destabilized by ca. 10 kcal/mol with respect to the canonical state. Such a specific solvation term is large enough to revert the intrinsic preferences of 6SeG, which in water should appear mainly in the canonical 1H tautomeric state (Table 4).

The interaction properties of 6-Seleno-Guanine. H-bonding calculations for the most stable tautomeric forms of 6SeG with cytosine in the gas phase (see Table 2) show that the "enol" tautomers are unlikely to play any role in stabilizing transitions in DNA structures due to its poorer H-bonding capabilities compared with the reference 1H tautomer. The higher stability of G·C base pair over 6SeG·C base pair (around 2 kcal/mol) shown in Table 2 agrees well with previous DFT results [63] and points towards a potential source of destabilization in DNA duplexes containing 6SeG. There is not however preference for G over 6SeG, when the purine is involved in reverse-Hoogsteen pairing (where the C=Se bond is not directly involved in hydrogen bonding). Finally, stacking energies calculated for two stacked base pairs, (G·C//G·C) and (G·C//6SeG·C) (Table 2), indicate that the presence of a 6-selenoguanine have only a marginal effect in stacking regarding the natural stacked tetramer.

The effect of 6SeG in the DNA duplex. MD simulation of a duplex DNA containing a G→6SeG mutation leads to a stable trajectory, which samples a B-like conformation very close to that of the parent duplex. The structural changes introduced by the presence of the selenium atom are very local and are evident (see Suppl. Table S2) in a small (around 0.1 Å) increase in rise, a moderate (2 degrees) decrease in roll and a significant enlargement of opening (6 degrees) and stretch (around 0.2 Å). These local changes can be rationalized as a consequence of the larger van der Waals sphere of the Se atom and produce some changes in the geometry of the grooves, the most important one: a reduction (around 0.4 Å) of the width of the minor groove, which leads to sizeable changes in the electrostatic profile around the mutation site (see Figure S1A). Besides this, the presence of the selenium atom lying in the major groove induces a notable widening (around 0.7 Å from its parent duplex) and a worse interacting region

with cationic probes due to the electronegativity reduction accounting for the oxygen substitution. Changes in the grooves produce local differences in the pattern of hydration around the mutation site, which is evident in the increase of the localized waters, which form a well defined spine of hydration along the minor groove that is broken by the guanine amino groups in the native duplex (Figure S1B).

We have assessed the influence of the selenium substitution in the stability of duplex DNA by means of MD/TI calculations (see *Methods*). All G↔6SeG mutations were smooth without discontinuities that could point to hysteresis artifacts and forward and reverse simulations yielded similar free energy estimations reinforcing our confidence in the results. Results (shown in Table 5) suggest that the presence of 6SeG strongly destabilizes the DNA duplex (around 5.2 kcal/mol). Although no experimental estimate of the destabilization free energy has been reported, to our knowledge, experimental studies by Salon et al. [30], confirms that the G→6SeG destabilizes very significantly the duplex (up to 11 degrees per modification) which points in the same direction as our quantitative theoretical estimations. Energy analysis of the trajectories (Table 6) reveals the weaker hydrogen bonding interactions involving 6SeG···C compared with the native G···C base pair (around 6 kcal/mol less stable for the 6SeG···C base pair interaction), in good agreement with our QM hydrogen-bonding calculations (Table 2). This destabilizing energy is not compensated by stacking (around 2 kcal/mol more stable based on our calculations), leading to an overall loss of stability of the DNA duplex, that might be tolerated only if the parent duplex is stable enough to provide a Tm> room temperature.

The effect of 6SeG in DNA triplexes. We analyzed two antiparallel triplexes, one containing 6SeG at the central Watson-Crick position (WC; see Figure 3) and the other with 6SeG placed at the reverse Hoogsteen (rH) position. Trajectories of both triplexes were stable, sampling standard antiparallel B-like triplexes (RMSd around 1.5-1.7 Å from the average structure, indistinguishable in terms of general geometrical characteristic to the parent one (containing only natural nucleobases). When 6SeG was placed both in the WC and in the reverse-Hoogsteen positions, and as a consequence of the steric repulsion of selenium atom, an increase in rise around 0.3 Å together with an enlargement of the distance between the bases of the corresponding triad of 0.2-0.3 Å was observed (see Table S3). Interestingly, the presence of the 6SeG in the reverse

Hoogsteen position does not influence the WC duplex regarding the helical parameters (Table S3), but indeed affects the helical properties of the reverse Hoogsteen duplex at the substitution site.

Local changes induced by the presence of 6SeG are reflected in subtle alterations in the widths of the three grooves,around the substitution site. As discussed elsewhere [64] binding of a triplex-forming oligonucleotide (TFO) into the major groove of a DNA duplex creates two new grooves: one which, despite of its bigger dimensions, does not show any characteristic chemical feature (MM groove) and the other which resembles the DNA duplex minor groove (mM groove) which can potentially interact with small molecules [64]. The presence of the 6-selenoguanine both in the WC and the rH positions makes the minor groove (m) notably narrower (around 1 Å) than in the non-modified triplex. The substitution of a central guanine both in the WC and in the reverse Hoogsteen positions widens the mM groove, especially for the last one for which the distance between phosphates of opposite strands shows a widening of this groove (around 0.7 Å regarding the canonical triplex structure;see Figure S2). From the computed cMIP profiles (Figure S2), it becomes also clear how the presence of 6SeG would improve the interaction region with small cationic molecules in the mM groove. As expected from the cMIP profiles, all triplexes are very well hydrated, with high density water regions both in the m- and mM- grooves. No dramatic differences are found related to the G→6SeG substitution, except a disruption in the spine of hydration in the mM groove around the 6SeG site when this is placed in the TFO strand, which can be thought to be related to the poorer hydrogen bonding properties from the C=Se bond.

Our MD/TI calculations strongly suggest that the presence of 6SeG at the WC position of the triplex destabilizes the structure by 6.6 kcal/mol (see Table 5) with respect to the single strand, which means that the duplex→triplex transition is disfavored by 1.4 kcal/mol upon G→6SeG in the WC strand. This indicates that the presence of 6SeG at the WC position instead of G not only destabilizes the WC binding with respect to the G·C pair, but also the triplex formation. Analysis of the interaction energies (see Table 6) shows the poorest hydrogen bonding specially for the WC duplex (around 6 kcal/mol compared to the parent G·C) due to the presence of the bulky and

less electronegative selenium group, which are involved in the interactions with both the WC and the Hoogsteen edges.

Quite interestingly, the presence of 6SeG at the rH position of the triplex destabilizes around 4.2 kcal/mol the triplex with respect to the single strand, which means that the thermodynamics of the relevant duplex→triplex transition is even slightly favored in the presence of 6SeG (see Table 5). The energy analysis reveals that when 6SeG is inserted in the rH position, the hydrogen bond interactions are only slightly destabilized (note in this case Se is not directly involved in interactions; see Figure 3). Therefore, the 6SeG substitution in the TFO strand shows a smaller destabilization regarding the WC positioning and even less than for the single→duplex transition since the WC duplex structure and its interactions remain almost unaffected.

The effect of 6SeG in DNA quadruplexes. We used the thrombin binding aptamer (TBA, PDB code: 148D) as a model of the DNA quadruplex, and introduced an increased number of G→6SeG changes in one of the tetrads. MD simulations show that only one of such substitutions is tolerated. Two or more lead to spontaneous unfolding of the quadruplex in the time length of the trajectory (see Figure 4), due to severe steric clashes in the interior channel of the quadruplex that lead to the eviction of the central K^+ cation and accordingly [65-67] to the disruption of the structure. A single substitution is tolerated, but the resulting aptamer shows significant alterations in the structure (see Figure 4) and loses stability (MD/TI suggest a stability penalty of 5.6 kcal/mol with respect to the single strand; see Table 5). Clearly (see Table 6) the strong reduction of the hydrogen bonding interactions and the eviction of the monovalent cation from the central channel are the main responsibles of the destabilization.

Biotechnological potential of 6SeG. While the medical use of 6SeG is well establish, the biotechnological possibilities of the derivative have not been yet explored. Certainly, our QM and MD/TI studies rule out any potential role of 6SeG as a nucleobase with multiple modes frames based on the coexistence of multiple tautomers [40], and in fact, 6SeG should have a similar affinity pattern than G. Present calculations show that the effect of 6SeG on the stability of the nucleic acid structure is dependent on the number and the arrangement of the strands, which opens the possibility to use 6SeG to displace the equilibrium between different structures. For example, our free

energy calculations (see Table 5) suggest that insertion of 6SeG in a poly-purine single strands (substituting a G) will affect very little the ability of the poly-purine single strand to act as (antiparallel) triplex forming oligonucleotide (TFO), while it will destabilize, or even disrupt intra-molecular quadruplex structures, which capture the TFO, acting as inhibitors for the formation of triplexes (see Figure 5). Calculations suggest then the substitution of G by 6SeG can be then an excellent approach to more effective TFOs.

The introduction of heavy atom derivatives of nucleobases might lead to modified DNAs with improved conductor properties [25], due to the reduction of the HOMO/LUMO gap. To explore whether or not 6SeG containing DNAs can display improved conduction properties, we analyzed the M06-2X/6-31++G(d,p) HOMO and LUMO energies of the isolated monomers, H-bonded dimers, triads and G-quartets and stacked complexes of G and 6SeG. Results in Table 7 show that the HOMO/LUMO gap for the 6SeG monomer is much smaller than the corresponding standard guanine with a reduction around 1.6 eV that results from the destabilization of the HOMO orbital upon seleno modification. A similar tendency is found when hydrogen bonded pairs are taken into account since the corresponding HOMO/LUMO gap for the H-bonded 6SeG·C dimer is slightly smaller (~0.2 eV) than for the corresponding G·C dimer. Both in the G·C dimer and the 6SeG·C dimer, the HOMO orbitals are localized in purine bases and the LUMO orbitals in the pyrimidine rings (Figure S3). The population analysis of the different H-bonded triads shows that when the 6SeG is in rH position the HOMO/LUMO gap is the smallest among the different triads (~0.4 eV respect to non-modified C·G#G triad) due to the destabilization of the HOMO level. The HOMO orbital is localized in the rH purine and the LUMO orbital in the cytosine (Figure S3). The presence of one 6SeG in a G-tetrad notably reduces the HOMO/LUMO gap around 1.2 eV regarding the native guanine quartet. Electron density plots show greater delocalization of HOMO and LUMO orbitals for the native G4 while both HOMO and LUMO orbitals are confined in the 6SeG nucleobase (Figure S3).

Regarding the stacked base pairs, we have considered three different fragments of one all-6SeG with two stacked 6SeG·C $(6SeG·C)_2$, one all-G with two natural stacked G·C pairs $(G·C)_2$ and, another system with a 6SeG·C base pair and a natural G·C base pair (G-C/6SeG-C). Starting geometries were obtained from average MD

geometries and refined later (see Methods). The molecular orbitals analysis shows a reduction around 0.3 eV of the HOMO/LUMO gap when having one or two stacked 6SeG·C base pairs with respect to the non-modified $(G{\cdot}C)_2$ (see Table 7). In summary, these results suggest that a reduction of the HOMO/LUMO gap when replacing G by 6SeG in the DNA duplex and second, a lower ionization potential for the selenoguanine (using the Koopmans' theorem), which would be more easily oxidized than the native guanine, which in fact has the lowest ionization potential among natural coding DNA bases. HOMO energies for dimer and stacked base pairs, suggest that DNA containing 6SeG could become a better hole trap during charge transfer. Indeed, natural nucleobases are known to undergo fast hole transfer reactions preventing from hole migration within stacked genomic DNA. Therefore, we have calculated the relative charge transfer rate by means of Marcus theory obtaining a k_{ET} (G→G)/k_{ET}(G→6SeG) ratio around 1.7. It is known that the absolute charge transfer rate between two adjacent guanines[68] is ~4.3 ns^{-1} thus, the estimation of the absolute rate for the G→6SeG charge transfer would be around 2.6 ns^{-1} still in the range of the fastest natural hole transfer reactions in DNA (the A → A absolute charge transfer rate is around 1.2 ns^{-1}[68]).

Overall these results suggest that 6SeG impacts the HOMO/LUMO gap and accordingly the transfer properties of DNA, providing interesting possibilities for the design of DNA systems with improved conductimetric properties. Although our kinetic estimations suggest that G→6SeG hole transfer rate would be slower than G→G transfer, electron transfer would be still remarkable.

ACKNOWLEDGMENTS

This work has been supported by grant BIO2012-32868 (MO) and the Consolider E-Science project from MINECO-Spain, the European Research Council (ERC-Advanced Grant, MO), the Instituto Nacional de Bioinformática (INB; MO), and the Fundación Marcelino Botin (MO). MO is an ICREA Academia Researcher.

Bibliography

(1) Benner, S. A. *Trends in Biotechnology* **1994**, *12*, 158–163.
(2) Matsuda, S.; Henry, A. A.; Romesberg, F. E. *J Am Chem Soc* **2006**, *128*, 6369–6375.
(3) Leconte, A. M.; Hwang, G. T.; Matsuda, S.; Hari, Y.; Romesberg, F. E. *J Am*

Chem Soc **2008**, *130*, 2336–2343.
(4) Bohon, J.; de los Santos, C. R. *Nucleic Acids Research* **2003**, *31*, 1331–1338.
(5) Jiang, J.; Sheng, J.; Carrasco, N.; Huang, Z. *Nucleic Acids Research* **2007**, *35*, 477–485.
(6) Liang, J.; Wang, Z.; He, X.; Li, J.; Zhou, X.; Deng, Z. *Nucleic Acids Research* **2007**, *35*, 2944–2954.
(7) Connolly, B. A.; Newman, P. C. *Nucleic Acids Research* **1989**, *17*, 4957–4974.
(8) Rao, T. V.; Haber, M. T.; Sayer, J. M.; Jerina, D. M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 907–910.
(9) Sintim, H. O.; Kool, E. T. *J Am Chem Soc* **2006**, *128*, 396–397.
(10) Faustino, I.; Aviño, A.; Marchán, I.; Luque, F. J.; Eritja, R.; Orozco, M. *J Am Chem Soc* **2009**, *131*, 12845–12853.
(11) Coleman, R. S.; Siedlecki, J. M. *J Am Chem Soc* **1992**, *114*, 9229–9230.
(12) Jing, Y.; Kao, J. F.; Taylor, J. S. *Nucleic Acids Research* **1998**, *26*, 3845–3853.
(13) Massey, A.; Xu, Y. Z.; Karran, P. *Current Biology* **2001**, *11*, 1142–1146.
(14) Marathias, V. M.; Jerkovic, B.; Bolton, P. H. *Nucleic Acids Research* **1999**, *27*, 1854–1858.
(15) Lahoud, G.; Timoshchuk, V.; Lebedev, A.; Arar, K.; Hou, Y. M.; Gamper, H. *Nucleic Acids Research* **2008**, *36*, 6999–7008.
(16) Somerville, L.; Krynetski, E. Y.; Krynetskaia, N. F.; Beger, R. D.; Zhang, W.; Marhefka, C. A.; Evans, W. E.; Kriwacki, R. W. *Journal of Biological Chemistry* **2003**, *278*, 1005–1011.
(17) Špačková, N.; Cubero, E.; Sponer, J.; Orozco, M. *J Am Chem Soc* **2004**, *126*, 14642–14650.
(18) Bohon, J.; de los Santos, C. R. *Nucleic Acids Research* **2005**, *33*, 2880–2886.
(19) Glaab, W. E.; Risinger, J. I.; Umar, A.; Barrett, J. C.; Kunkel, T. A.; Tindall, K. R. *Carcinogenesis* **1998**, *19*, 1931–1937.
(20) Krynetskaia, N. F.; Krynetski, E. Y.; Evans, W. E. *Mol. Pharmacol.* **1999**, *56*, 841–848.
(21) Presta, M.; Belleri, M.; Vacca, A.; Ribatti, D. *Leukemia* **2002**, *16*, 1490–1499.
(22) Waters, T. R.; Swann, P. F. *Biochemistry* **1997**, *36*, 2501–2506.
(23) Karran, P.; Bignami, M. *BioEssays* **1994**, *16*, 833–839.
(24) Salon, J.; Sheng, J.; Jiang, J.; Chen, G.; Caton-Williams, J.; Huang, Z. *J Am Chem Soc* **2007**, *129*, 4862–4863.
(25) Vázquez-Mayagoitia, A.; Huertas, O.; Brancolini, G.; Migliore, A.; Sumpter, B. G.; Orozco, M.; Luque, F. J.; Di Felice, R.; Fuentes-Cabrera, M. *J Phys Chem B* **2009**, *113*, 14465–14472.
(26) Hassan, A. E. A.; Sheng, J.; Zhang, W.; Huang, Z. *J Am Chem Soc* **2010**, *132*, 2120–2121.
(27) Sun, H.; Sheng, J.; Hassan, A. E. A.; Jiang, S.; Gan, J.; Huang, Z. *Nucleic Acids Research* **2012**.
(28) Lin, L.; Sheng, J.; Momin, R. K.; Du, Q.; Huang, Z. *Nucleosides Nucleotides Nucleic Acids* **2009**, *28*, 56–66.
(29) Melvin, J. B.; Haight, T. H.; Leduc, E. H. *Cancer Res.* **1984**, *44*, 2794–2798.
(30) Salon, J.; Jiang, J.; Sheng, J.; Gerlits, O. O.; Huang, Z. *Nucleic Acids Research* **2008**, *36*, 7009–7018.
(31) Alhambra, C.; Luque, F. J.; Portugal, J.; Orozco, M. *Eur. J. Biochem.* **1995**, *230*, 555–566.
(32) Watson, J. D.; Crick, F. H. *Nature* **1953**, *171*, 737–738.
(33) Topal, M. D.; Fresco, J. R. *Nature* **1976**, *263*, 289–293.

(34) Colominas, C.; Luque, F. J.; Orozco, M. *J Am Chem Soc* **1996**, *118*, 6811–6821.
(35) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J Chem Phys* **1997**, *106*, 9639–9646.
(36) Truhlar, D. G. *Chemical Physics Letters* **1998**, *294*, 45–48.
(37) Jurečka, P.; Hobza, P. *J Am Chem Soc* **2003**, *125*, 15608–15613.
(38) Sponer, J.; Jurečka, P.; Hobza, P. *J Am Chem Soc* **2004**, *126*, 10142–10151.
(39) Cieplak, P.; Bash, P.; Singh, U. C.; Kollman, P. A. *J Am Chem Soc* **1987**, *109*, 6283–6289.
(40) Blas, J. R.; Luque, F. J.; Orozco, M. *J Am Chem Soc* **2004**, *126*, 154–164.
(41) Tomasi, J.; Mennucci, B.; Cances, E. *Journal of Molecular Structure: THEOCHEM* **1999**, *464*, 211–226.
(42) Orozco, M.; Luque, F. J. *Chem Rev* **2000**, *100*, 4187–4226.
(43) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. *Journal of Molecular Structure: THEOCHEM* **2005**, *727*, 29–40.
(44) Boys, S. F.; Bernardi, F. *Molecular Physics* **1970**, *19*, 553–566.
(45) Zhao, Y.; Truhlar, D. G. *Theor Chem Account* **2007**, *120*, 215–241.
(46) Marcus, R. A. *J Chem Phys* **1956**, *24*, 966.
(47) Senthilkumar, K.; Grozema, F. C.; Guerra, C. F.; Bickelhaupt, F. M.; Lewis, F. D.; Berlin, Y. A.; Ratner, M. A.; Siebbeles, L. D. A. *J Am Chem Soc* **2005**, *127*, 14894–14903.
(48) Saneyoshi, H.; Mazzini, S.; Aviño, A.; Portella, G.; González, C.; Orozco, M.; Marquez, V. E.; Eritja, R. *Nucleic Acids Research* **2009**, *37*, 5589–5601.
(49) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. *Comparison of simple potential functions for simulating liquid water*; The Journal of chemical …, 1983.
(50) Shields, G. C.; Laughton, C. A.; Orozco, M. *J Am Chem Soc* **1997**, *119*, 7463–7469.
(51) Soliva, R.; Laughton, C. A.; Luque, F. J.; Orozco, M. *J Am Chem Soc* **1998**, *120*, 11226–11233.
(52) Darden, T. A.; York, D.; Pedersen, L. *J Chem Phys* **1993**, *98*, 10089.
(53) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J Comp Phys* **1977**, *23*, 327–341.
(54) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys J* **2007**, *92*, 3817–3829.
(55) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J Am Chem Soc* **1995**, *117*, 5179–5197.
(56) Aduri, R.; Psciuk, B.; Saro, P.; Taniga, H. *J Chem Theory Comput* **2007**, *3*, 1464–1475.
(57) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821–7839.
(58) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. *Nucleic Acids Research* **2009**, *37*, 5917–5929.
(59) Gelpí, J. L.; Kalko, S. G.; Barril, X.; Cirera, J.; la Cruz, de, X.; Luque, F. J.; Orozco, M. *Proteins* **2001**, *45*, 428–437.
(60) Plekan, O.; Feyer, V.; Richter, R.; Coreno, M.; Vall-Llosera, G.; Prince, K. C.; Trofimov, A. B.; Zaytseva, I. L.; Moskovskaya, T. E.; Gromov, E. V.; Schirmer, J. *J. Phys. Chem. A* **2009**, *113*, 9376–9385.

(61) Venkateswarlu, D.; Leszczynski, J. *J. Phys. Chem. A* **1998**, *102*, 6161–6166.
(62) Alhambra, C.; Luque, F. J.; Estelrich, J.; Orozco, M. *J. Org. Chem.* **1995**, *60*, 969–976.
(63) Wang, J.; Gu, J.; Leszczynski, J. *J Comput Chem* **2012**, *33*, 1587–1593.
(64) Shields, G. C.; Laughton, C. A.; Orozco, M. *J Am Chem Soc* **1998**, *120*, 5895–5904.
(65) Borzo, M.; Detellier, C.; Laszlo, P.; Paris, A. *J Am Chem Soc* **1980**, *102*, 1124–1134.
(66) Špačková, N.; Berger, I.; Sponer, J. *J Am Chem Soc* **1999**, *121*, 5519–5534.
(67) Gu, J.; Leszczynski, J.; Bansal, M. *Chemical Physics Letters* **1999**, *311*, 209–214.
(68) Conron, S. M. M.; Thazhathveetil, A. K.; Wasielewski, M. R.; Burin, A. L.; Lewis, F. D. *J Am Chem Soc* **2010**, *132*, 14388–14390.

4.1.2.1 Results and discussion

We explore the chemical properties of 6-selenoguanine (6SeG) and the influence in duplex, triplex and G-quadruplex structures by means of high-level quantum mechanics calculations and free energy calculations combined with molecular dynamics simulations. 6-Selenoguanine, like other related antimetabolites like thiopurines, has been longtime used against several cancers like lymphomas or hepatomas (Lin et al. 2009). Moreover, their applications have been extended to solving X-ray nucleic acids structures (6SeG can be incorporated to these molecules and display anomalous dispersion (Salon et al. 2008)) and, from the present work we suggest that 6SeG can also exhibit remarkable conductor properties in DNA structures.

Tautomeric properties of 6-selenoguanine Accurate gas phase and solvation calculations were performed to evaluate the tautomeric preferences of 6SeG. From gas phase calculations, the selone tautomer is not the most stable since selenol tautomers (6c and 6t) are 6-7 kcal/mol more stable. However, MST/SCRF hydration calculations show the selenol forms are destabilized by ca. 10 kcal/mol regarding the selone tautomer, suggesting that canonical tautomers are expected to be the dominant species in aqueous solution.

6-Selenoguanine in DNA structures Molecular dynamics simulations were used to evaluate the influence of 6SeG in DNA duplex, antiparallel triplex and G-quadruplex structures. Our QM calculations on monomer and H-bonded dimers suggest that selenol tautomeric forms are unlikely to have significant influence in the DNA environment. In fact, DNA duplex structures incorporating 6SeG yield stable trajectories and only very local deviations from canonical B-form values are observed. Energy analysis of MD trajectories reveals the weaker hydrogen bonding interactions in 6SeG compared with the G·C base pairs.

The influence of 6-Selenoguanine in antiparallel triplexes was studied, when modifications were inserted either at the central Watson-Crick position (WC) or at the TFO strand (rH). The analysis of the interaction energies around the 6SeG residue shows that, for the WC modification, the less electronegative selenium atom destabilizes base-base interactions through both the WC and the Hoogsteen edges while for the rH modification, the 6SeG substitution in the TFO strand shows only small destabilization since selenium atom is not directly involved in hydrogen bonds but its presence produces a small local distortion.

Finally, we explored the effect of the substitution of G by 6SeG in G-quadruplexes

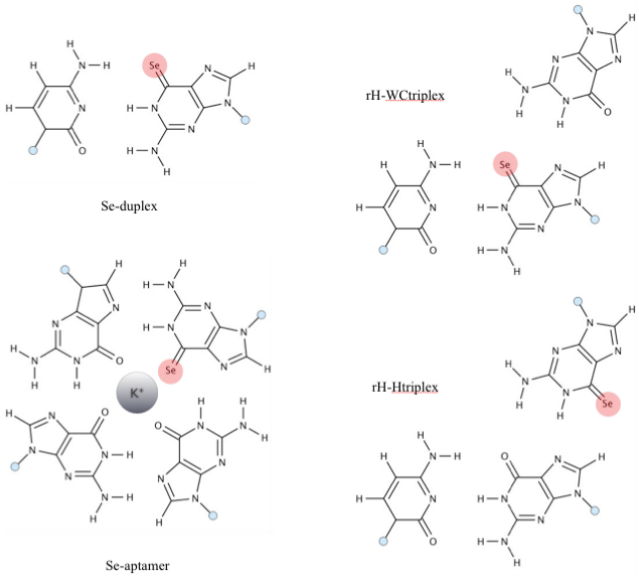


Figure 4.2: Interactions of 6SeG in DNA base pair duplex, triplex triads and G-tetrad.

using the thrombin binding aptamer (TBA; PDB code: 148D) as a model of G-quadruplex. Our MD simulations clearly show that only one substitution is tolerated while two or more generate an structural opening after few nanoseconds of simulation. According to the energy analysis, the major source of destabilization come from the reduction of hydrogen bonding interactions and from the eviction of the K^+ atom from the central channel.

	duplex	Se-duplex	triplex	WC-Se-triplex	rH-Se-triplex	aptamer	Se-aptamer
H-bonding	-27.3 ± 1.7	-21.5 ± 1.6	-45.1 ± 1.5	-39.4 ± 1.5	-42.9 ± 1.6	-98.1 ± 1.8	-80.1 ± 1.5
			(-27.4, -17.7)	(-22.6, -16.8)	(-27.2, -15.7)		
Stacking	-24.0 ± 1.0	-25.9 ± 1.0	-49.0 ± 1.1	-47.7 ± 1.0	-47.1 ± 1.2	-19.2 ± 1.5	-17.2 ± 1.3
K^+-tetrads						-147.8 ± 1.8	-150.5 ± 1.7
Total	-51.3 ± 1.2	-47.4 ± 1.1	-94.1 ± 1.1	-87.1 ± 1.1	-90.0 ± 1.2	-265.1 ± 1.3	-247.8 ± 1.2

Table 4.4: Interaction energies (H-bond, stacking) and standard deviations for duplex, triplex and aptamer structures around the mutation site and K^+-nucleobases interaction energies for the two tetrads of the aptamer and the modified aptamer with single mutation. Energies are in kcal/mol. H-bonding energies correspond to base pairs at the mutation site while stacking energies correspond to the central three base pairs around the mutation site. Values in parentheses correspond to Watson-Crick and reverse-Hoogsteen hydrogen bond interactions.

The free energy values associated to the G→6SeG mutation were calculated for the different DNA structures. Our MD/TI calculations suggest: i) 6SeG insertion destabilizes with respect to the native guanine due to the weak hydrogen bonding interactions and, ii) 6SeG substitution in a G-rich oligo will form an antiparallel triplex with a matching sequence instead of forming a G-quadruplex structure, a fact that

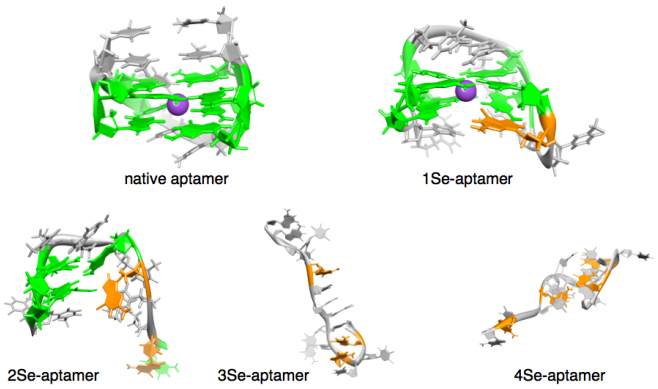


Figure 4.3: Final structures of MD simulations of the wild-type and progressive number of 6SeG substitutions in the TBA molecule. For clarity 6SeG residues are depicted in orange, G's in green and K^+ ion in purple.

usually prevents from the formation of stable triplexes, increasing the possibilities of oligos in anti-gene based therapies.

	ΔΔG (kcal/mol)
Se-duplex	5.2 ± 0.9
WC-Se-triplex	6.6 ± 2.0
rH-Se-triplex	4.2 ± 2.8
Se-aptamer	5.6 ± 1.0

Table 4.5: Relative free energy values (in kcal/mol) for the associated G→6SeG mutation.

Biotechnological properties The potential utility of heavy nucleobase derivatives in DNA in order to improve conductivity have been suggested in previous studies (Vazquez-Mayagoitia et al. 2009), which suggested that the reduction of the HOMO/LUMO gap has been related to a better electronic transfer between nucleobases specially important in the case of purine bases. QM calculations show that the 6SeG monomer has a lower gap with respect to their native counterpart (∼1.6 eV). A similar tendency can be observed when calculating the HOMO/LUMO gap for the hydrogen bonded 6SeG·C.

HOMO energies in DNA nucleobases are also related with hole transfer reactions. Indeed, natural nucleobases are known to undergo fast natural transfer reactions preventing from hole migration with stacked genomic DNA. Relative charge transfer rate was calculated with the Marcus theory for G→G and G→6SeG systems. The k_{ET}(G→G) / k_{ET}(G→6SeG) ratio was computed to be around 1.7. The estimation of the absolute rate between two stacked guanines is ∼4.3 ns^{-1} (Conron et al. 2010) and therefore, the

absolute rate for the G//6SeG stacked system would be around 2.6 ns^{-1} which it will be faster than the A→A charge transfer (Conron et al. 2010), providing interesting possibilities for the design of DNA systems with improved conductimetric properties.

		HOMO	LUMO	gap
	G	-5.00	-0.94	4.05
	6SeG	-4.26	-1.84	2.42
H-bonded	G·C	-4.34	-1.95	2.39
	6SeG·C	-4.15	-1.98	2.17
	C·G#G	-4.00	-2.31	1.69
	C·6SeG#G	-4.11	-2.30	1.81
	C·G#6SeG	-3.58	-2.34	1.24
	G4	-4.55	-0.85	3.70
	G3·6SeG	-4.08	-1.61	2.47
Stacked	$(G·C)_2$	-4.44	-2.53	1.91
	(G·C/6SeG·C)	-3.96	-2.35	1.61
	$(6SeG·C)_2$	-4.11	-2.46	1.65

Table 4.6: Energies of HOMO and LUMO orbitals, and HOMO/LUMO gap for isolated bases and base pairs calculated with SVWN5/6-31++G(d,p) (in eV).

4.1.2.2 Conclusions

We have explored the structural and recognition properties of the 6-selenoguanine in the context of different DNA structures. The most stable selone tautomeric form is expected to be the preferred both in hydrated systems and in DNA environments. The substitution of G by 6SeG is reasonably tolerated when inserted in DNA duplex, triplex and even with single mutation in G-quadruplex structures. The free energy calculations show that the weaker hydrogen bonding interactions related to the presence of 6SeG alter the thermodynamic stability of all studied systems. Indeed, the lowest destabilizing effect is found when the mutation is placed in a G-rich triplex forming oligonucleotide (TFO) strand. Therefore, a G-rich single strand with one 6SeG in its sequence will favor the formation of the triplex structure (ideally with a perfect match with the corresponding DNA duplex) instead of forming a G-quadruplex structure. The weak interaction of selone groups in G-tetrads facilitate the eviction of the central K^+ ion, which is shown to play an essential role in the G-quadruplex stability.

Finally, the introduction of selenium in the guanine nucleobase reduces the HOMO /LUMO gap of monomer and hydrogen bonded dimers suggesting that the insertion of 6SeG in DNA duplex structures may improve the conductimetric properties of DNA structures.

4.1.2.3 References

Conron, S.M.M. et al., 2010. Direct measurement of the dynamics of hole hopping in extended DNA G-Tracts. An unbiased random walk. Journal of the American Chemical Society, 132(41), pp.14388–14390.

Lin, L. et al., 2009. Facile Synthesis and Anti-Tumor Cell Activity of Se-Containing Nucleosides. Nucleosides, nucleotides & nucleic acids, 28(1), pp.56–66.

Salon, J. et al., 2008. Derivatization of DNAs with selenium at 6-position of guanine for function and crystal structure studies. Nucleic Acids Research, 36(22), pp.7009–7018.

Vazquez-Mayagoitia, A. et al., 2009. Ab initio Study of the Structural, Tautomeric, Pairing, and Electronic Properties of Seleno-Derivatives of Thymine. The Journal of Physical Chemistry B, 113(43), pp.14465–14472.

4.1.3 Functionalization of the 3'-Ends of DNA and RNA Strands with N-ethyl-N-coupled Nucleosides: A Promising Approach To Avoid 3'-exonuclease-Catalyzed Hydrolysis of Therapeutic Oligonucleotides.

Montserrat Terrazas, Adele Alagia, **Ignacio Faustino**, Modesto Orozco and Ramón Eritja.

ChemBioChem. 2013, 14(4), pp.510–520.

ChemPubSoc Europe

DOI: 10.1002/cbic.201200611

Functionalization of the 3′-Ends of DNA and RNA Strands with N-ethyl-N-coupled Nucleosides: A Promising Approach To Avoid 3′-Exonuclease-Catalyzed Hydrolysis of Therapeutic Oligonucleotides

Montserrat Terrazas,*[a] Adele Alagia,[a] Ignacio Faustino,[b, c] Modesto Orozco,[b, c] and Ramon Eritja*[a]

The development of nucleic acid derivatives to generate novel medical treatments has become increasingly popular, but the high vulnerability of oligonucleotides to nucleases limits their practical use. We explored the possibility of increasing the stability against 3′-exonucleases by replacing the two 3′-terminal nucleotides by N-ethyl-N-coupled nucleosides. Molecular dynamics simulations of 3′-N-ethyl-N-modified DNA:Klenow fragment complexes suggested that this kind of alteration has negative effects on the correct positioning of the adjacent scissile phosphodiester bond at the active site of the enzyme, and accordingly was expected to protect the oligonucleotide from degradation. We verified that these modifications conferred complete resistance to 3′-exonucleases. Furthermore, cellular RNAi experiments with 3′-N-ethyl-N-modified siRNAs showed that these modifications were compatible with the RNAi machinery. Overall, our experimental and theoretical studies strongly suggest that these modified oligonucleotides could be valuable for therapeutic applications.

Introduction

Over the past three decades, the inhibition of gene expression by synthetic oligonucleotides has been a widely explored field.[1] Relevant established classes of oligonucleotide agents include antisense oligonucleotides, short interfering RNAs (siRNAs),[2] aptamers,[3] and DNA/RNAzymes.[4] Unfortunately, the application of oligonucleotides as therapeutic agents in vivo faces some key problems.[5] One of the most important ones is that ordinary DNA and RNA are highly vulnerable to serum nucleases; this leads to short half-lives of the oligonucleotides in serum.

Much research effort has been focused on overcoming these limitations.[2,5,6] In particular, major efforts have been made to increase the biostability of oligonucleotide-based agents without compromising their biological activity.[5] This research has crystallized in the synthesis of a wide variety of modified oligonucleotides that contain chemical modifications involving the sugar ring[7] and/or the phosphate backbone.[7a,g,8] Among them, the phosphorothioate modification[9] provides high levels of nuclease protection and has been widely and successfully employed.[5] In contrast to the large effort made to create modified backbones with improved biostability, there has been little exploration of the potential impact of nucleobase modifications.[10]

An alternative solution is the incorporation of modified nucleotides in the 3′-dinucleotide overhangs of siRNAs.[11] Incorporation of two peptide nucleic acid (PNA) units has been found to improve the biostability of these oligonucleotides without impairing the siRNA activity.[11a] Positive results have also been obtained for carbohydrate conjugates,[11b] for siRNAs bearing C-5 polyamine-substituted nucleosides,[11c] and for siRNAs containing terminal bis(hydroxymethyl)benzene[11d] and biaryl units.[11e] Many of the studied modifications increased the biostability of oligonucleotides, but in some cases modified oligonucleotides were found to have negative effects on activity.[12] Thus, the search for efficient oligonucleotide chemistries remains a focus of continued study.

We have created and analyzed a new class of modification aimed at increasing the stability of oligonucleotides against 3′-exonucleases (the predominant nuclease activity present in serum[13]) without affecting biological action. In particular, rational design showed the possibility of blocking the hydrolytic activity of 3′-exonucleases by creating a new nucleotide scaffold characterized by its lack of a phosphodiester bond linking the two 3′-terminal nucleotide building blocks. Our approach is based on the replacement of the two 3′-terminal nucleotides

[a] Dr. M. Terrazas, A. Alagia, Prof. Dr. R. Eritja
Institute for Research in Biomedicine (IRB Barcelona) and
Institute for Advanced Chemistry of Catalonia (IQAC)
Spanish Research Council (CSIC)
Cluster Building, Baldiri i Reixac 10, 08028 Barcelona (Spain)
E-mail: montserrat.terrazas@irbbarcelona.org
recgma@cid.csic.es

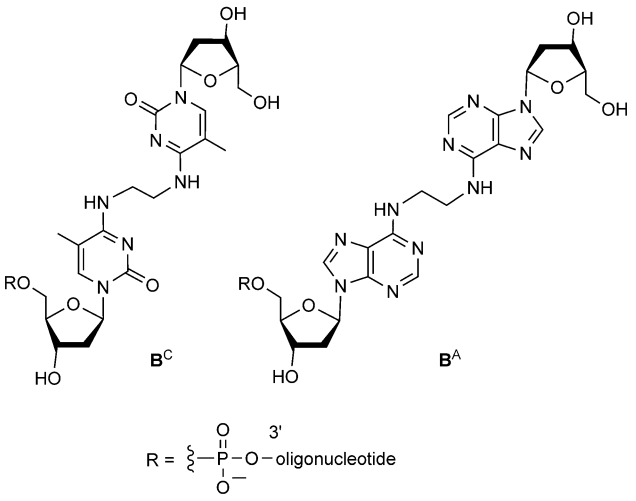
[b] I. Faustino, Prof. Dr. M. Orozco
Joint IRB–BSC Program on Computational Biology
Institute for Research in Biomedicine (IRB Barcelona)
Baldiri i Reixac 10, 08028 Barcelona (Spain)

[c] I. Faustino, Prof. Dr. M. Orozco
Department of Biochemistry, University of Barcelona
Diagonal 647, 08028 Barcelona (Spain)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cbic.201200611.

WILEY ONLINE LIBRARY © 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

of a natural oligonucleotide strand (linked through a 3′–5′ phosphodiester bond) by two nucleoside units linked together by an alkyl chain through the exocyclic amino group of the nucleobase. The resulting dimeric nucleosides (N^4-ethyl-N^4 2′-deoxy-5-methylcytidine derivatives (B^C) and N^6-ethyl-N^6 2′-deoxyadenosine derivatives (B^A)) are connected to the oligonucleotide strand through a normal 3′–5′ phosphodiester bond. Molecular dynamics (MD) simulations of a 3′-B^C-substituted DNA strand in complex with the Klenow fragment of *Escherichia coli* DNA polymerase I predicted strong resistance to 3′-exonuclease-catalyzed hydrolysis due to steric clashes between the ethyl linker and amino acid residues at the active site of the enzyme.



In agreement with the results of our calculations, functionalization of the 3′-ends of DNA and RNA strands with B^C and B^A modifications completely blocked the hydrolytic activity of 3′-exonucleases. Moreover, comparative studies involving 3′-B^C-modified oligonucleotides and their phosphorothioate (PS)-modified versions revealed that the N-ethyl-N modification confers higher 3′-exonuclease protection than phosphorothioate bonds. Finally, RNA interference (RNAi) experiments with B^C- and B^A-modified siRNAs targeting a luciferase gene and an antiapoptotic gene demonstrated that this class of alteration is compatible with the RNAi machinery.

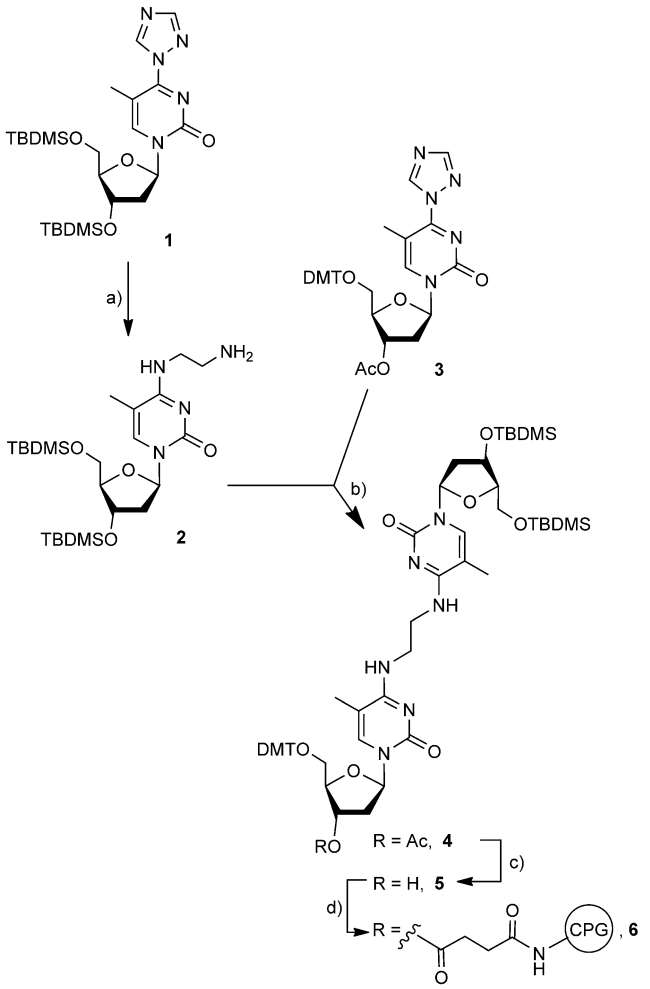
Results and Discussion

Synthesis of coupled *N*-ethyl-*N* dimeric pyrimidine and purine nucleosides

In order to incorporate N-ethyl-N-coupled nucleosides at the 3′-end of oligonucleotide strands, we protected the 5′-OH of one of the units of the dimer with a dimethoxytrityl (DMT) group and the 5′- and 3′-hydroxy groups of the second building block with a *tert*-butyldimethylsilyl (TBDMS) group (see compounds **5** and **10**). This leaves the 3′-OH of the first nucleoside unit free to participate in the functionalization of the solid support for oligonucleotide synthesis (via a succinate derivative).

The N^4-ethyl-N^4 pyrimidine nucleotide building blocks (B^C) were synthesized by following a similar method to the one described previously for the preparation of phosphoramidites of asymmetric coupled nucleosides used in DNA crosslinks, which involves an O^4-triazolyl intermediate.[14]

Thus, the 3′,5′-di-*O*-TBDMS-protected O^4-triazolyl nucleoside **1** (Scheme S1) was converted to the aminoethyl derivative **2** by treatment with ethylenediamine (Scheme 1). Treatment of amino nucleoside **2** with the 3′-*O*-Ac-5′-*O*-DMT-protected O^4-triazolyl nucleoside **3** gave the dimeric derivative **4** (Scheme 1), which was deacetylated to **5** with methanolic ammonia. To enable attachment to the solid support, **5** was first treated with succinic anhydride. The resulting succinate derivative was then linked to long-chain aminoalkyl controlled-pore glass (LCAA-CPG) to create the solid support **6** linked to **5**. The succinate moiety was coupled to the free amino groups on the CPG by using 2,2′-dithio-bis-(5-nitropyridine), 4-dimethylaminopyri-

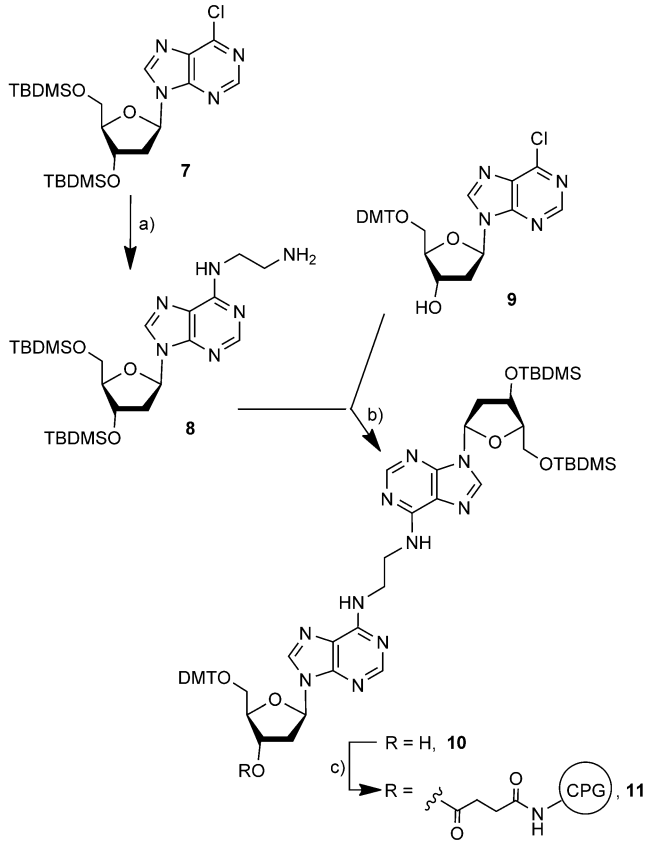


Scheme 1. Synthesis of coupled N^4-ethyl-N^4 dimeric pyrimidine nucleosides. a) ethylenediamine, pyridine, RT, 89%; b) **3**, Et_3N, pyridine, RT, 92%; c) NH_3/MeOH, RT, 86%; d) i: succinic anhydride, iPr_2NEt, 4-dimethylaminopyridine (DMAP), CH_2Cl_2, ii: LCAA-CPG, PPh_3, DMAP, 2,2′-dithio-bis-(5-nitropyridine), CH_2ClCH_2Cl/CH_3CN.

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

dine and triphenylphosphine[15] to generate an amide bond between the deoxynucleoside and the CPG. The loading of the 3′,5′-di-*O*-TBDMS-5′-*O*-DMT-protected dimer on CPG was quantified by following acid-catalyzed detritylation at 498 nm by using a UV–visible spectrophotometer[16] (the resulting CPG solid support had a loading of 29.5 μmol g^{-1}).

We next prepared the N^6-ethyl-N^6 purine dimers (B^A) conveniently protected for the functionalization of the solid support (Scheme 2). The synthesis of purine dimers derived from adenosine has already been described.[17] However, we developed a more convenient synthesis of asymmetric 2′-deoxyadenosine dimers. 3′,5′-Di-*O*-TBDMS-protected 6-chloropurine nucleoside **7** (Scheme S2) was converted to aminoethyl nucleoside **8** by following a similar procedure to the one employed for the preparation of aminoethyl nucleoside **2** (Scheme 2). Subsequently, we prepared the dimer by treating compound **8** with the 5′-*O*-DMT-protected 6-chloronucleoside **9**. When we performed the reaction in the presence of triethylamine in pyridine at room temperature, dimeric nucleoside **10** was obtained in low yield (18%). However, replacement of pyridine by acetonitrile/dichloromethane (1:1) led to a significant increase in the yield of the dimer (54%). CPG was functionalized with **10** according to the same procedure as used for the preparation of B^C-functionalized CPG (the resulting CPG solid support had a loading of 28.6 μmol g^{-1}).



Scheme 2. Synthesis of coupled N^6-ethyl-N^6 dimeric purine nucleosides. a) ethylenediamine, pyridine, RT, 99%; b) **9**, Et_3N, CH_3CN/CH_2Cl_2 (1:1), RT, 3 days, 54%; c) i: succinic anhydride, iPr_2NEt, DMAP, CH_2Cl_2, ii: LCAA-CPG, PPh_3, DMAP, 2,2′-dithio-bis-(5-nitropyridine), CH_2ClCH_2Cl/CH_3CN.

Synthesis of DNA and siRNA sense and guide strands carrying dimeric species at the 3′-ends

We next conjugated the dimeric nucleotide building blocks to the 3′-ends of oligonucleotides. We synthesized a 18-nt DNA strand containing no modifications (**12**), its 3′-B^C-modified and 3′-B^A-modified versions (**13** and **14**, respectively) and its DNA complement (**15**; Table 1).

Table 1. Sequences of synthesized oligonucleotides.

ON	Sequence
12	3′-AGGTCTTGTTTCCTTTGC-5′
13	3′-**B**^{C}AGGTCTTGTTTCCTTTGC-5′[a]
14	3′-**B**^{A}AGGTCTTGTTTCCTTTGC-5′[a]
15	3′-GCAAAGGAAACAAGACCT-5′
16	3′-TTAAAAAGAGGAAGAAGUCUA-5′
17	3′-**B**^{C}AAAAAGAGGAAGAAGUCUA-5′[a]
18	3′-**B**^{A}AAAAAGAGGAAGAAGUCUA-5′[a]
19	5′-UUUUUCUCCUUCUUCAGAUTT-3′
20	5′-UUUUUCUCCUUCUUCAGAU**B**^{C}-3′[a]
21	5′-UUUUUCUCCUUCUUCAGAU**B**^{A}-3′[a]
22	3′-TTAGGAAAGAAAGAAAGCUAU-5′
23	5′-UCCUUUCUUUCUUUCGAUATT-3′
24	3′-TTGAAGUAGUGAUAGAGGGCC-5′
25	3′-**B**^{C}GAAGUAGUGAUAGAGGGCC-5′[a]
26	5′-CUUCAUCACUAUCUCCCGGTT-3′
27	3′-TTUGCACUGUGCAAGCCUCUU-5′
28	5′-ACGUGACACGUUCGGAGAATT-3′

[a] **B**^{C}: N^4-ethyl-N^4-coupled 2′-deoxy-5-methylcytidine monomer, **B**^{A}: N^6-ethyl-N^6-coupled 2′-deoxyadenosine monomer.

To further investigate if these terminal modifications are accepted in a therapeutically relevant gene-silencing process, we focused our attention on siRNAs. RNA interference (RNAi)[18] has received considerable attention over the past decade for its high potency and potential inhibition of a wide variety of overexpressed genes. We designed and synthesized siRNAs that target the site 501–519 of *Renilla* luciferase mRNA (Table 1 and Figure 3B, below). We prepared sense and guide siRNA strands containing no modifications and B^C or B^A units in place of the natural 3′-dinucleotide overhang TpT (unmodified, 3′-B^C- and 3′-B^A-modified sense strands **16**–**18**, respectively, and unmodified, 3′-B^C- and 3′-B^A-modified guide strands **19**–**21**). Modified oligonucleotide strands could be synthesized by using a DNA/RNA synthesizer without substantial alteration of standard synthesis protocols. Moreover, as controls for RNAi studies, we also prepared scrambled versions of sense and guide siRNA strands **16** and **19** (**22** and **23**, respectively; Table 1).

Finally, we also prepared unmodified and modified sense and guide siRNA strands for RNAi studies targeting the endogenous gene Bcl-2 (unmodified and 3′-B^C-modified sense strands **24** and **25**, and unmodified guide strand **26**; Table 1) and scrambled versions as negative controls (sense and guide strands **27** and **28**).

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Effect of 3′-terminal N-ethyl-N modifications on the thermal and structural properties of oligonucleotides

CD spectra of selected double-stranded oligonucleotides (siRNAs and DNAs containing no modifications or a B^C unit at the 3′-end of one of the strands composing the duplex) clearly indicated that the overall conformation of typical A-form and B-form double helical geometries are retained in the 3′-B^C-modified siRNA and in the 3′-B^C-modified double-stranded DNA, respectively (see Figure S27).

Moreover, thermal-denaturation studies of siRNA and DNA duplexes containing B^C or B^A units at their 3′ termini (Figure 3B, below, and Table S2) showed that the presence of a 3′-terminal dimeric modification did not cause any significant effect on the T_m of the duplex. The T_m values of 3′-modified siRNAs and 3′-modified DNAs were comparable to that of their unmodified versions.

Finally, MD simulations of a single-stranded RNA (5′-AUCUGAAGAAGGAGAAAAA^{19}TT-3′) and its corresponding 3′-B^C-modified version indicated that the presence of a 3′-B^C-modification does not alter the global geometry of the oligonucleotides (a plot of RMSD values calculated along the trajectories with respect to the starting structures can be found in Figure S30A). Only an induced local effect of the dimeric nucleoside on the sugar pucker phase angle of the neighboring ribonucleotide unit A19 (C3′-*endo* to C2′-*endo* transition) was observed (Figure S30C).

Effect of 3′-terminal N-ethyl-N modifications on the hydrolytic activity of 3′-exonucleases

We then investigated the blocking effect of dimeric nucleotide-containing oligonucleotides on the active site of 3′-exonucleases. As a model system we used the 3′–5′-exonucleolytic active site of the Klenow fragment (KF) of *E. coli* DNA polymerase I, which catalyzes the hydrolysis of a phosphodiester bond with the aid of two divalent metal ions.[19] Structure, exonuclease functions, and mechanistic aspects of the KF have been extensively studied in the past.[19,20] In particular, amino acid sequence-homology studies provided evidence that most of the residues at the 3′–5′-exonuclease active site of DNA polymerase I (Asp355, Glu357, Leu361, Asp424, Phe473, Tyr497, and Asp501) are conserved among the 3′-exonucleolytic domain of several prokaryotic and eukaryotic polymerases,[20] thus suggesting that structural information on the 3′–5′-exonuclease active site of KF might be applicable to other enzymes that catalyze 3′–5′-exonuclease reactions. Such observations have prompted several research groups to use the KF as a model to study the effect of oligonucleotide modifications on the 3′-exonuclease reaction.[21] Indeed, recent studies have revealed that the structures of one of the most abundant mammalian 3′-exonucleases (TREX1) bound to DNA are closely related to the structures of KF:DNA complexes.[22]

Starting from the X-ray coordinates of a single-stranded DNA 3′-*S*-phosphorothiolate trimer bound to the 3′–5′-exonucleolytic active site of the KF in the presence of Zn^{2+} and Mn^{2+} ions (PDB ID: 2KFN),[19a] we performed 50 ns MD simulations of 5′-ApB^C:KF and 5′-ApT1pT2:KF complexes (here B^C = subunit B^C1–ethyl–subunit B^C2 and A = 2′-deoxyadenosine; Figure 1). In the case of the native 5′-ApT1pT2:KF complex, the deoxy-3′-*S*-phosphorothiolate was replaced by thymidine. However, for the 5′-ApB^C:KF complex, we started from two different conformations of the B^C unit and used as template the

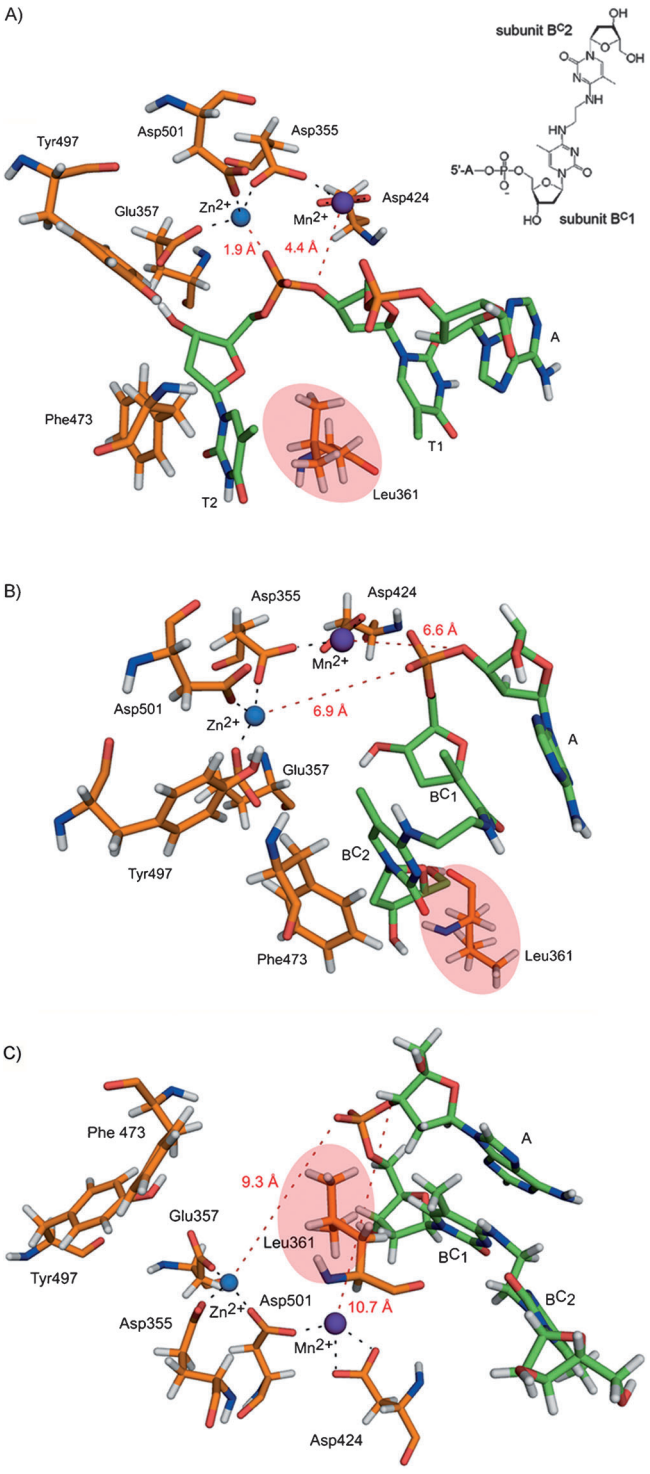


Figure 1. Representative snapshots from the MD trajectory (50 ns) showing the position of relevant KF amino acid residues. A) The unmodified DNA trimer ApT1pT2. The B) stacked and C) extended 3′-B^C-modified DNA trimer ApB^C1-ethyl-B^C2.

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

original trimer oligonucleotide: an intramolecular stacked conformation and an extended conformation of the dimer. In the first case, the starting conformation for the simulation is stabilized by nucleotide–nucleotide and nucleotide–protein interactions (such as stacking between the three nucleobases and stacking of B^C2 against Phe473), whereas in the second case, intramolecular stacking between B^C1 and B^C2 is lost, and the remarkably favorable nucleotide–protein interactions do not occur.

The final MD structure of the ApT1pT2 (unmodified):KF complex agreed well with the crystal structure (rmsd = 3.6 Å), and all the key relevant protein–DNA interactions were maintained (Figure 1A). As in the crystal structure, the side chain of residue Leu361 is wedged between the nucleobases of the two last 3′-nucleotides (T1 and T2), Phe473 stacks against the nucleobase of T2, and the 3′-terminal phosphodiester bond is well accommodated in the active site, with one of the nonbridging oxygen atoms interacting with the Zn^{2+} ion.[19] Previous studies have revealed that this divalent ion facilitates the precise positioning of the scissile phosphodiester bond with respect to an incoming nucleophile (water or a hydroxide ion; not shown). The second metal ion (Mn^{2+}) interacts directly with the bridging oxygen of the phosphodiester linkage. It has been suggested that this ion stabilizes the negative charge that comes to reside on the leaving 3′-oxygen after nucleophilic attack.[19]

As B^C is located at the 3′-end of the oligonucleotide and the B^C1 and B^C2 subunits are linked together by an ethyl chain through the exocyclic amino group of the nucleobase, the phosphodiester bond that is susceptible to undergoing cleavage is the one that links B^C1 to the neighboring natural nucleotide (in this case, 2′-deoxyadenosine; A). MD simulations of 5′-ApB^C:KF complexes showed that modification of B^C has a negative effect on the correct positioning of this phosphodiester bond at the 3′-exonuclease active site (Figure 1B and C). In both cases (stacked and extended forms, respectively), the bridging and nonbridging oxygens of this linkage are positioned far away from the catalytically important Zn^{2+} and Mn^{2+} ions.

Interestingly, in the case of the stacked 5′-ApB^C:KF complex, as the MD simulation progressed Leu361 was gradually shifted 10.4 Å away from its initial ($t = 0$) position (Figure 1B). Analysis of the MD trajectory suggests that this is due to a steric clash between the Leu361 side chain and the dimer (B^C) ethyl linker. This steric factor might force the phosphodiester bond between the deoxyadenosine unit (A) and the B^C1 unit to move away from the metal ions. Remarkably, the stacking interaction between the nucleobase of B^C2 and Phe473 was not lost over the course of the simulation.

Although the final snapshot from the MD trajectory of the extended 5′-ApB^C:KF complex (Figure 1C) did not reveal significant changes in the structure of the 3′-exonuclease active site, analysis of the simulations suggest that steric clashes between Leu361 and the ethyl linker might block the entry of the dimer into the active site, leading to a loss of interaction between the phosphodiester and metal ions.

Previous analyses based on site-directed mutagenesis revealed that mutation of several amino acid residues at the 3′-exonuclease active site of KF (among them Leu361), caused a decrease in the hydrolytic activity of the enzyme.[23] Thus, in agreement with those observations, our results indicate that Leu361 might play an important role in the catalytic activity of the enzyme. Hydrophobic interaction of Leu361 with the two terminal nucleobases of a natural oligonucleotide might force the 3′-terminal phosphodiester bond to accommodate well in the active site, and thus interact with the Zn^{2+} and Mn^{2+} ions (Figure 1A). However, steric clashes between this residue and the dimer ethyl linker might interfere with the correct positioning of the adjacent phopshodiester bond (linking B^C1 to the neighboring natural nucleotide; in this case nucleotide A, located in the 5′-end) at the active site (Figure 1B and C). Due to this steric factor, the enzyme might not be able to bypass the dimeric nucleotide unit and cleave this phosphodiester bond.

In view of the promising results obtained, we evaluated the effect of the 3′-N-ethyl-N modification on the stability of the corresponding oligonucleotide derivatives against the KF. As this enzyme is highly specific for DNA, oligodeoxynucleotides **12** (unmodified), **13** (3′-B^C-modified), and **14** (3′-B^A-modified) were incubated with KF at 37 °C, and the degradation patterns were assayed by gel electrophoresis. In order to compare the 3′-N-ethyl-N dimers with the phosphorothioate (PS) modification,[9] the KF stabilities of a 3′-PS-modified version of **ON13** (**ON29**; in which two natural thymidine units linked by a phosphorothioate bond replace the B^C unit; Figure 2A) and a fully PS-modified DNA (**30**; Figure S29A) were also evaluated. Oligonucleotide **12** was rapidly degraded, whereas DNAs containing a dimeric nucleoside (B^C or B^A) at the 3′-end (**13** and **14**, respectively) showed strongly enhanced stability, with no visible degradation even after 72 h of incubation with KF, thus confirming the strong 3′-exonuclease resistance predicted by our calculations. In contrast to this, the integrated intensities of the gel bands showed that ~85 and ~50% of 3′-PS-modifed (**29**) and fully PS-modified (**30**) DNAs, respectively, were hydrolyzed after 72 h (Figure 2A and S29A). Thus, although a single 3′-terminal PS modification confers high levels 3′-exonuclease resistance, it does not block the 3′-exonucleolytic activity of the enzyme.

To extend our investigations to other classes of oligonucleotides, we performed nuclease digestion studies with snake venom phosphodiesterase (SNVPD), which is another relevant 3′-exonuclease that degrades both ssDNA and ssRNA.[24] It has been reported that SNVPD has catalytic properties similar to those of the nucleotide pyrophosphatases/phosphodiesterases present in plasma, which catalyze the 3′-exonuclease degradation of oligonucleotides.[25] As observed with the KF, the 3′-exonuclease activity of SNVPD is also blocked by the dimeric nucleoside. In fact, even after three days of incubation with SNVPD, 3′-B^C- and 3′-B^A-modified RNAs **17** and **18** remained untouched, whereas unmodified RNA **16** was completely hydrolyzed after 1 h. When we treated a 3′-PS-modified version of **ON17** (**ON31**) with SNVPD, significant degradation was observed after 1 h (Figure 2B). The degradation profile for an RNA strand containing four PS modifications at the 3′-end (**32**,

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

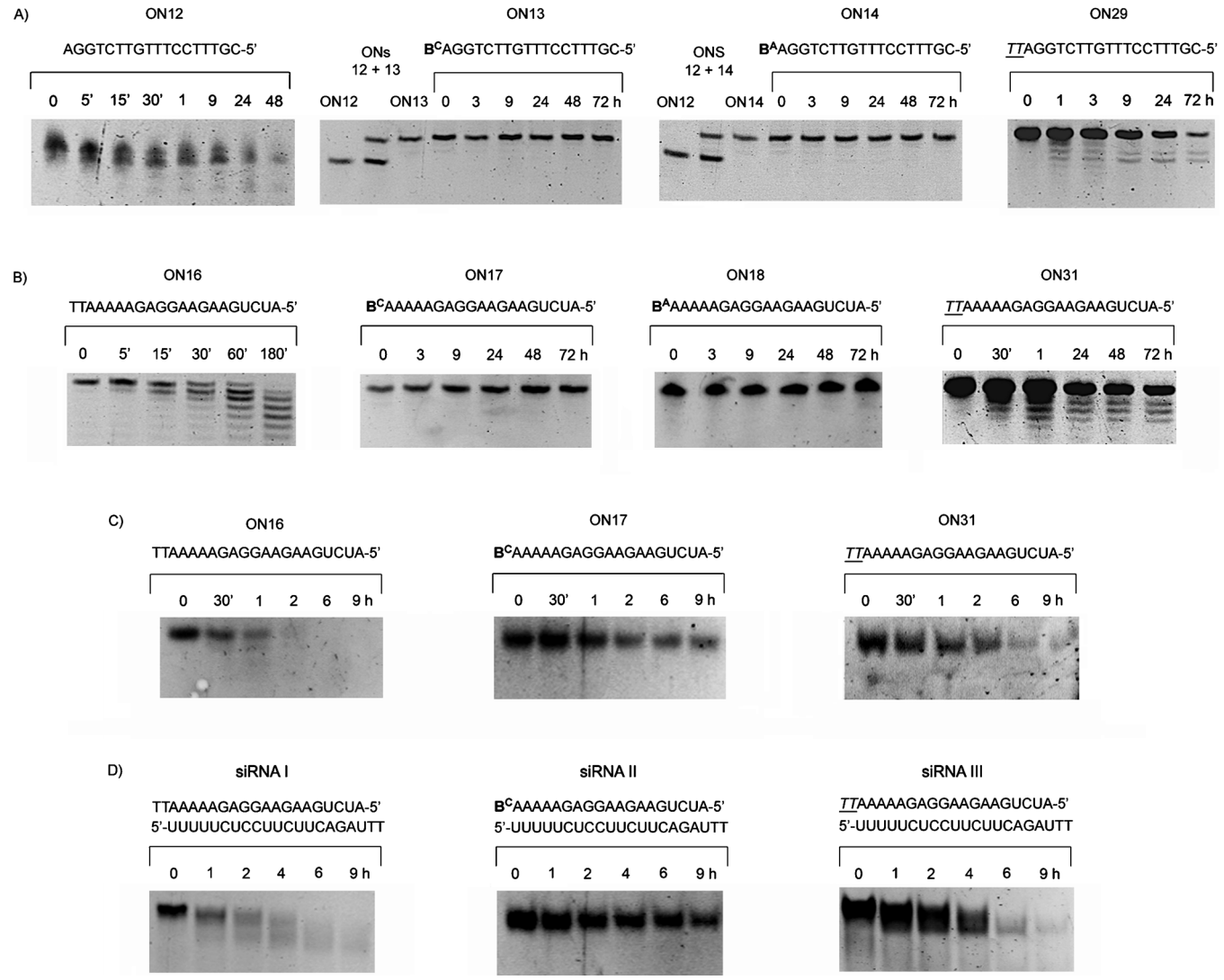


Figure 2. A) 20% denaturing polyacrylamide gels depicting the time course of the KF-catalyzed degradation of unmodified, 3′-B^C-modified, 3′-B^A-modified, and 3′-PS-modified single-stranded oligodeoxynucleotides **12**–**14** and **29**, respectively. B) 20% denaturing polyacrylamide gels depicting the time course of the SNVPD-catalyzed degradation of unmodified, 3′-B^C-modified, 3′-B^A-modified, and 3′-PS-modified single-stranded RNAs **16**–**18** and **31**, respectively. The oligonucleotides were incubated with SNVPD (10 mU) at 37 °C. C) and D) 20% non-denaturing polyacrylamide gels of unmodified, 3′-B^C-modified, and 3′-PS-modified single-stranded RNAs **16**, **17** and **31** (C) and unmodified, 3′-B^C-sense-modified, and 3′-PS-sense-modified double-stranded siRNAs I, II and III (D) incubated in PBS containing 40% human serum at 37 °C. All oligonucleotides were withdrawn at indicated points, separated, and visualized with SYBR green II. Underlined capital letters indicate phosphorothioate modification.

Figure S29B) was very similar to that observed for **ON31**. Even a fully PS-modified RNA (**33**, Figure S29B) underwent degradation.

Finally, the SNVPD susceptibilities of unmodified and 3′-modified DNA derivatives **12**–**14**, **29**, and **30** (Figure S29C) were very similar to those obtained for RNAs **16**–**18**, **31**, and **33**.

Serum stability of 3′-N-ethyl-N-modified single- and double-stranded siRNAs

We then evaluated the effect of the 3′-N-ethyl-N modification on the serum stability of oligonucleotides. Unmodified and selected modified single- and double-stranded siRNAs were incubated in PBS containing 40% human serum, and the reaction mixtures were analyzed by gel electrophoresis under non-denaturing conditions. Unmodified ssRNA **16** showed low stability. Complete degradation of the original RNA was observed after 2 h in serum (Figure 2C). Interestingly, ssRNA with a B^C unit at the 3′-end (**17**) displayed strongly enhanced stability. The integrated intensities of the gel bands showed that ~15% of the original RNA population remained intact after 9 h in serum. In contrast to this, ~95% of the 3′-PS-modified RNA **31** was hydrolyzed after 6 h (Figure 2C). A double-stranded siRNA with only a B^C unit at the 3′-end of the sense strand (siRNA II) demonstrated even higher stability, with ~25% of the original siRNA remaining intact for 9 h (Figure 2D). In contrast, 3′-PS-sense modified siRNA III was completely degraded after 6 h in serum.

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

RNAi activities of B^C- and B^A-modified siRNAs targeting the *Renilla* luciferase gene

To evaluate whether B^C- and B^A-modified siRNAs regulate gene expression through the RNAi pathway, we carried out separate RNAi studies in SH-SY5Y human neuroblastoma cells with *Renilla* luciferase siRNAs containing B^Cs at the 3′-ends of the sense and/or guide strands (II, V and VII), B^As at the 3′-ends of one or both strands (IV, VI and VIII), and with the unmodified and the scrambled siRNAs I and sc1, respectively (Figure 3B). We chose a luciferase model system because it allowed rapid determination of RNAi activity. The cells were first transfected with dual reporter plasmids that express *Renilla* luciferase (the target) and nontargeted firefly luciferase as an internal control. The effects of the different siRNAs on luciferase expression were evaluated after dosing with double-stranded siRNAs (16 pg–210 ng; 2 pM–26 nM) in the cell medium, and measurement of luminescence responses after 22 h. The results, showing *Renilla* luciferase activity normalized to firefly luciferase, are represented in Figure 3A.

Remarkably, all the siRNAs used in this study were potent inhibitors of luciferase activity with sub-nanomolar IC_{50} values. Interestingly, siRNAs containing a B^C or a B^A unit at the 3′-end of the sense strand (II and IV) displayed gene-silencing activity significantly higher than that of unmodified siRNA (I). The most significant differences were observed when very low concentrations (8 and 2 pM) of siRNAs were employed (significant differences were assessed by Bonferroni test, see data in Figure S28). For example, at a concentration of 8 pM, there was (34 ± 9) and (37 ± 9)% gene knockdown for II and IV respectively, versus (16 ± 9)% for I ($p<0.05$; Figure S28).

Replacing the natural 3′-TT-guide overhang with a B^C or a B^A unit (siRNAs V–VIII) caused a slight decrease in activity, but the gene-silencing activity of these siRNAs was still quite remarkable: (40 ± 4), (37 ± 2), (34 ± 10), and (38 ± 11)% for V–VIII, respectively, versus (51 ± 4)% for I; in all cases a concentration of 16 pM was used.

RNAi activities of B^C-modified siRNAs targeting Bcl-2

The inhibition of expression of antiapoptotic genes is a key strategy in the treatment of cancer. In particular, the Bcl-2 gene has attracted the attention of many groups.[26] Because RNA interference has been shown to inhibit the expression of virtually any gene in cell culture, it could represent a promising method for the treatment of cancer. In order to extend our RNAi studies to therapeutically relevant genes, the effect of B^C-modified siRNAs on the inhibition of expression of Bcl-2 gene was examined.

One of the most promising siRNA designs, corresponding to siRNA II, was selected from our previous *Renilla* luciferase gene-silencing results and used to target Bcl-2 mRNA in MCF-7 human breast adenocarcinoma cells with use of a previously described sequence[26a] (Table 1). We prepared the unmodified Bcl-2 siRNA IX and its B^C-modified version X (Figure 3E) containing a B^C modification at the same position as in the *Renilla* luciferase siRNA II (a B^C unit replacing the natural 3′-TT overhang of the sense strand). The effect of the siRNAs on the Bcl-2 mRNA and protein levels were assessed by quantitative real-time PCR and western blot, respectively, 24 h after transfection at a siRNA concentration of 60 nM. As shown in Figure 3C, Bcl-2 mRNA levels were significantly down-regulated by unmodified (IX) and modified (X) siRNAs. Modified siRNA X displayed activity comparable to that of wild-type siRNA IX. Changes in Bcl-2 protein levels 24 h after the treatment of MCF-7 cells with either of the two Bcl-2 siRNAs corresponded well with a siRNA-induced reduction in Bcl-2 mRNA levels (Figure 3C and D). The reduction in Bcl-2 protein levels reached approximately 93 or 94% (for siRNAs IX or X, respectively) 24 h after Bcl-2 siRNA treatment. No down-regulation was seen with the control siRNA (sc2; Figure 3C and D), thus indicating a specific mode of action.

Conclusions

In this work, we have developed a new class of oligonucleotide modification that confers extraordinary resistance against 3′-exonuclease activity. This class of alteration involves minimal and selective modification. Although many examples of nuclease-resistant oligonucleotides have been reported, most of them require extensive modification. Moreover, 3′-terminal N-ethyl-N-coupled nucleosides conferred higher 3′-exonuclease resistance than the phosphorothioate modification and were accepted by the RNAi machinery. As our modification is located at the end of the oligonucleotide strand, it could be compatible with conventional oligonucleotide chemistries (like LNA and PS modifications) at internal positions to give rise to therapeutic oligonucleotide analogues with better efficiency and even higher nuclease stability.

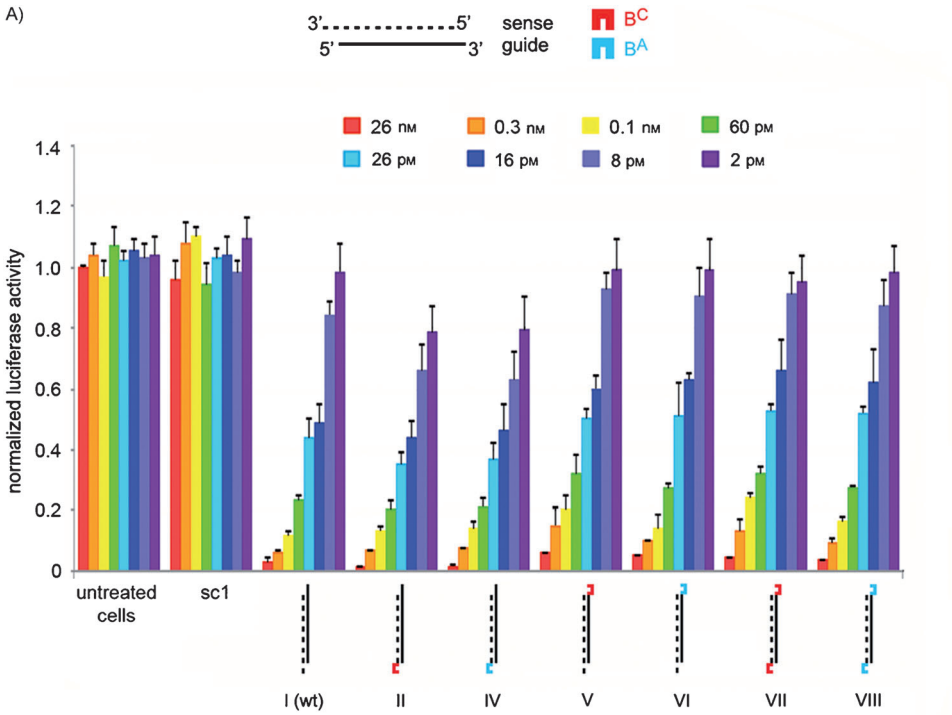
A detailed computational study has revealed a possible explanation for the strong 3′-exonuclease stability of these oligonucleotide analogues, which can be exploited to rationally design even more effective modifications. Our theoretical studies suggest that disruption of the interactions between the two 3′-terminal nucleobases and the amino acids at the 3′-exonuclease active site cause a negative effect on the correct positioning of the adjacent scissile phosphodiester bond. Such studies not only provide a deeper insight into the role of nucleobase–protein interactions on 3′-exonuclease function, but can also help to design new potent analogues, such as dimers containing different classes of linkers, and/or nucleobases of different shape and size.

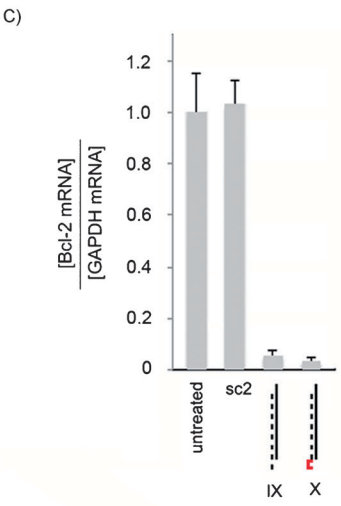
Finally, it is noteworthy that our N-ethyl-N dimeric nucleosides are conjugated to the oligonucleotide chain through the 5′-oxygen of one of the two nucleoside monomers composing the dimer. This leaves the 5′- and 3′-oxygen atoms of the second monomer free to participate in the conjugation (prior to oligonucleotide synthesis) with lipids or peptides aimed at facilitating the delivery of the oligonucleotide into the cell.

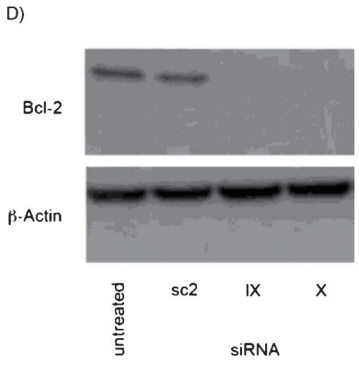
Experimental Section

Computational methods: Molecular dynamics simulations of the naked ssRNA and complex ssDNA–Klenow fragment were used to

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim







B)

siRNA	ON	Sequence	T_m [°C]	IC_{50} [pM]
I	16 19	TTAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUTT	61.1	25
II	17 19	B^CAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUTT	60.9	13
IV	18 19	B^AAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUTT	60.8	13
V	16 20	TTAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUB^C	60.8	30
VI	16 21	TTAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUB^A	61.6	28
VII	17 20	B^CAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUB^C	60.8	33
VIII	18 21	B^AAAAAAGAGGAAGAAGUCUA-5' 5'-UUUUUCUCCUUCUUCAGAUB^A	61.0	28
sc1	22 23	TTAGGAAAGAAAGAAAGCUAU-5' 5'-UCCUUUCUUUCUUUCGAUATT	-	-

E)

siRNA	ON	Sequence
IX	24 26	TTGAAGUAGUGAUAGAGGGCC-5' 5'-CUUCAUCACUAUCUCCCGGTT
X	25 26	B^CGAAGUAGUGAUAGAGGGCC-5' 5'-CUUCAUCACUAUCUCCCGGTT
sc2	27 28	TTUGCACUGUGCAAGCCUCUU-5' 5'-ACGUGACACGUUCGGAGAATT

Figure 3. A) Plot of gene-specific RNAi activity for B^C- and B^A-modified siRNAs targeting the *Renilla* luciferase mRNA in SH-SY5Y cells. Various amounts of siRNAs were added as shown. B) Sequences of unmodified, B^C-modified, and B^A-modified siRNAs targeting the *Renilla* luciferase mRNA, and T_m data. C) Down-regulation of Bcl-2 mRNA induced by siRNAs IX, X and sc2 (60 nM) in MCF-7 cells. Relative Bcl-2 levels were assessed by quantitative RT-PCR. D) Representative immunoblots of Bcl-2 and β-actin (internal control) proteins from cells treated as described in (C). E) Sequences of unmodified and B^C-modified siRNAs targeting the Bcl-2 mRNA. The upper strand is the sense strand in the 3′→5′ direction (same as the target sequence), the lower strand is the guide strand in the 5′→3′ direction (complementary to the target). B^C: N^4-ethyl-N^4-coupled 2′-deoxy-5-methylcytidine monomer. B^A: N^6-ethyl-N^6-coupled 2′-deoxyadenosine monomer. Sc1 and sc2: scrambled sequences. Bars indicate mean ± S.D. ($n = 6$ independent treatments).

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

explain the ability of the 1,2-di(5-methylcytidin-N^4-yl)ethane to block 3′-exonuclease digestion from the structural point of view. For comparison, we prepared all systems containing normal phosphoribonucleotides and our modified compounds. We substituted the original deoxy-3′-*S*-phosphorothiolate (US1 residue in the X-ray structure, PDB ID: 2kfn) by the standard thymidine (leading the hanging trimer d(5′-ApTpT-3′), which was later modified to create the oligomer d(5′-ApB^C-3′) in which B^C stands for the cytidinyl dimer). In the simulations of naked RNA, the chosen sequence was r(5′-AUCUG AAGAA GGAGA AAAAB^C-3′). The 1,2-dicytidin-N^4-yl-ethane residue was parameterized by using the RED program,[27] which automates the calculation of RESP charges at the HF/6–31G(d) level of theory. The antechamber module was used to obtain the remaining force field parameters according to GAFF force field procedures.[28]

Single-stranded RNA initial structures were built with the canonical A-form parameters frame with the NAB module[29] included in the Ambertools package (v. 1.5); the KF–ssDNA complex with the modified nucleobase was built with the xLeap module of AMBER. The systems were immersed in a TIP3P water periodic box, extended 12 Å away from any solute atom. Sodium counterions were added in order to neutralize the negative charges of the oligomers, and, in the case of the KF complex structures, Mn^{2+} and Zn^{2+} ions[19a] already present in the original X-ray structure were kept in the exonuclease binding pocket. The minimization and equilibration processes were run according to the protocol successfully used in previous works.[30] Unrestrained MD simulations were run in the isothermic–isobaric (NPT; 298 K; 1 atm) ensemble by using periodic boundary conditions and the Particle Mesh Ewald.[31] The SHAKE algorithm,[32] which allowed us to use a 2 fs of integration step, was used. Parm99[33] and Parm99SB[34] AMBER force fields for nucleic acids and proteins, respectively, together with latest force field improvements for nucleic acids, *parmbsc0*[35] and *chiOL3*,[36] were used in the MD simulations. All the MD trajectories were extended to 50 ns. Postprocessing analysis of the trajectories was carried out with the ptraj module.

Synthetic protocols: See the Supporting Information.

3′-Exonuclease digestions

Snake venom phosphodiesterase: Each single-stranded oligomer (RNA or DNA; 120 pmol) was incubated with phosphodiesterase I from *Crotalus adamanteus* venom (SNVPD; 340 ng, 10 mU or 680 ng, 20 mU) in a buffer containing Tris·HCl (56 mM, pH 7.9) and $MgSO_4$ (4.4 mM, total volume = 40 μL) at 37 °C. At appropriate times, aliquots of the reaction mixture (5 μL) were taken and added to a solution of urea (15 μL, 9 M), and the mixtures were analyzed by electrophoresis on a 20% polyacrylamide gel containing urea (7 M). Oligonucleotide bands were visualized with the SYBR Green II reagent (Sigma–Aldrich) according to the manufacturer's instructions.

Klenow fragment of E. coli polymerase I: Single-stranded DNA oligomers (300 pmol) were incubated with *E. coli* polymerase I (3.2 U) in Tris buffer (50 mM, pH 8.0) containing NaCl (50 mM) and $MgCl_2$ (10 mM) at 37 °C (total volume = 45 μL). At appropriate times, aliquots (5 μL) were removed and added to a solution of urea (15 μL, 9 M), and the samples were analyzed by electrophoresis on 20% polyacrylamide gel containing urea (7 M).

Stability of RNA and DNA oligonucleotides in PBS containing human serum: Each oligonucleotide (300 pmol) was incubated in PBS containing 40% of human serum (total volume = 75 μL) at 37 °C. At appropriate times, aliquots of the reaction mixture (5 μL) were separated and added to a glycerol loading solution (15 μL), and the samples were run on a 20% polyacrylamide gel under non-denaturing conditions.

UV-monitored thermal-denaturation studies: Absorbance versus temperature curves of duplexes were measured at 1 μM strand concentration in phosphate buffer (10 mM, pH 7.0) containing EDTA (0.1 mM) and NaCl (100 mM). Experiments were performed in Teflon-stoppered quartz cells of 1 cm path length on a JACSO V-650 spectrophotometer equipped with thermoprogrammer. The samples were heated to 90 °C, allowed to cool slowly to 25 °C, and then warmed during the denaturation experiments at a rate of 0.5 °C min^{-1} to 85 °C; the absorbance was monitored at 260 nm. The data were analyzed by the denaturation curve-processing program, MeltWin v. 3.0. Melting temperatures (T_m) were determined by computer fit of the first derivative of absorbance with respect to 1/*T*.

CD measurements: CD spectra were recorded on a Jasco J-810 spectropolarimeter equipped with a Julabo F/25HD temperature controller under the same buffer conditions and with the same oligonucleotide concentrations as for UV melting curves. All spectra were recorded at room temperature between 220 and 320 nm by using a 100 nm min^{-1} scan rate. The graphs were analyzed by using Origin software.

Luciferase siRNA assays: SH-SY5Y cells were regularly passaged to maintain exponential growth. The cells were seeded 1 day prior to the experiment in a 24-well plate at a density of 150 000 cells per well in complete Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal bovine serum (FBS; 500 μL per well). Following overnight culture, the cells were treated with luciferase plasmids and siRNAs. Two luciferase plasmids—*Renilla* luciferase (pRL-TK) and firefly luciferase (pGL3) from Promega—were used as reporter and control, respectively. Co-transfection of plasmids and siRNAs was carried out with Lipofectamine 2000 (Life Technologies) as described by the manufacturer for adherent cell lines; pGL3-control (1.0 μg), pRL-TK (0.1 μg), and siRNA duplex (2 pM–26 nM) formulated into liposomes were added to each well with a final volume of 600 μL. After an incubation period of 5 h, cells were rinsed once with PBS and fed with fresh DMEM (600 μL) containing 10% FBS. After a total incubation time of 22 h, the cells were harvested and lysed with passive lysis buffer (100 μL per well) according to the instructions of the Dual-Luciferase Reporter Assay System (Promega). The luciferase activities of the samples were measured with a MicroLuma*Plus* LB 96V (Berthold Technologies) with a delay time of 2 s and an integration time of 10 s. The following volumes were used: 20 μL of sample and 30 μL of each reagent (Luciferase Assay Reagent II and Stop and Glo Reagent). The inhibitory effects generated by siRNAs were expressed as normalized ratios between the activities of the reporter (*Renilla*) luciferase gene and the control (firefly) luciferase gene. IC_{50} values were calculated by using GraphPad Prism software with the sigmoidal dose–response function.

Assessment of Bcl-2 mRNA levels by quantitative real-time PCR: MCF-7 cells were seeded 1 day prior to transfection in 60 mm dishes at a density of 620 000 cells per dish in complete DMEM containing 10% FBS. Following overnight culture, siRNA duplexes (60 nM per dish) formulated into liposomes were added to each dish with a final volume of 6 mL. Co-transfection of siRNAs was carried out by using Lipofectamine 2000. After an incubation time of 5 h, the transfection medium was changed to complete DMEM containing 10% FBS. After an incubation time of 24 h, the cells were harvested, and total RNA was isolated by using an RNeasy

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Mini kit (Qiagen). Purified RNA was used as a template to assess the gene expression level of Bcl-2 through quantitative reverse-transcription (qRT-PCR). First, reverse transcription was performed by using the high-capacity cDNA reverse transcriptase kit (Applied Biosystems). After cDNA synthesis, qRT-PCR was carried out by using the Power SYBR Green PCR Master Mix kit (Applied Biosystems). For both Bcl-2 and GAPDH, custom primers were purchased with sequences Bcl-2, forward: 5′-GGTGA ACTGG GGGAG GATTG T, reverse: 5′-CTTCA GAGAC AGCCA GGAGA A; GAPDH, forward: 5′-TGCAC CACCA ACTGC TTAG, reverse: 5′-GATGC AGGGA TGATG TTC. The data were normalized to GAPDH, which was selected as internal control.

Analysis of Bcl-2 protein knockdown by western blot: MCF-7 cells were seeded 24 h before transfection in 60 mm dishes at a density of 620 000 cells per dish in DMEM containing 10% FBS. Following overnight culture, siRNA duplexes (60 nM per dish) formulated into liposomes were added to each dish with a final volume of 6 mL. Co-transfection of siRNAs was carried out by using Lipofectamine 2000. After an incubation time of 5 h, the transfection medium was changed to complete DMEM containing 10% FBS. After an incubation time of 24 h, the cells were harvested with PBS and lysed by incubation in RIPA buffer (150 mM NaCl, 1% Triton X-100, 0.5% sodium deoxycolate, 0.1% SDS, 50 mM Tris, pH 8.0) containing protease inhibitors (Roche) at 4 °C for 1 h. Cell debris was removed by centrifugation at 8000 *g* for 20 min at 4 °C, and the protein concentration was determined by using the BCA assay (Pierce). Protein (30 μg) was resolved by SDS-PAGE and transferred to a poly(vinylidene difluoride) membrane (Immobilon-P, Millipore). The membrane was blocked with 5% skim milk in Tris-buffered saline containing 0.1% Tween for 1 h at room temperature and subsequently probed with anti-Bcl-2 monoclonal antibody (Dako, Glostrup, Denmark; diluted 1:500 in blocking buffer) overnight at 4 °C. β-Actin was selected as internal control and was detected by incubation with anti-actin monoclonal antibody (Sigma–Aldrich) in blocking buffer (1:3500) for 1 h at room temperature. Horseradish peroxidase-labeled polyclonal goat anti-mouse secondary antibody (Thermo Scientific) was incubated in the blocking solution (1:1000) for 1 h at room temperature. The intensities of the bands were analyzed by using ImageJ 1.45 software (Rasband, W.S., ImageJ, U.S. National Institutes of Health, Bethesda, MD, USA, http://imagej.nih.gov/ij/, 1997–2011).

Statistical analysis: Data were analyzed by using the GraphPad Prism 5 program (GraphPad Software). Significant differences were assessed by ANOVA to compare three or more groups followed by Bonferroni test. In all figures, $*p < 0.05$.

Acknowledgements

This research was supported by the European Union (MULTIFUN, NMP4-LA-2011–262943), the Spanish Ministry of Education (CTQ2010–20541), and the Generalitat de Catalunya (2009/SGR/208). M.O. thanks the Spanish Ministry of Science and Innovation (BIO2009–10964 and Consolider E-Science), the Instituto Nacional de Bioinformática (INB), the European Research Council (SimDNA ERC Advanced Grant), and the Fundación Marcelino Botín. M.T. acknowledges a Juan de la Cierva contract (MICINN, Spain) for financial support, and A.A. acknowledges IRB Barcelona for a predoctoral fellowship. We thank Drs. Isaac Gállego and Francisco Miguel Torres (IRB Barcelona) for helpful advice and discussions.

Keywords: computational chemistry · DNA · exonuclease resistance · N-ethyl-N-coupled nucleosides · RNA

[1] P. C. Zamecnik, M. L. Stephenson, *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 280 – 284.
[2] J. K. Watts, G. F. Deleavey, M. J. Damha, *Drug Discovery Today* **2008**, *13*, 842 – 855.
[3] A. D. Keefe, S. Pai, A. Ellington, *Nat. Rev. Drug Discovery* **2010**, *9*, 537 – 550.
[4] J. Mulhbacher, P. St-Pierre, D. A. Lafontaine, *Curr. Opin. Pharmacol.* **2010**, *10*, 551 – 556.
[5] G. Deleavey, M. J. Damha, *Chem. Biol.* **2012**, *19*, 937 – 954.
[6] N. M. Bell, J. Micklefield, *ChemBioChem* **2009**, *10*, 2691 – 2703.
[7] a) Y.-L. Chiu, T. M. Rana, *RNA* **2003**, *9*, 1034 – 1048; b) T. Dowler, D. Bergeron, A.-L. Tedeschi, L. Paquet, N. Ferrari, M. J. Damha, *Nucleic Acids Res.* **2006**, *34*, 1669 – 1675; c) F. Czauderna, M. Fechtner, S. Dames, H. Aygün, A. Klippel, G. J. Pronk, K. Giese, J. Kaufmann, *Nucleic Acids Res.* **2003**, *31*, 2705 – 2716; d) J. Elmén, H. Thonberg, K. Ljungberg, M. Frieden, M. Westergaard, Y. Xu, B. Wahren, Z. Liang, H. Ørum, T. Koch, C. Wahlestedt, *Nucleic Acids Res.* **2005**, *33*, 439 – 447; e) C. Wahlestedt, P. Salmi, L. Good, J. Kela, T. Johnsson, T. Hökfelt, C. Broberger, F. Porreca, J. Lai, K. Ren, M. Ossipov, A. Koshkin, N. Jakobsen, J. Skouv, H. Oerum, M. H. Jacobsen, J. Wengel, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 5633 – 5638; f) M. Terrazas, S. M. Ocampo, J. C. Perales, V. E. Marquez, R. Eritja, *ChemBioChem* **2011**, *12*, 1056 – 1065; g) Y. S. Sanghvi in *Current Protocols in Nucleic Acid Chemistry* (Eds.: S. L. Beaucage, D. E. Bergstrom, P. Herdewijn, A. Matsuda), Wiley, Hoboken, **2011**, pp. 4.1.1 – 4.1.22.
[8] a) F. Eckstein, *Biochimie* **2002**, *84*, 841 – 848; b) P. Li, Z. A. Sergueeva, M. Dobrikov, B. R. Shaw, *Chem. Rev.* **2007**, *107*, 4746 – 4796.
[9] a) F. Eckstein, *Annu. Rev. Biochem.* **1985**, *54*, 367 – 402; b) F. Eckstein, *Antisense Nucleic Acid Drug Dev.* **2000**, *10*, 117 – 121.
[10] a) M. Terrazas, E. T. Kool, *Nucleic Acids Res.* **2009**, *37*, 346 – 353; b) M. Terrazas, R. Eritja, *Mol. Diversity* **2011**, *15*, 677 – 686.
[11] a) N. Potenza, L. Moggio, G. Milano, V. Salvatore, B. Di Blasio, A. Russo, A. Messere, *Int. J. Mol. Sci.* **2008**, *9*, 299 – 315; b) Y. Ikeda, D. Kubota, Y. Nagasaki, *Bioconjugate Chem.* **2010**, *21*, 1685 – 1690; c) M. M. Masud, T. Masuda, Y. Inoue, M. Kuwahara, H. Sawai, H. Ozaki, *Bioorg. Med. Chem. Lett.* **2011**, *21*, 715 – 717; d) Y. Ueno, Y. Watanabe, A. Shibata, K. Yoshikawa, T. Takano, M. Kohara, Y. Kitade, *Bioorg. Med. Chem.* **2009**, *17*, 1974 – 1981; e) K. Yoshikawa, A. Ogata, C. Matsuda, M. Kohara, H. Iba, Y. Kitade, Y. Ueno, *Bioconjugate Chem.* **2011**, *22*, 42 – 49; f) A. Somoza, M. Terrazas, R. Eritja, *Chem. Commun.* **2010**, *46*, 4270 – 4272; g) D. V. Morrissey, J. A. Lockridge, L. Shaw, K. Blanchard, K. Jensen, W. Breen, K. Hartsough, L. Machemer, S. Radka, V. Jadhav, N. Vaish, S. Zinnen, C. Vargeese, K. Bowman, C. S. Shaffer, L. B. Jeffs, A. Judge, I. MacLachlan, B. Polisky, *Nat. Biotechnol.* **2005**, *23*, 1002 – 1007.
[12] a) J. B. Bramsen, M. B. Laursen, A. F. Nielsen, T. B. Hansen, C. Bus, N. Langkjær, B. R. Babu, T. Højland, M. Abramov, A. Van Aerschot, D. Odadzic, R. Smicius, J. Haas, C. Andree, J. Barman, M. Wenska, P. Srivastava, C. Zhou, D. Honcharenko, S. Hess, E. Müller, G. V. Bobkov, S. N. Mikhailov, E. Fava, T. F. Meyer, J. Chattopadhyaya, M. Zerial, J. W. Engels, P. Herdewijn, J. Wengel, J. Kjems, *Nucleic Acids Res.* **2009**, *37*, 2867 – 2881; b) J. Harborth, S. M. Elbashir, K. Vandenburgh, H. Manninga, S. A. Scaringe, K. Weber, T. Tuschl, *Antisense Nucleic Acid Drug Dev.* **2003**, *13*, 83 – 105.
[13] a) J.-P. Shaw, K. Kent, J. Bird, J. Fishback, B. Froehler, *Nucleic Acids Res.* **1991**, *19*, 747 – 750; b) S. Choung, Y. J. Kim, S. Kim, H.-O. Park, Y.-C. Choi, *Biochem. Biophys. Res. Commun.* **2006**, *342*, 919 – 927.
[14] a) A. M. Noronha, D. M. Noll, C. J. Wilds, P. S. Miller, *Biochemistry* **2002**, *41*, 760 – 771; b) M. B. Smeaton, E. M. Hlavin, A. M. Noronha, S. P. Murphy, C. J. Wilds, P. S. Miller, *Chem. Res. Toxicol.* **2009**, *22*, 1285 – 1297; c) C. J. Wilds, A. M. Noronha, S. Robidoux, P. S. Miller, *J. Am. Chem. Soc.* **2004**, *126*, 9257 – 9265.
[15] K. C. Gupta, P. Kumar, D. Bhatia, A. K. Sharma, *Nucleosides Nucleotides* **1995**, *14*, 829 – 832.
[16] T. Atkinson, M. Smith in *Oligonucleotide Synthesis: A Practical Approach* (Ed.: M. J. Gait), IRL Press, Oxford, **1984**, pp. 35 – 81.
[17] a) J. Žemlička, *Biochemistry* **1980**, *19*, 163 – 168; b) J. Žemlička, J. Owens, *J. Org. Chem.* **1977**, *42*, 517 – 523.

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

[18] a) A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, C. C. Mello, *Nature* **1998**, *391*, 806 – 811; b) S. M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, T. Tuschl, *Nature* **2001**, *411*, 494 – 498; c) S. M. Elbashir, W. Lendeckel, T. Tuschl, *Genes Dev.* **2001**, *15*, 188 – 200; d) D. Bumcrot, M. Manoharan, V. Koteliansky, D. W. Y. Sah, *Nat. Chem. Biol.* **2006**, *2*, 711 – 719.

[19] a) C. A. Brautigam, S. Sun, J. A. Piccirilli, T. A. Steitz, *Biochemistry* **1999**, *38*, 696 – 704; b) J. F. Curley, C. M. Joyce, J. A. Piccirilli, *J. Am. Chem. Soc.* **1997**, *119*, 12691 – 12692; c) C. A. Brautigam, T. A. Steitz, *J. Mol. Biol.* **1998**, *277*, 363 – 377; d) L. S. Beese, T. A. Steitz, *EMBO J.* **1991**, *10*, 25 – 33.

[20] A. Bernad, L. Blanco, J. M. Lázaro, G. Martín, M. Salas, *Cell* **1989**, *59*, 219 – 228.

[21] M. Teplova, S. T. Wallace, V. Tereshko, G. Minasov, A. M. Symons, P. D. Cook, M. Manoharan, M. Egli, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 14240 – 14245.

[22] M. Brucet, J. Querol-Audí, M. Serra, X. Ramirez-Espain, K. Bertlik, L. Ruiz, J. Lloberas, M. J. Macias, I. Fita, A. Celada, *J. Biol. Chem.* **2007**, *282*, 14547 – 14557.

[23] V. Derbyshire, N. D. F. Grindley, C. M. Joyce, *EMBO J.* **1991**, *10*, 17 – 24.

[24] S. M. Linn, R. S. Lloyd, R. J. Roberts, *Nucleases*, Cold Spring Harbor Laboratory Press, New York, **1993**.

[25] a) M. Wójcik, M. Cieślak, W. J. Stec, J. W. Goding, M. Koziolkiewicz, *Oligonucleotides* **2007**, *17*, 134 – 145; b) R. Gijsbers, J. Aoki, H. Arai, M. Bollen, *FEBS Lett.* **2003**, *538*, 60 – 64.

[26] a) R. T. Lima, L. M. Martins, J. E. Guimarães, C. Sambade, M. H. Vasconcelos, *Cancer Gene Ther.* **2004**, *11*, 309 – 316; b) A. A. Pandyra, R. Berg, M. Vincent, J. Koropatnick, *J. Pharmacol. Exp. Ther.* **2007**, *322*, 123 – 132; c) C. W. Beh, W. Y. Seow, Y. Wang, Y. Zhang, Z. Y. Ong, P. L. R. Ee, Y.-Y. Yang, *Biomacromolecules* **2009**, *10*, 41 – 48; d) U. Akar, A. Chaves-Reyez, M. Barria, A. Tari, A. Sanguino, Y. Kondo, S. Kondo, B. Arun, G. Lopez-Berestein, B. Ozpolat, *Autophagy* **2008**, *4*, 669 – 679.

[27] F.-Y. Dupradeau, D. A. Case, C. Yu, R. Jimenez, F. E. Romesberg, *J. Am. Chem. Soc.* **2005**, *127*, 15612 – 15617.

[28] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157 – 1174.

[29] "Molecular Modeling of Nucleic Acids", T. Macke, D. A. Case, *ACS Symp. Ser.* **1998**, *682*, 379 – 393.

[30] a) J. R. Blas, F. J. Luque, M. Orozco, *J. Am. Chem. Soc.* **2004**, *126*, 154 – 164; b) I. Faustino, A. Aviño, I. Marchán, F. J. Luque, R. Eritja, M. Orozco, *J. Am. Chem. Soc.* **2009**, *131*, 12845 – 12853.

[31] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089 – 10092.

[32] J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comp. Phys.* **1977**, *23*, 327 – 341.

[33] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179 – 5197.

[34] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins Struct. Funct. Bioinf.* **2006**, *65*, 712 – 725.

[35] A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham, C. A. Laughton, M. Orozco, *Biophys. J.* **2007**, *92*, 3817 – 3829.

[36] P. Banáš, D. Hollas, M. Zgarbová, P. Jurečka, M. Orozco, T. E. Cheatham, J. Šponer, M. Otyepka, *J. Chem. Theory Comput.* **2010**, *6*, 3836 – 3849.

Received: September 24, 2012
Published online on January 29, 2013

© 2013 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

4.1.3.1 Results and discussion

Development of synthetic oligonucleotides (ONs) has received much attention in the past years due to their ability to silence the expression of undesired overexpressed genes. Among the stablished therapies, siRNAs have been widely studied due to their high potency and sequence-specificity. However, the application of oligonucleotides *in vivo* faces important limitations (Watts et al. 2008; Deleavey & Damha 2012). For example, their high vulnerability to degradation by serum nucleases. Much effort has been made to overcome these limitations but, in many cases, modifications that increase nuclease stability, cause negative effects on RNAi activity. We are interested in creating chemical modifications able to increase nuclease stability without disrupting RNAi activity. In particular, in this work, we explored 3'-terminal modifications by replacing standard nucleobases by N-ethyl-N-coupled nucleosides. We focused special attention on 3'-exonucleases because they are the predominant nuclease activity present in serum. By means of molecular dynamics simulations and *in vitro* experiments, modified siRNAs resistant to nuclease digestion were evaluated against Bcl-2 *in vitro*, an antiapoptotic gene which is overexpressed in several cancer processes.

Effect of 3'-terminal N-ethyl-N modification on naked oligonucleotides Single- and double-stranded oligonucleotides with 3' N-ethyl-N modification were evaluated to check their stability. MD simulations of non-modified ssRNA and its corresponding 3'-modified version yielded stable trajectories with small relative structural deviations related to the presence of terminal modifications, except for some local effect at the neighboring ribose sugar puckering A19.

Effect of 3'-terminal N-ethyl-N modification on hydrolytic activity of 3'-exonucleases We used the structure of the Klenow fragment (KF) of DNA polymerase I from *E. coli* to validate and explore the role of end modifications in the susceptibility of single stranded oligos to the action of 3'-exonucleases. Particularly, we studied the effect of terminal modification in the place of the catalytic moiety in the binding site of the enzyme by means of MD simulations.

The MD simulation performed with the native oligo shows how the Leu361 wedges between the last two 3' nucleobases (T1 and T2 in the native molecule) and the Phe473 stacks with nucleobase T2 (Figure 5.5a). Both residues accommodate the substrate in the active site and fix it to facilitate the hydrolysis. However, the bulky N-ethyl-N modified nucleobases at the 3'-end of the oligonucleotide strand prevents the enzymatic action by a set of different mechanisms. On one side, the Leu361 is unable to stick

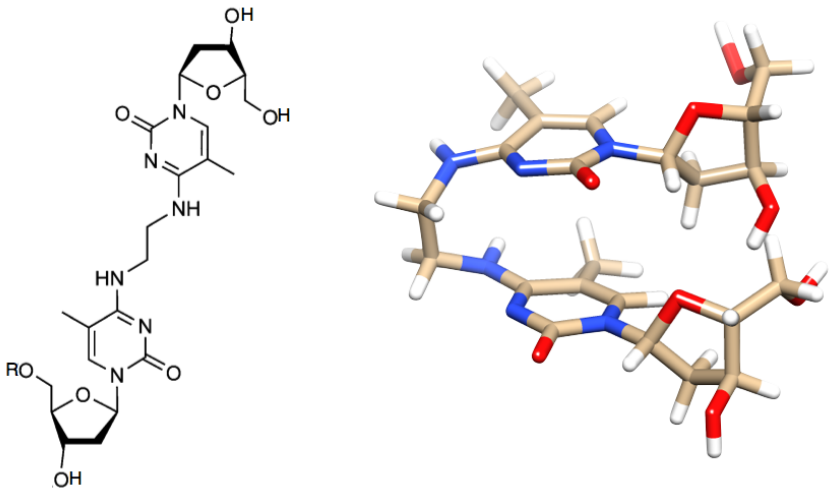


Figure 4.4: Schematic representation of N^4-ethyl-N^4 dimeric pyrimidine nucleosides (left) and stacked conformation of N^4-ethyl-N^4 dimer.

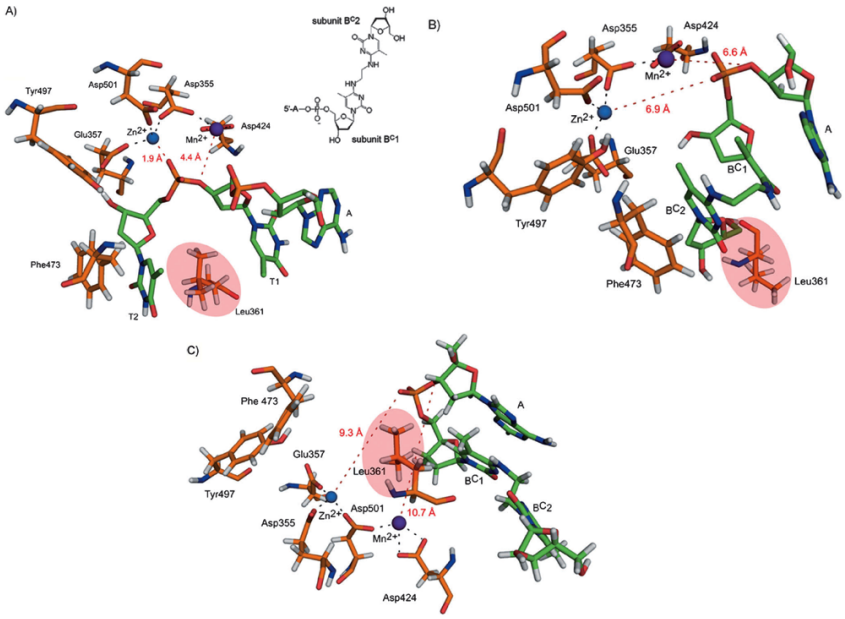


Figure 4.5: Representative snapshots from the MD trajectory (50 ns) showing the position of relevant KF amino acid residues. A) The unmodified DNA trimer ApT1pT2. The B) stacked and C) extended 3'-B^C-modified DNA trimer ApB^C1-ethyl-B^C2.

between the last two nucleobases and falls apart (10.4 Å) from its initial position although the Phe473 which, provides the right conformation by stacking with one of the modified nucleobases, maintains that interaction (Figure 5.5b). On the other side, the lack of a scissile phosphodiester bond close enough to the catalytic metal ions prevents hydrolysis to occur. MD simulations were also run with an extended N-ethyl-N dimer as initial conformation. The corresponding results were similar to the obtained with the stacked dimer conformation (Figure 5.5c) which suggest that both conformations would prevent hydrolysis distorting the binding site of the enzyme.

Nuclease digestion studies both with KF and another common 3'-exonuclease, snake venom phophodiesterase (SNVPD), confirmed the MD results showing strong stability for those 3' N-ethyl-N modified oligonucleotides even after 72h incubation. Comparative studies with phosphorothioate modification (PS) that was previously reported to strongly inhibit 3'-exonuclease activity show poorer blocking activity.

Unmodified and modified single- and double-stranded siRNAs were incubated in human serum. While native ssRNA showed complete degradation after 2h in serum, ∼15% of the initial 3' N-ethyl-N modified ssRNA was still observed after 9h in serum. On the other hand, 3' PS modified ssRNA exhibits almost complete degradation (∼95%) after 6h in serum.

Finally, we showed that the terminal modification does not affect siRNA activities. For this purpose, gene expression studies on MCF-7 human breast adenocarcinoma cells with modified siRNAs were performed to assess the in-cell activity of these modified interference RNAs against the Bcl-2 mRNA. The Bcl-2 mRNA levels after 24h transfection showed similar activities for both wild-type and modified siRNAs (∼93%) confirming the application of the 3' N-ethyl-N modification in RNA interference strategies.

4.1.3.2 Conclusions

This work presents a promising oligonucleotide modification that blocks the 3'→5' degradation by 3'-exonucleases. The modification consists of an N-ethyl-N linker between nucleobases through the opposite side of the lacking phosphate group. These chemical features are the key components that may impede the correct accommodation of the 3'-end modification in the active site. Molecular dynamics simulations suggest that the two 3'-terminal nucleobases and the amino acids at the 3'-exonuclease active site cause a negative effect for the substrate accommodation, since no scissile phosphodiester bond is found in the proximities of the catalytic center formed by coordinated metal ions.

In vitro assays with different 3'-exonucleases showed the resistance inferred by attaching the 3' N-ethyl-N modification to single and double stranded oligonucleotides. Besides this, extensive studies on the susceptibility to enter in the RNAi machinery show that both wild-type and 3' modified siRNAs have very similar activities.

4.1.3.3 References

Beese, L.S. & Steitz, T.A., 1991. Structural basis for the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I: a two metal ion mechanism. The EMBO Journal, 10(1), p.25.

Brautigam, C.A. & Steitz, T.A., 1998. Structural principles for the inhibition of the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I by phosphorothioates. Journal of Molecular Biology, 277(2), pp.363–377.

Brautigam, C.A. et al., 1999. Structures of normal single-stranded DNA and deoxyribo-3'-S-phosphorothiolates bound to the 3‘-5' exonucleolytic active site of DNA polymerase I from Escherichia coli. Biochemistry, 38(2), pp.696–704.

Deleavey, G.F. & Damha, M.J., 2012. Designing chemically modified oligonucleotides for targeted gene silencing. Chemistry & biology, 19(8), pp.937–954.

Eckstein, F., 1985. Nucleoside Phosphorothioates. Annual review of biochemistry, 54(1), pp.367–402.

Eckstein, F., 2000. Phosphorothioate oligodeoxynucleotides: what is their origin and what is unique about them? Antisense and Nucleic Acid Drug Development, 10(2), pp.117–121.

Watts, J., Deleavey, G.F. & Damha, M., 2008. Chemically modified siRNA: tools and applications. Drug Discovery Today, 13(19-20), pp.842–855.

4.2 Flexibility of nucleic acids

4.2.1 Toward a consensus view of duplex RNA flexibility.

Ignacio Faustino, Alberto Pérez, and Modesto Orozco.
Biophysical Journal. 2010, 99(6), pp.1876–1885.

Toward a Consensus View of Duplex RNA Flexibility

Ignacio Faustino,[†] Alberto Pérez,[†] and Modesto Orozco[†‡§*]

[†]Joint Institute of IRB/BSC Program on Computational Biology, Institute of Research in Biomedicine, Barcelona, Spain and Barcelona Supercomputing Centre, Barcelona, Spain; [‡]National Institute of Bioinformatics, Barcelona, Spain; and [§]Departament de Bioquímica, Facultat de Biología, Barcelona, Spain

ABSTRACT The structure and flexibility of the RNA duplex has been studied using extended molecular dynamics simulations on four diverse 18-mer oligonucleotides designed to contain many copies of the 10 unique dinucleotide steps in different sequence environments. Simulations were performed using the two most popular force fields for nucleic acids simulations (AMBER and CHARMM) in their latest versions, trying to arrive to a consensus picture of the RNA flexibility. Contrary to what was found for DNA duplex (DNA_2), no clear convergence is found for the RNA duplex (RNA_2), but one of the force field seems to agree better with experimental data. MD simulations performed with this force field were used to fully characterize, for the first time to our knowledge, the sequence-dependent elastic properties of RNA duplexes at different levels of resolutions. The flexibility pattern of RNA_2 shows similarities with DNA_2, but also surprising differences, which help us to understand the different biological functions of both molecules. A full mesoscopic model of RNA duplex at different resolution levels is derived to be used for genome-wide description of the flexibility of double-helical fragments of RNA.

INTRODUCTION

Nucleic acids exist in nature as two main polymers (DNA and RNA), which despite having quite similar chemical composition display quite different structure and very different biological function. While DNA carries genetic information and is usually found as a right-handed double helix, RNA is much more versatile and can display very different secondary and tertiary structures, allowing it to engage in a very different range of biological functions, from carrying genetic information to gene regulation or catalysis (1–4). DNA is synthesized to define a perfectly paired self-complementary duplex (DNA_2), where one strand recognizes the other by means of A·T and G·C Watson-Crick pairings (5). RNAs, in developed organisms, exist mostly as single strands which adopt compact structures where the strand recognizes itself to maximize the amount of duplex, which (when possible) will be formed by Watson-Crick A·U and G·C pairings. For some cases, as in that of microRNA, such hairpins are processed to generate pure antiparallel duplexes—which are then recognized by microRNA processing proteins (6) in having a key role in the control of cell function.

It has been known since the 1950s that a right-handed duplex with 10 basepairs per turn (known as the "B-form") is the most stable conformation for DNA_2 under physiological conditions, and that a more compact ,11 basepairs per turn, right-handed duplex (known as the "A-form") is the preferred conformation for the RNA duplex (RNA_2). RNA_2 is more stable than DNA_2, except in the case of sequences very rich in A and T(U) pairings, where it has been shown (7) that DNA_2 can be more stable. Structural analysis shows that DNA_2 can display a large variety of structures close to the B-form (8), which combined with the larger polymorphism of the DNA backbone (S↔E repuckering or BI↔BII transitions) has been traditionally used to support the idea that canonical B-DNA duplex is more flexible than RNA_2 (and A-DNA_2). However, as already discussed elsewhere (9), flexibility is a quite dangerous concept with little meaning when disconnected from the geometrical perturbation used to define it, and certainly this interpretation of experimental data is not unique (see below).

Molecular dynamics (MD) simulations using state-of-the-art simulation conditions and last-generation force fields are a perfect complement to experimental techniques in the definition of nucleic acids flexibility (9,10). Thus, different groups (11) have made extensive analysis of the flexibility properties of DNA_2 (12–14), which has provided very valuable descriptors to help us understand not only gene structure, but also regulatory mechanisms or chromatin organization (15–18). These studies have provided detailed information on fine details of DNA_2 deformability, such as the sequence- (at the basepair step level) and perturbation-dependent stiffness of DNA_2 (19,20). Massive multigroup projects are currently under development (11) to obtain a refined database of the stiffness of the DNA duplex considering the local helical deformations at the tetramer level, which can yield refined parameters for mesoscopic modeling of DNA. In comparison with the large amount of information for DNA_2 flexibility derived from MD simulations, little is known for RNA_2 (9,10,21–24). Recently, using the AMBER force field, our group studied a 12-mer sequence of DNA_2 and RNA_2, finding that in global terms, RNA_2 was more rigid than DNA_2—mostly due to a higher backbone flexibility for DNA (9,25,26). However, that situation reverses when only the very first deformation modes

Submitted April 22, 2010, and accepted for publication June 25, 2010.

*Correspondence: modesto@mmb.pcb.ub.es

Editor: Kathleen B. Hall.

© 2010 by the Biophysical Society

0006-3495/10/09/1876/10 $2.00

doi: 10.1016/j.bpj.2010.06.061

were considered. Findings supported by a parallel database analysis (23) suggest that the DNA_2 deformation space is wide and complex, while that of RNA_2 is narrow and simple, becoming dominated by a very few numbers of very low-frequency movements.

These conclusions have been challenged in an even more recent work by Priyakumar and MacKerell (22), who presented carefully analyzed CHARMM-based data pointing in an opposite direction: i.e., that RNA_2 is in global terms more flexible than DNA_2. To clarify this issue and, for the first time to our knowledge, obtain a full characterization of the sequence-dependent properties of RNA_2, we decided to perform a CHARMM27 (27,28) and AMBER-ff99/Parmbsc0 (29–31) study of several long and representative RNA and DNA duplexes to derive the corresponding flexibility descriptors. The desired objective of our work was the derivation of a consensus picture of RNA duplex flexibility, as we did previously for proteins (32,33) and DNA duplex (34). Unfortunately, results reported here show that, contrary to the situation with parent DNA_2s, no convergence between force fields has been reached and that extra caution is required when deriving conclusions from MD simulations on RNA duplexes. Analysis of long trajectories (150 ns) for four 18-mer duplexes containing a variety of sequences (18,34) reinforce our confidence in the results obtained with the Parmbsc0 revision of the AMBER ff99/Parmbsc0 force field (in the following noted simply as Parmbsc0), while some aspects of the structural and flexibility patterns reported by CHARMM27 simulations seem difficult to fit to available experimental data—supporting previous claims of different groups on CHARMM27-based simulations of RNA duplexes (35,36). After a careful study of all trajectories and extensive comparison with available experimental data, a first atlas of the flexibility of RNA_2 is derived. Such an atlas can be used to describe the flexibility of RNA duplexes as well as to understand and quantify, in a fast and efficient way, important aspects of RNA biology, such as the indirect-recognition mechanisms for protein binding.

METHODS

System selection

To make conclusions as general as possible, and following previous works (34), four sequences of DNA_2 and RNA_2 were selected, namely,

SEQ1: x(GCCYAYAAACGCCYAYAA)·x(YYAYAGGCGYYYAYAGGC),
SEQ2: x(CYAGGYGGAYGACYCAYY)·x(AAYGAGYCAYCCACCYAG),
SEQ3: x(CACGGAACCGGYYCCGYC)·x(GACGGAACCGGYYCCGYG),
SEQ4: x(GGCGCGCACCACGCGCGG)·x(CCGCGCGYGGYGCGCGCC),

where x stands for ribo (r) or deoxyribo (d) backbones and Y stands for either U (RNA) or T (DNA)). These four sequences are sufficiently long to reduce the importance of end-effects in simulations and contain many copies of the 10 unique base steps, namely,

(x(GG)·x(CC),
x(GC)·x(GC),
x(GA)·x(YC),
x(GY)·x(AC),
x(AG)·x(CY),
x(AA)·x(YY),
x(AY)·x(AY),
x(CG)·x(CG),
x(CA)·x(YG),
x(YA)·x(YA),

where x is equal to r or d and Y equal to T or U, making possible a statistically reliable analysis of the sequence-dependence in duplex flexibility.

System preparation and production runs

Following the protocol described elsewhere (37–39), all the systems were neutralized by adding a suitable number of Na^+ ions to be surrounded by ~9200 TIP3P water molecules. Solvent boxes were manipulated to guarantee that an equal number of water molecules exist for DNA and RNA duplexes of the same sequence. All the systems were partially optimized, thermalized, and equilibrated using a standard multistep procedure (38,40). RNA_2 simulations were equilibrated by 10 ns before the 150-ns production runs, while DNA_2 simulations started from the end of previous 100-ns trajectories (34) and were extended for an additional 50 ns to complete 150 ns trajectories, which, based on our experience with microsecond-long simulations (39), should be large enough as to capture well the near-equilibrium dynamic properties of the different duplexes. All simulations were performed in the isothermal-isobaric ensemble (T = 298 K, P = 1 atm) using periodic boundary conditions and particle-mesh Ewald treatments to account for long-range electrostatic effects (41). SHAKE (42) was used to maintain all bonds involving hydrogen atoms at their equilibrium values, which allowed the use of a 2-fs integration step. Simulations were carried out with NAMD (43) and PMEMD (44) computer programs (see details in the Supporting Material) after checking carefully (see Fig. S1 in the Supporting Material) that no differences can be expected from the use of these two different MD codes.

Analysis of trajectories

The 150-ns-long trajectories of the four DNA and four RNA duplexes were processed to obtain information on the structural and flexibility properties of both nucleic acids at global and local levels. Structural analysis was performed using standard procedures (45) on the central 14-mer portion using CURVES+ (46) for the helical analysis, the *ptraj* module of AMBER9 for energetic analysis, and VMD (47), as well as in-house programs. The expected pattern of interaction of the average DNA and RNA duplexes was determined from classical molecular interaction potentials (48). The global deformability of DNA and RNA duplexes was characterized by means of entropy calculations using pseudoharmonic modes (49,50) (see Methods in the Supporting Material). The global patterns of deformation of duplexes were studied by means of essential dynamic algorithms adapted to nucleic acids (14,37) using the same atom-compression rules as for entropies when different duplex types of different sequences were compared. The essential dynamic analysis processes the Cartesian coordinates compiled from the dynamics into a set of eigenvectors ($\{v_i\}$) and eigenvalues ($\{\lambda_i\}$); the first provides information on the nature of the essential deformation movements, while the second informs on the stiffness associated with such deformations (K_i),

$$K_i = \frac{k_B T}{\lambda_i}, \quad (1)$$

where k_B is the Boltzmann's constant and T is the temperature.

Comparison between the essential deformation modes in two trajectories (A and B) was performed using similarity indexes (14) (see the Supporting Material) that were computed for a common set of atoms.

The stiffness associated to pure global deformations (Θ) was determined by inversion of the associated variance,

$$K_{\Theta} = \frac{k_B T}{\left\langle (\Theta - \Theta_0)^2 \right\rangle}, \quad (2)$$

where $\langle \ldots \rangle$ stands for a Boltzmann's averaging and the index 0 refers to the equilibrium value of the global coordinate. Four global helical values were considered (tilt, roll, stretch, and twist) for all the possible fragments of the central 14-mer of the four DNA and RNA duplexes. These descriptors were combined to get the most commonly used global stiffness parameters: the stretch modulus (C), the twisting (T), and the bending (B) persistence lengths. For bending (which has roll and tilt contributions), both isotropic and anisotropic persistence lengths were computed. Additional technical details on the methodology used to derive global stiffness indexes can be found in Lankas et al. (20).

The analysis of local helical flexibility was carried out from the study of the stiffness matrix ($\in$) associated to helical deformations (twist (w), roll (r), tilt (t), rise (s), slide (l), and shift (f)) at the dinucleotide level, and determined by inversion of the MD-associated covariance matrix ($\subset$),

$$\Xi = E(\Delta X)^{-2} = k_B T \subset^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{wr} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{wt} & k_{rt} & k_t & k_{st} & k_{tl} & k_{tf} \\ k_{ws} & k_{rs} & k_{st} & k_s & k_{ls} & k_{lf} \\ k_{wl} & k_{rl} & k_{tl} & k_{ls} & k_l & k_{lf} \\ k_{wf} & k_{rf} & k_{tf} & k_{lf} & k_{lf} & k_f \end{pmatrix}, \quad (3)$$

where E is the energy associated to the deformation ΔX and k stands for the different stiffness constants defining the 36 elements of the stiffness matrix.

Stiffness parameters for a given dinucleotide step were computed individually for each sequence and averaged later to avoid the derivation of artifactual soft parameters due to nonneighbor effects on equilibrium geometry (11,34). By using this procedure, the stiffness parameters for a given step agree well independently of the dinucleotide environment, suggesting that stiffness parameters are less dependent on remote neighbors than equilibrium geometry parameters.

Analysis of structural databases

The structures of all DNA and RNA duplexes deposited on the 2009 version of NDB (23,34,51) were filtered to define a set of naked DNA_2 and RNA_2 from which we derived dinucleotide structural data. Thus, we eliminated from the database all oligos bound to proteins or drugs as well as those containing mismatchings and noncoding nucleobases or those solved with poor quality (resolution <3.5 Å). To reduce the impact of packing effects, all oligos shorter than eight basepairs were also eliminated. The remaining duplexes were studied using helical descriptors at the dinucleotide level identical to those used in trajectory analysis, as suggested by Olson et al. (15). Dinucleotide steps showing extreme helical values (outside three standard deviations from the average of any of the six helical parameters) were excluded from the analysis, because these deformations cannot be explained with the harmonic model considered here (a very small number of cases were eliminated due to these criteria). Note that the experimental data derived in this way can be reasonably accurate in terms of average values, but caution is necessary with the standard deviations (from which stiffness parameters are derived). This is due to scarcity of experimental data for several dinucleotide steps and additional errors in experimental estimates that are expected from systematic biases in refinement used considering simple force fields with reduced rough experimental information.

RESULTS AND DISCUSSION

Reliability of RNA trajectories

A crucial point in any MD study is the validation of the trajectories in terms of their ability to reproduce known experimental data. The 8 × 150 ns RNA_2 trajectories studied here were stable, sampling in all cases helical (or pseudohelical) structures. However, contrary to the situation found in the DNA duplex, where small deviations from the canonical helical conformation were found (Fig. S2 and Fig. S3) and excellent convergence between Parmbsc0 and CHARMM27 simulation was achieved, nonnegligible differences are found here between both RNA force fields. Significant structural alterations are evident in some of the CHARMM27 simulations (specially for seq2 and seq4), whose root-mean-square deviation (RMSD), with respect to canonical A-form, becomes very large (Fig. 1), even when the floppy ends are eliminated from the RMSD calculation (Fig. 1 and Fig. S4). Graphical analysis of the structure evolution clearly illustrates that CHARMM27 leads to local unfoldings of the helix, which are reversible in some cases but irreversible in many others (see Figs. 2 and 3 and Fig. S5, Fig. S6, and Fig. S7). The distortions detected in these simulations are very large in terms of twist for some steps, which take often completely artifactual values (Fig. S7), leading to a complete lost of helicity in some fragments of the structure (see Fig. 2) and dramatic changes in the major groove geometry (see Fig. S5). Local base geometry is largely altered in CHARMM27 RNA_2 simulations, affecting both r(A·U) and r(G·C) pairs with either temporary or permanent losses of interstrand hydrogen bonds (see examples in Fig. 3 and Fig. S6), which leads to major losses of helicity in the opening region (something that is not detected in Parmbsc0 simulations). It is worth noting here that reversible base opening is not necessarily an artifactual behavior, but the process is experimentally known (7,52) to happen in the millisecond (and not in the nanosecond) timescale, and the population of opened conformations is expected to be very small at room temperature. Finally, it must be noted that, while severe distortions happen in some parts of the helix, other regions with similar composition remain close to the expected conformation—indicating the stochastic nature of the distortion process found in these calculations, and the fact that, for some small regions, CHARMM27 results could be used to derive local helical parameters.

The four 150-ns trajectories of RNA_2 done with the Parmbsc0 force field yield very stable trajectories, with sampling conformations close to the expected A-form (Figs. 1 and 2 and Fig. S4). Hydrogen-bond pattern is strictly maintained (Fig. 3), with the exception of some fraying movements at the ends (especially in r(A·U) steps), agreeing with experimental findings on the stability of RNA_2 hydrogen bonds and on the very slow kinetics of basepair opening (7,52). No local unfolding, untwisting,

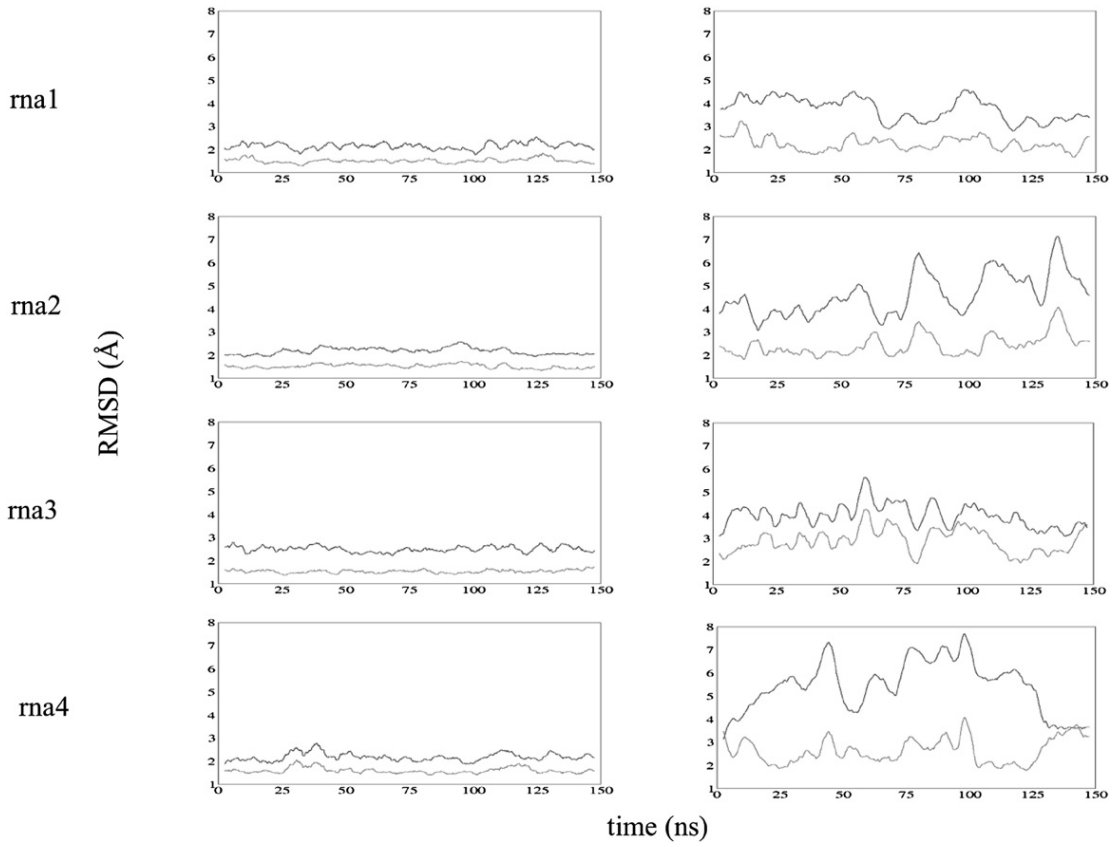


FIGURE 1 Smoothed RMSD (in Å) from A-RNA fiber conformation (in *gray*) and average structure (in *black*) for RNA/Parmbsc0 (on the *left*) and CHARMM27 simulations (*right*) for the central 14-mer of the four RNA sequences.

or any other artifactual distortion of the helix is detected during the four simulations reported here (Fig. 2). In summary, all our analysis suggest that Parmbsc0 simulations of RNA duplexes sample near-equilibrium geometries close to those expected for the corresponding sequences

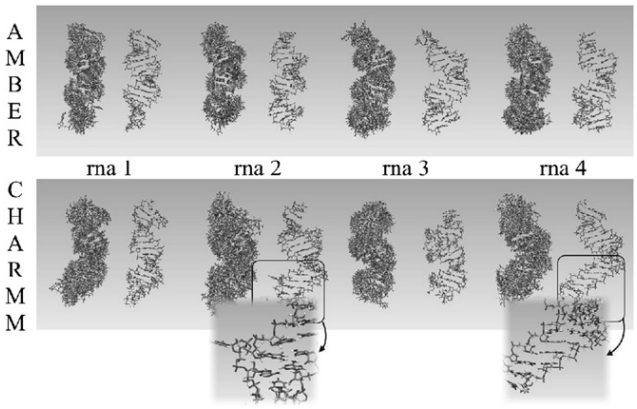


FIGURE 2 Ensemble at different times of simulation (on the *left*) and final (on the *right*) structures of the four RNA sequences for Parmbsc0 (at the *top*) and CHARMM27 (at the *bottom*) simulations. In the case of CHARMM27 simulations, some fragments undergo irreversible nonhelical transition. (*Insets*) Detailed images of some of the major distortions.

based on the known experimental evidence, and can be used safely to derive mechanical descriptors of RNA_2.

General structure of the RNA duplex

The four Parmbsc0-derived trajectories were analyzed to obtain a representation of the average structural properties of the RNA duplex in solution. The RNA_2 helix appears as a compact structure, with helical parameters very close to those of crystal structures (see Fig. S8 and Table S1), except for twist—where the force field underestimates experimental values at ~3° (as happens for DNA_2 with the same force field). Analysis of backbone conformations reports results similar to those expected for an A-duplex as found in experimental structures (Fig. S9). Sugar puckering is strictly fixed in the North region for the four RNA/Parmbsc0 duplexes, agreeing well with available experimental data (see Fig. S9). Very interestingly, the variability in backbone conformation obtained in RNA_2 simulations seems smaller than that detected during equivalent DNA_2 simulations (see Fig. S9), mainly due to the restricted sampling around the δ-, ζ/ε-, and α/γ-torsions, pointing to the higher rigidity in the RNA backbone. Clearly, CHARMM27 structures should be taken with

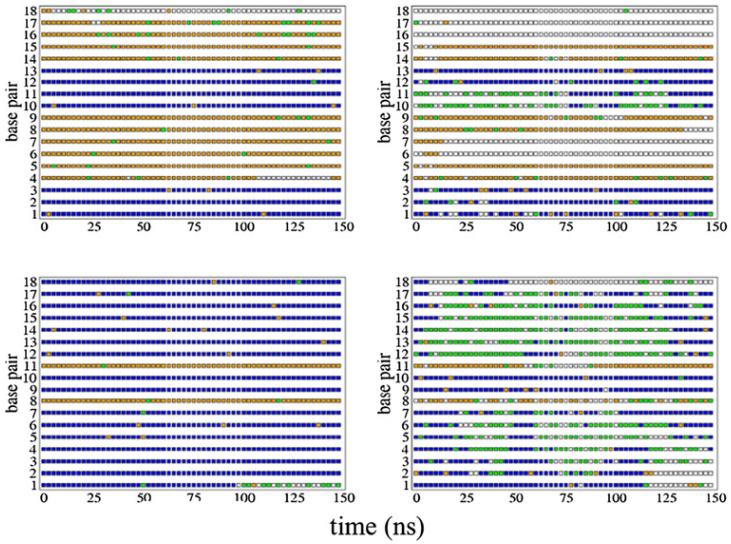


FIGURE 3 Comparison of averaged interstrand hydrogen-bonding interactions for every basepair along sequences 1 (on the *top*) and 4 (at the *bottom*) and along the time of simulation for Parmbsc0 (on the *left*) and CHARMM27 trajectories (*right*) for RNA. (*Blue* means three standard Watson-Crick hydrogen bondings; *orange* stands for two of them; *green* means only one hydrogen interaction; and *white*, no standard interaction between one base and its complement.)

caution due to the existence of abnormal distortions. However, if the dinucleotide analysis is performed using CHARMM27 trajectories from which all the steps affected by base opening (i.e., those losing hydrogen bonds and their neighbors during at least 50% of the time of simulation) are removed, the helical equilibrium results are indistinguishable from the Parmbsc0 ones (see Table S1), reinforcing the confidence in our theoretical results.

Di-nucleotide equilibrium geometries

Despite the uncertainties implicit with the use of the nearest-neighbor model (11), dinucleotide equilibrium helical parameters appear as a simple and intuitive approach to roughly characterize sequence-effects on duplex geometry. Analysis of Parmbsc0 trajectories allowed us to derive average helical equilibrium geometries for each of the 10 unique RNA_2 dinucleotide steps in 3–8 different sequence environments (see Fig. 4). With the data presented here is not possible to perform a systematic evaluation of nonneighbor effects (4- or 6-mer) in di-nucleotide geometry, as done for DNA in a recent article (11). However, analyses of data suggest that the influence of the environment on the equilibrium geometry of di-nucleotides is moderate. For example, no bimodal distributions are found in any of the cases, and for the analysis of the four-dinucleotide step where we have at least four different environments, standard deviations in average are very small, even for twist (standard deviations range between 0.4 and 0.7°; AC: ± 0.5; CG: ± 0.4; CC: ± 0.7; GA: ± 0.5°).

Analysis of Fig. 4 strongly suggest that sequence effects even smaller that in DNA_2 (see Fig. S10), are still quite significant for RNA_2, as shown in ranges of sequence-dependent variability of 6° or 11° in twist and roll. Profiles obtained with Parmbsc0 agree pretty well with data derived from structures solved experimentally, even though caution is needed with the latter because some of the steps are poorly represented in the experimental databases (for example, AU average was taken from only 15 steps extracted from only 13 different duplexes; see Table S1). Also interestingly, CHARMM27 results also agree well with Parmbsc0 data—provided open steps are neglected from the analysis, reinforcing the hypothesis that opening is the main reason for structural distortion in CHARMM27 simulations of RNA duplexes. Finally, it is worth noting that some sequence-dependent geometry profiles (like twist and roll) are similar in DNA_2 and RNA_2, but most of them are clearly different, suggesting that backbone restrictions influence significantly the arrangement of nucleotide pairs in the duplexes and that sequence-effects are not orthogonal to backbone effects.

Molecular interactions

Classical molecular interaction potentials (48) (see Fig. S11) allowed us to trace the different interaction properties of DNA_2 and RNA_2 for a common sequence. As expected

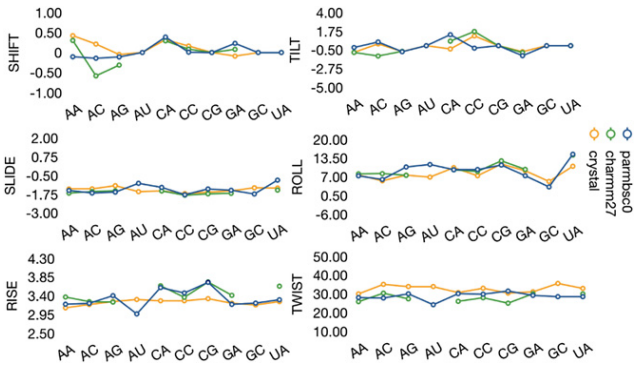


FIGURE 4 Average helical parameters for RNA for the 10 unique representative basepair steps for translational (shift, slide, and rise; in Å) and rotational (tilt, roll, and twist; in degrees) parameters. The CHARMM27 values were taken after removing open steps.

from previous studies (9) a dramatic change in preferred interaction sites is detected in RNA_2 compared to DNA_2, probably due to the different geometries of the grooves (Fig. S11). Such a change is reflected in a complete alteration of the pattern of hydration. Thus, the central 14-mer RNA duplex binds (see Table S2), on average, ~13 water molecules more than the equivalent DNA duplex; 26.6 water molecules per basepair in DNA_2 and 27.6 for RNA_2. The presence of the polar 2′ hydroxyl group and the large electronegative cloud in the major groove are the main factors responsible for the higher hydrophilicity of RNA_2.

Global flexibility

The concept of flexibility is not well defined at either the global or the local levels, but a reasonable approach is to consider that one duplex (of a given length) is more flexible than another one when, for a common set of atoms, it shows higher intramolecular entropy. Parmbsc0 results in Table 1 clearly demonstrate that DNA_2 is more flexible than RNA_2 due to the greater deformability of its backbone, something that could already be expected from the dihedral analysis noted above (see Fig. S9). Analysis of the frequencies associated with each deformation movement (see Fig. S12) reveals that the larger flexibility of DNA_2 is due to the higher order deformation modes, because the lower modes, those explaining a larger percentage of variance, are in fact very similar for RNA_2 and DNA_2 (a result already found in previous simulations (9) and which is also clear from the inspection of the stiffness associated to the different deformation modes (Fig. S12)). These findings confirm our previous claims that while RNA_2 has a quite simple dynamics and explores very deeply only a small number of modes, the DNA_2 explores a much larger number of deformation modes.

According to Parmbsc0 calculations, major deformability of RNA_2 is dominated by twisting and bending motions, which are similar, but not identical to those of the corresponding DNA duplexes as noted in similarity indexes (see Table S3). Contrary to the situation in DNA_2, where similarity between Parmbsc0 and CHARMM27 was very high (similarity indexes at ~0.9), there is a mediocre correspondence between both force fields for most RNA_2 duplexes considered here (see Table S4). Clearly the large distortions in some CHARMM27 trajectories are responsible for the reduced similarity between the flexibility patterns computed from the two force fields.

As an additional analysis of the global flexibility of RNA duplexes, we derived global helical stiffness parameters from the oscillation of helical parameters (for the 14-mer central duplex (see Methods)). The parameters obtained by this analysis are quite robust to the length of the oligo considered (see Fig. S13) and provides quite intuitive information on the global helical flexibility, which can be directly compared with experimental measures of stress. The results (again only Parmbsc0 results make sense here) shown in Table 2 indicate that the deformations of the global roll, tilt, and twist for RNA_2 are more difficult than for DNA_2. This is in agreement not only with previous theoretical suggestions (9) but with experimental results (53–55) suggesting a bending persistence length at ~54 ± 2 nm for DNA and 20–30% higher in the case of RNA duplex, which confirms the reliability of our results on short oligos. However, it is also worth noting that DNA_2 becomes stiffer than RNA_2 for global stretch deformations, which again warns against a too-simplistic or too-general use of the concept "flexibility."

Sequence-dependent flexibility

The analysis of trajectories by basepair step resolution harmonic models (see Methods) provides very useful descriptors of local flexibility which can be integrated so

TABLE 1 Intramolecular entropies (in kcal/mol·K) for DNA and RNA (*italics*)

	Schlitter and Klähn method	Andricioaei and Karplus method
Seq 1	3.35	3.07
	2.82	*2.54*
Seq 2	3.31	3.03
	2.90	*2.62*
Seq 3	3.37	3.09
	2.93	*2.65*
Seq 4	3.33	3.05
	2.95	*2.66*
Averages	3.34	3.06
	2.90	*2.62*

Performed for 150-ns simulation time computed with the Schlitter and Klähn (49) and Andricioaei and Karplus (50) methods, taking into account the all common atoms of the central 14-mer for the four sequences considered.

TABLE 2 DNA and RNA (in *italics*)

	Bending (nm)			Stretching (pN)	Twisting (nm)
	Roll	Tilt	Isotropic		
Seq 1	64.99	55.67	59.97	2318.36	86.38
	77.45	*71.74*	*74.49*	*1662.39*	*177.19*
Seq 2	71.83	46.38	56.37	2540.19	93.05
	102.13	*62.32*	*77.40*	*1745.22*	*191.66*
Seq 3	74.35	37.11	49.51	2262.47	90.93
	105.67	*67.57*	*82.43*	*1860.05*	*198.35*
Seq 4	66.49	57.83	61.86	2475.36	81.34
	99.67	*71.42*	*83.21*	*1448.42*	*195.18*
Averages	69.42 ± 4.41	49.25 ± 9.49	56.93 ± 5.45	2399.09 ± 30.27	87.93 ± 5.20
	96.23 ± 12.76	*68.26 ± 4.39*	*79.38 ± 4.16*	*1679.02 ± 173.79*	*190.60 ± 9.35*

Bending (anisotropic and isotropic) (B) and twisting (C) persistence lengths (in nanometers) and stretch modulus (S, in picoNewtons) for every sequence, and the corresponding averages and standard deviations from Parmbsc0 simulations. Analysis done by taking the values from sequences of 11 basepairs.

as to derive complete pictures of flexibility of long DNA fragments (14,56–58). As noted above, we cannot reach consensus values here due to the distortions detected in CHARMM27 trajectories, but it is worth noting that if snapshots with open steps are eliminated from CHARMM27 trajectories (see above), the derived stiffness parameters are not far from those obtained with Parmbsc0 simulations (see Fig. 5). This, and the good agreement with the few stiffness parameters that can be derived, with statistical significance, from the geometrical variability in experimental structure (only steps with at least 40 structures were used here), allows us to conclude that even without fill consensus, our Parmbsc0-derived stiffness parameters are likely to be very reliable descriptors of RNA duplex flexibility. Note that local flexibility descriptors presented here are based in the nearest-neighbor approximation, and accordingly might be contaminated by nonneighbor effects (4- or 6-mer). However, analysis of our data show that:

1. No sequence-dependent bimodality as that found in DNA_2 is detected in RNA duplex, and
2. Di-nucleotide force constants (stiffness parameters) are rather robust (in general <9% variation) to change 4- or 6-mer effects (as in fact happens for DNA_2 where the 4- or 6-mer effects are focalized in the equilibrium values rather than in the stiffness parameters).

Accordingly, with all the required cautions, we believe that Parmbsc0 estimates of the stiffness parameters can be used to describe RNA_2 deformability.

Stiffness parameters (Fig. 5, original numerical data in Table S5 and Table S6) derived from Parmbsc0 trajectories reveal the large sequence-dependence of local flexibility in RNA_2, as noted in variation of up to 400% in local helical stiffness parameters. Certain steps are intrinsically rigid (like GC or CC), while others seem very flexible (like AU and UA), but in general, the concept of rigid and flexible steps should be taken with extreme caution, because the relative deformability depends dramatically on the nature of the deformation considered. For example, AC is stiffer than CC for rise deformation but simultaneously much softer for tilt deformation (see Fig. 5 and Table S5). Once again, the concept of flexibility and rigidity without link to the nature of the deformation movement is meaningless. It is worth noting that the relative ordering of stiffness for the different steps in RNA_2 is very similar to that found previously for DNA_2 (see Fig. 5), with the exception of r(A·U) steps, which are unexpectedly soft in RNA_2, probably due to the lack of the 5-Me group. For all the steps (except in the d(AT)/d(AU) pair), RNA_2 is stiffer than DNA_2 both in terms of rotational and translational deformations, in agreement with the results found in global helicity and entropy analysis. However, the analysis of Fig. 5 reveals that the difference in DNA_2 versus RNA_2 stiffness is smaller than the sequence-related variability, which implies that it could be possible to design sequences of RNA_2 with softer than average DNA_2 sequences.

Intrapair flexibility

An important, but often neglected pattern of flexibility is related to the relative movement of paired nucleobases. This is defined by three relative translations (stagger, stretch, and shear) and rotations (propeller twist, opening, and buckle). Of particular interest is the analysis of opening for two reasons:

1. Most of Parmbsc0 versus CHARMM27 differences arise from massive openings when the latter force field is used; and
2. NMR data (7,52) suggest that opening is often easier for RNA_2 than for DNA_2, which could be interpreted as a support to claims that RNA_2 is more flexible than DNA_2.

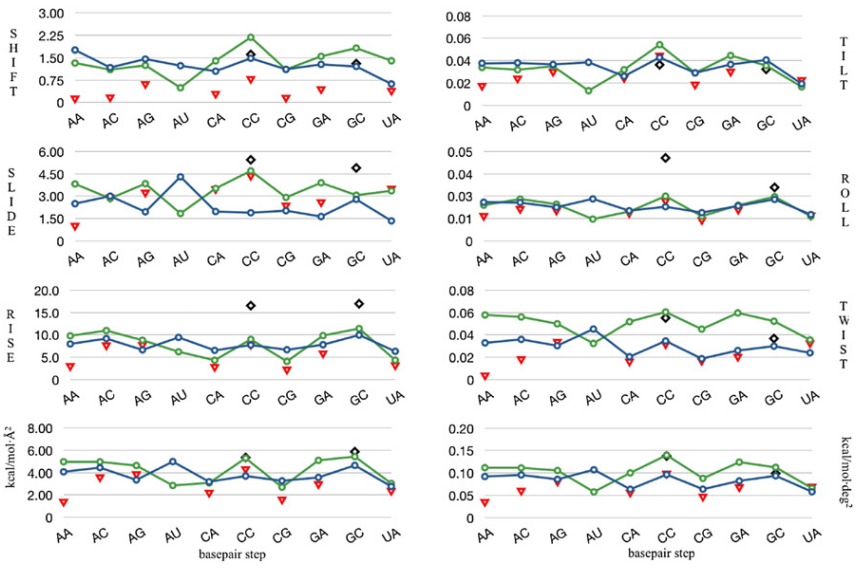


FIGURE 5 Stiffness constants (translational ones in kcal/mol·Å^2 and rotational ones in kcal/mol·deg^2) for the 10 representative dinucleotide steps associated to the different deformation modes comparing DNA/Parmbsc0 (in *blue*), RNA/Parmbsc0 (in *green* with *lines*), RNA/CHARMM27 (in *red triangles*), and derived for analysis of x-ray structural data (in *black diamonds*) values. (*Bottom*) Summation of stiffness constants for translational helical parameters (*left*), and the same for rotational helical parameters (*right*).

TABLE 3 Opening force constants and their averages

	Seq 1		Seq 2		Seq 3		Seq 4		Averages	
d(A·T)	0.017	*0.015*	0.017	*0.016*	0.020	*0.019*	0.022	*0.022*	0.019	*0.018*
r(A·U)	0.016	*0.005*	0.017*	*0.003*	0.019	*0.003*	0.015	*0.004*	0.017	*0.003*
d(G·C)	0.068	*0.051*	0.063	*0.047*	0.059	*0.044*	0.064	*0.048*	0.063	*0.047*
r(G·C)	0.062	*0.003*	0.057	*0.007*	0.052†	*0.006*	0.056	*0.003*	0.058	*0.005*

Measured in kcal/mol deg^2 for d(A·T), r(A·U), d(G·C), and r(G·C) for the four sequences. The ending two basepairs have been removed for the analysis. Values corresponding to Parmbsc0 simulations are in roman style while those corresponding to CHARMM27 are in italics.
*Basepair 3 was removed due to fraying effects.
†Basepair 16 was removed due to fraying effects.

Results in Table 3, which are integrated for all the A·T / A·U and G·C steps, indicates that while CHARMM27 and Parmbsc0 opening force-constants are reasonably close for DNA_2 (especially for d(A·T) steps), they are completely different for RNA_2. This reflects the tendency of CHARMM27 simulations to open the pairs in the multi-nanosecond timescale, something that contrasts with NMR-measured average base opening times (~1 ms; (7,51)). Clearly in this case, analysis should be restricted to Parmbsc0 values, which shows very clearly that in any sequence, environment openings of basepairs is easier (by ~10%) for RNA_2 than for DNA_2. Note that this finding seems counterintuitive, because RNA helices are generally more stable than DNA_2, but agrees well with previous quantum mechanical calculations (23,59), which suggested that the general conformation of the RNA duplex induces a slight reduction in stability of the purine·pyrimidine hydrogen bonding compared with DNA_2. There is also beautiful agreement with NMR data (7,52), which points toward an easier opening in RNA than in DNA duplex, despite the globally larger stiffness of RNA_2. Our simulations are underlying again the complexity and richness of the concept of flexibility and the danger of making general claims about flexibility based on a single physical descriptor.

CONCLUSIONS

Contrary to the situation in DNA duplex, where the two most widely used force fields provide similar structural and mechanical information, no general consensus picture of RNA_2 can be reached (to our knowledge), because of the large distortions occurring in RNA duplexes simulated with the CHARMM27 force field (which seems to be related to the loss of hydrogen bonding during simulations). It is, however, worth noting that if corrupted segments are eliminated from the analysis, reasonable agreement is found between CHARMM27 and Parmbsc0 simulations—pointing to a potential convergence between force fields in the near future.

Extreme caution is required when talking about flexibility, because this depends on the sequence and the level of resolution considered (global structure, base steps, or basepairs). However, if entropy or global stiffness parameters are accepted as global measures of flexibility, the DNA duplex appears clearly as an overall more flexible structure. At the basepair step level, the situation becomes more complicated, because even if DNA_2 is in general more flexible, sequence-dependent variability can induce larger changes in flexibility/rigidity than those originated by the nature of the oligonucleotide considered. Very interestingly, upon examining this at the basepair level, some deformation movements, such as opening, can be more difficult for DNA_2 than for RNA_2. Such finding is in excellent agreement with experimental and high level QM data. However, it is also quite counterintuitive, when one considers the overall larger rigidity and stability of the RNA duplex.

Parmbsc0 simulations allowed us to derive (to our knowledge), for the first time, not only global helical stiffness parameters (which agree well with experimental data), but sequence-adapted stiffness (and equilibrium) parameters for the RNA duplex. Results reveal a dependence of physical deformability with the sequence, which is similar, but not identical to that found in DNA_2. The apparently minor change T→U from DNA_2 to RNA_2 leads to major changes in the relative flexibility of the A·X step (X = T or U) between the two nucleic acids, which raises an interesting hypothesis on the motivations of nature for use different pyrimidines in DNA and RNA and suggest intriguing hypothesis on the physical mechanisms of epigenetic gene control mechanisms. Mesoscopic parameters presented here for the first time can be used to describe near-equilibrium geometric deformations of any RNA structure showing double-helix fragments, as is the case of interference RNA binding to the RISC complex. This opens the field for correlation of sequence effects with biological properties of RNA duplexes.

SUPPORTING MATERIAL

Thirteen figures, six tables, and two equations are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)00805-2.

The authors are indebted to the Barcelona Supercomputing Center for computational resources. We also thank Drs. Carlos González and David Torrents for many useful discussions.

This work was supported by the Spanish Ministry of Science (grant No. BIO2009-10964), the Consolider E-Science project, the COMBIOMED ISCIII project, and the Fundación Marcelino Botin. I.F. is a PhD fellow of the Spanish Ministry of Science and A.P. is a Juan de la Cierva fellow.

REFERENCES

1. Brenner, S., F. Jacob, and M. Meselson. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature.* 190:576–581.
2. Joyce, G. F. 1989. RNA evolution and the origins of life. *Nature.* 338:217–224.
3. Cech, T. R., S. H. Damberger, and R. R. Gutell. 1994. Representation of the secondary and tertiary structure of group I introns. *Nat. Struct. Biol.* 1:273–280.
4. Cech, T. R., and B. L. Bass. 1986. Biological catalysis by RNA. *Annu. Rev. Biochem.* 55:599–629.
5. Watson, J. D., and F. H. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* 171:737–738.
6. Khvorova, A., A. Reynolds, and S. D. Jayasena. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell.* 115:209–216.
7. Snoussi, K., and J. L. Leroy. 2001. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry.* 40:8898–8904.
8. Saenger, W. 1984. Principles of Nucleic Acid Structure. Springer-Verlag, New York.
9. Noy, A., A. Pérez, …, M. Orozco. 2004. Relative flexibility of DNA and RNA: a molecular dynamics study. *J. Mol. Biol.* 343:627–638.
10. Hashem, Y., and P. Auffinger. 2009. A short guide for molecular dynamics simulations of RNA systems. *Methods.* 47:187–197.
11. Lavery, R., K. Zakrzewska, …, J. Sponer. 2010. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* 38:299–313.
12. Lavery, R., H. Sklenar, …, B. Pullman. 1986. The flexibility of the nucleic acids. II. The calculation of internal energy and applications to mononucleotide repeat DNA. *J. Biomol. Struct. Dyn.* 3:989–1014.
13. Lankaš, F., R. Lavery, and J. H. Maddocks. 2006. Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure.* 14:1527–1534.
14. Orozco, M., A. Pérez, …, F. J. Luque. 2003. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* 32:350–364.
15. Olson, W. K., A. A. Gorin, …, V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95:11163–11168.
16. Dixit, S. B., D. L. Beveridge, …, P. Varnai. 2005. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II. Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* 89:3721–3740.
17. Young, M. A., G. Ravishanker, …, H. M. Berman. 1995. Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. *Biophys. J.* 68:2454–2468.
18. Lankas, F., J. Sponer, …, T. E. Cheatham, 3rd. 2003. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* 85:2872–2883.
19. Matsumoto, A., and W. K. Olson. 2002. Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophys. J.* 83:22–41.
20. Lankas, F., J. Sponer, …, J. Langowski. 2000. Sequence-dependent elastic properties of DNA. *J. Mol. Biol.* 299:695–709.
21. Cheatham, T., and P. Kollman. 1997. Molecular dynamics simulations: highlight the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J. Am. Chem. Soc.* 119:4805–4825.
22. Priyakumar, U. D., and A. D. Mackerell, Jr. 2008. Atomic detail investigation of the structure and dynamics of DNA·RNA hybrids: a molecular dynamics study. *J. Phys. Chem. B.* 112:1515–1524.
23. Pérez, A., A. Noy, …, M. Orozco. 2004. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.* 32:6144–6151.
24. Zacharias, M. 2000. Comparison of molecular dynamics and harmonic mode calculations on RNA. *Biopolymers.* 54:547–560.
25. Noy, A., F. J. Luque, and M. Orozco. 2008. Theoretical analysis of antisense duplexes: determinants of the RNase H susceptibility. *J. Am. Chem. Soc.* 130:3486–3496.
26. Noy, A., A. Pérez, …, M. Orozco. 2005. Structure, recognition properties, and flexibility of the DNA.RNA hybrid. *J. Am. Chem. Soc.* 127:4910–4920.
27. Brooks, B. R., C. L. Brooks, 3rd, …, M. Karplus. 2009. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30:1545–1614.
28. MacKerell, Jr., A. D., N. Banavali, and N. Foloppe. 2000-2001. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers.* 56:257–265.
29. 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.
30. Spellmeyer, D., T. Fox, …, P. Kollman. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.
31. Cheatham, 3rd, T. E., P. Cieplak, and P. A. Kollman. 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* 16:845–862.
32. Rueda, M., C. Ferrer-Costa, …, M. Orozco. 2007. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA.* 104:796–801.
33. Meyer, T., X. de la Cruz, and M. Orozco. 2009. An atomistic view to the gas phase proteome. *Structure.* 17:88–95.
34. Pérez, A., F. Lankas, …, M. Orozco. 2008. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* 36:2379–2394.
35. Van Wynsberghe, A. W., and Q. Cui. 2005. Comparison of mode analyses at different resolutions applied to nucleic acid systems. *Biophys. J.* 89:2939–2949.
36. McDowell, S. E., N. Spacková, …, N. G. Walter. 2007. Molecular dynamics simulations of RNA: an in silico single molecule approach. *Biopolymers.* 85:169–184.
37. Perez, A., J. Blas, …, J. Lopez-Bes. 2005. Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.* 1:790–800.
38. Soliva, R., C. A. Laughton, …, M. Orozco. 1998. Molecular dynamics simulations in aqueous solution of triple helices containing d (G·C·C) trios. *J. Am. Chem. Soc.* 120:11226–11233.
39. Pérez, A., F. J. Luque, and M. Orozco. 2007. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* 129:14739–14745.
40. Shields, G., C. A. Laughton, and M. Orozco. 1997. Molecular dynamics simulations of the d(TAT) triple helix. *J. Am. Chem. Soc.* 119:7463–7469.
41. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an $N\cdot\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* 98:10090–10092.
42. Ryckaert, J., G. Ciccotti, and H. Berendsen. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* 23:327–341.
43. Phillips, J. C., R. Braun, …, K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.
44. Case, D. A., T. Darden, …, P. A. Kollman. 2008. AMBER 10. University of California, San Francisco, CA.
45. Olson, W. K., M. Bansal, …, H. M. Berman. 2001. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* 313:229–237.
46. Lavery, R., M. Moakher, …, K. Zakrzewska. 2009. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* 37:5917–5929.
47. Nelson, M., W. Humphrey, …, A. Dalke. 1996. NAMD: a parallel, object-oriented molecular dynamics program. *Int. J. High Perform. Comput. Appl.* 10:251–268.
48. Gelpí, J. L., S. G. Kalko, …, M. Orozco. 2001. Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins.* 45:428–437.

49. Schlitter, J., and M. Klähn. 2003. A new concise expression for the free energy of a reaction coordinate. *J. Chem. Phys.* 118:2057–2060.

50. Andricioaei, I., and M. Karplus. 2001. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* 115:6289–6292.

51. Berman, H. M., C. Zardecki, and J. Westbrook. 1998. The nucleic acid database: a resource for nucleic acid science. *Acta Crystallogr. D Biol. Crystallogr.* 54:1095–1104.

52. Leroy, J. L., E. Charretier, …, M. Guéron. 1988. Evidence from base-pair kinetics for two types of adenine tract structures in solution: their relation to DNA curvature. *Biochemistry.* 27:8894–8898.

53. Abels, J. A., F. Moreno-Herrero, …, N. H. Dekker. 2005. Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys. J.* 88:2737–2744.

54. Kebbekus, P., D. E. Draper, and P. Hagerman. 1995. Persistence length of RNA. *Biochemistry.* 34:4354–4357.

55. Hagerman, P. J. 1988. Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.* 17:265–286.

56. Curuksu, J., M. Zacharias, …, K. Zakrzewska. 2009. Local and global effects of strong DNA bending induced during molecular dynamics simulations. *Nucleic Acids Res.* 37:3766–3773.

57. Goñi, J. R., C. Fenollosa, …, M. Orozco. 2008. DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics.* 24:1731–1732.

58. Goñi, J. R., A. Pérez, …, M. Orozco. 2007. Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.* 8:R263.

59. Pérez, A., J. Sponer, …, M. Orozco. 2005. Are the hydrogen bonds of RNA (AU) stronger than those of DNA (AT)? A quantum mechanics study. *Chemistry.* 11:5062–5066.

4.2.1.1 Results and discussion

Despite their chemical similarity, DNA and RNA duplex structures play very different biological roles. While double stranded DNA can display a large variety of structures close to the B-form with very dynamic sugar-phosphate conformations (Neidle 2007), RNA double stranded has been traditionally considered rigid. However, (Priyakumar & MacKerell 2008) presented results from CHARMM27 MD simulations that pointed in the opposing direction, stating that RNA_2 is more flexible than DNA_2. To shed some light on this topic, we study the flexibility of double stranded RNA molecules by means of molecular dynamics simulations on four different 18mer sequences. AMBER and CHARMM nucleic acid force fields were used to analyze the sequence dependent elastic properties and the results were ultimately compared with the obtained for DNA duplex (Pérez et al. 2008), were both force fields produced comparable results.

Reliability of the trajectories A common starting analysis of any MD trajectory is the RMSD calculation with respect to a reference structure. Usually, for relatively short nucleic acid molecules, the average RMSD value, taking the average structure as a reference, is around 3-4 Å. Parmbsc0 (Pérez et al. 2007) revision of the AMBER ff99 force field (Cornell et al. 1995) yielded stable trajectories for all the simulated RNA molecules. By contrast, CHARMM27 (Brooks et al. 2009; MacKerell et al. 2000) force field displayed an erratic behavior along the trajectories with higher RMSD values regarding the parmbsc0 simulations. Visual inspection together with the energetic analysis of the dsRNA molecules suggested that this unpredictable phenomenon is due to in some cases irreversible loss of hydrogen bonds between complementary bases at the central part of the RNA molecules. Based on these results, we decided to fully trust only on parmbsc0 simulations in contrast to what it was found for double stranded DNA molecules (Pérez et al. 2008).

Sequence dependent flexibility The analysis of the base pair step helical parameters provides a useful tool to evaluate the local flexibility in terms of dinucleotide steps. The MD-derived stiffness parameters were successfully compared with X-ray experimental data (only steps with at least 40 structures were used for comparison) showing clear differences among the ten dinucleotide steps. While GC and CC steps exhibit a rather rigid behavior for translational and rotational deformations, AU and UA steps are very flexible. But caution has to be taken since one step can be stiff for rise deformation but at the same time it can easily tolerate tilt rotations. The same analysis for DNA molecules show a similar order for the different steps but RNA stiffness values are higher than corresponding DNA values. Interestingly, the r(AU) step is unexpectedly

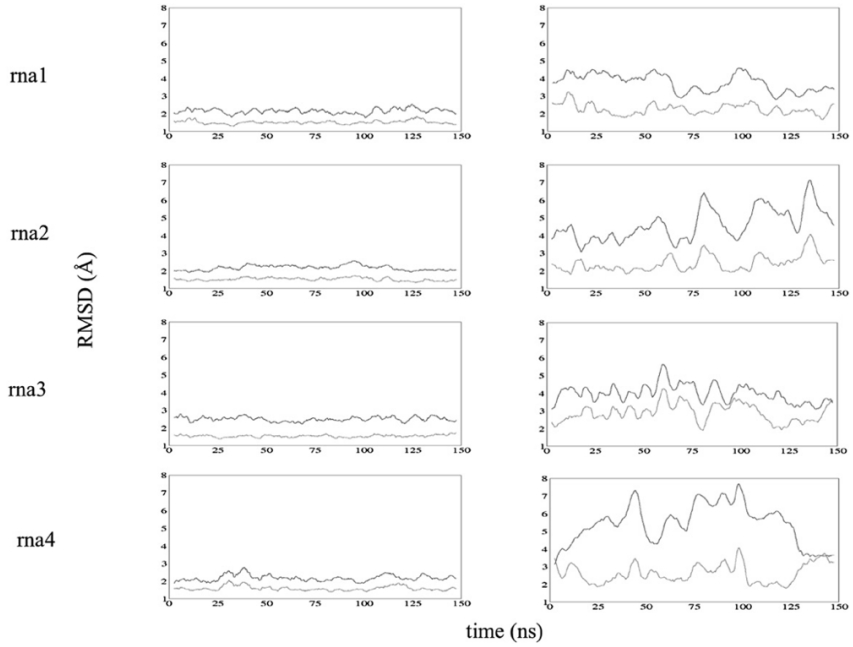


Figure 4.6: Smoothed RMSD (in Å) from A-RNA fiber conformation (in gray) and average structure (in black) for RNA/Parmbsc0 (on the left) and CHARMM27 simulations (right) for the central 14-mer of the four RNA sequences.

more flexible than the corresponding d(AT) step which is probably related to the lack of the 5-methyl group.

Global flexibility Parmbsc0 estimates of intramolecular entropies for DNA and RNA molecules show that dsDNA is more flexible than dsRNA. This higher flexibility is mainly due to the higher order deformation modes since lower order deformation modes are similar for both DNA and RNA molecules.

Global helical stiffness parameters allow also an insightful analysis of the global helical flexibility providing a direct comparison with experimental flexibility data (Table 5.7). The results from parmbsc0 show that global deformations of roll, tilt and twist in dsRNA are more difficult than for dsDNA, a conclusion that agrees with persistence length experimental measurements. However, dsRNA becomes more flexible for global stretching deformations than dsDNA which stresses on how cautious we have to be when talking about flexibility.

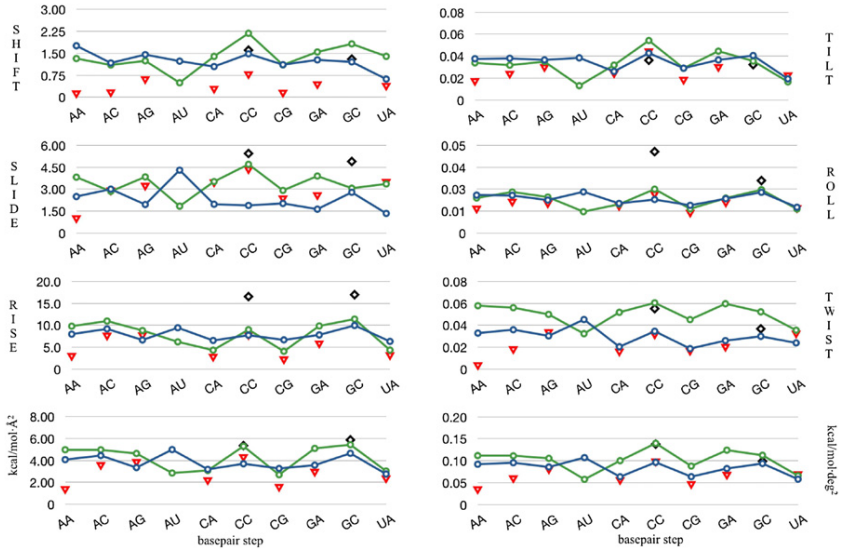


Figure 4.7: Stiffness constants (translational ones in kcal/mol·Å^2 and rotational ones in kcal/mol·deg^2) for the 10 representative dinucleotide steps associated to the different deformation modes comparing DNA/Parmbsc0 (in blue), RNA/Parmbsc0 (in green with lines), RNA/CHARMM27 (in red triangles), and derived for analysis of X-ray structural data (in black diamonds) values. (Bottom) Summation of stiffness constants for translational helical parameters (left), and the same for rotational helical parameters (right).

	Bending (nm)			Stretching	Twisting
	Roll	Tilt	Isotropic	(pN)	(nm)
Seq 1	64.99	55.67	59.97	2318.36	86.38
	77.45	*71.74*	*74.49*	*1662.39*	*177.19*
Seq 2	71.83	46.38	56.37	2540.19	93.05
	102.13	*62.32*	*77.40*	*1745.22*	*191.66*
Seq 3	74.35	37.11	49.51	2262.47	90.93
	105.67	*67.57*	*82.43*	*1860.05*	*198.35*
Seq 4	66.49	57.83	61.86	2475.36	81.34
	99.67	*71.42*	*83.21*	*1448.42*	*195.18*
Averages	69.42 ± 4.41	49.25 ± 9.49	56.93 ± 5.45	2399.09 ± 30.27	87.93 ± 5.20
	96.23 ± 12.76	*68.26 ± 4.39*	*79.38 ± 4.16*	*1679.02 ± 173.79*	*190.60 ± 9.35*

Table 4.7: DNA and RNA (italics) bending (anisotropic and isotropic) (B) and twisting (C) persistence lengths (in nm) and stretch modulus (S, in pN) for the four sequences. Standard deviations from parmbsc0 simulations are shown. Values come from analysis taking 11-mer sequences.

4.2.1.2 Conclusions

Definition of flexibility in nucleic acids is complex and local or global context should be previously specified. Analysis of RNA duplex deformations at the base pair step level have shown that dinucleotides steps do not equally behave exhibiting very different stiffness. We can therefore classify stiff and flexible steps depending on their capacity to tolerate translational and rotational deformations. Overall, the RNA dinucleotide steps are more rigid than the DNA steps at least for double stranded structures. Global flexibility analysis confirms that DNA duplexes are, in general, more flexible than RNA duplexes.

Unfortunately, while for DNA_2 both AMBER and CHARMM force fields provided similar results, in the case of RNA_2 systems, only parmbsc0 AMBER force field allowed us to derive reliable information. This artifactual behavior seems to be related to the loss of hydrogen bonds between complementary bases during the simulations related perhaps to a problem in the parameterization of the 2'-hydroxyl group (Denning et al. 2011). It is worth to note that this work with CHARMM27 and RNA systems contribute to a subsequent modification and partial improvement in the following CHARMM36 force field.

4.2.1.3 References

Brooks, B.R. et al., 2009. CHARMM: The biomolecular simulation program C. L. Brooks & D. A. Case, eds. Journal of Computational Chemistry, 30(10), pp.1545–1614.

Cornell, W.D. et al., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society, 117(19), pp.5179–5197.

Denning, E.J. et al., 2011. Impact of 2'-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA. Journal of Computational Chemistry, 32(9), pp.1929–1943.

MacKerell, A.D., Banavali, N.K. & Foloppe, N., 2000. Development and current status of the CHARMM force field for nucleic acids. Biopolymers, 56(4), pp.257–265.

Neidle, S., 2007. Principles of nucleic acid structure Academic Press, Academic Press.

Pérez, A. et al., 2007. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. Biophysical Journal, 92(11), pp.3817–3829.

Pérez, A. et al., 2008. Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Research, 36(7), pp.2379–2394.

Priyakumar, U.D. & MacKerell, A.D., 2008. Atomic detail investigation of the structure and dynamics of DNA.RNA hybrids: a molecular dynamics study. The Journal of Physical Chemistry B, 112(5), pp.1515–1524.

5

PhD Advisor Report

5.1 *Unique Tautomeric and Recognition Properties of Thioketothy-mines?*

Ignacio Faustino, Anna Aviñó, Iván Marchán, F. Javier Luque, Ramón Eritja y Modesto Orozco. *J Am Chem Soc.* 2009, 131(35), pp.12845–12853.

This paper explored the nature of thio-derivatives of thymine. We started the study under the assumption that a previous experimental study (also in JACS) by a world-leading research group described the existence of enol-derivatives of thioketothymines (S), since it was the only way to explain their surprising results of the equivalence of G·S and A·S pairings. A wide and systematic theoretical study convinced us that the equivalence of G·S and A·S pairings suggested in the literature cannot be explained based on "state of the art" calculations. Furthermore, we demonstrated that the tautomeric pattern of T and S is the same, also in contrast with previous experimental results. We performed then experimental measures, demonstrating the existence of an artifact in previous published results and showing how S behaves exactly as T, opening the possibility to incorporate these thio-derivatives into DNA as a thymidine derivative with interesting biotechnological applications. The paper was published in J Am Chem Soc, one of the most prestigious chemistry journals (IF above 10). The results led to the group who published the first results to publish a corrigendum in the same journal, which appeared before our paper was published. Ignacio did all the calculations shown in the paper, while the experimental validation of our calculations was done in R.Eritja's laboratory.

5.2 *Toward a consensus view of duplex RNA flexibility.*

Ignacio Faustino, Alberto Pérez y Modesto Orozco. *Biophysical Journal.* 2010, 99(6), pp.1876–1885.

This paper presents the first consensus study of the mechanical properties of RNA duplexes. By using molecular dynamics simulations, we characterized the structure and flexibility of a variety of canonical RNA duplexes using both parmbsc0 and charmm27 force fields. The study found important errors in charmm27 which, after our work, were corrected by the developer group, proving on the contrary the reliability of parmbsc0. We provided in this paper the first set of elastic parameters for RNA. Biophysical Journal is the reference journal for Biophysics in the world. The paper has collected since 2010 a total of 14 citations. The presented results are considered the standard in the field. Ignacio was the main responsible of all the calculations.

5.3 *Functionalization of the 3'-Ends of DNA and RNA Strands with N-ethyl-N-coupled Nucleosides: A Promising Approach To Avoid 3'-exonuclease-Catalyzed Hydrolysis of Therapeutic Oligonucleotides.*

Montserrat Terrazas, Adele Alagia, **Ignacio Faustino**, Modesto Orozco y Ramón Eritja. *ChemBioChem.* 2013, 14(4), pp.510–520.

This paper is the result of a collaboration between Eritja's laboratory and us. We present here the rational design, synthesis and experimental validation of the properties of N-ethyl-N-coupled nucleosides as elements to protect siRNAs from exonuclease degradation, while keeping external RNAs as substrates of the RISC degradation system. The designed modification shows an incredible power to protect oligos from nuclease degradation and an excellent ability to be processed by RISC giving to extremely good activity properties. The paper has been recently published in a prestigious chemical journal. Ignacio was the responsible of all the calculations, while all the experimental part was done by M. Terrazas in Eritja's laboratory.

5.4 *The DNA-forming properties of 6-selenoguanine.*

Ignacio Faustino, Carles Curutchet, F. Javier Luque y Modesto Orozco. Submitted.

This paper presents the first complete study of the impact of guanine to selenoguanine modification in different forms of DNA. Using state of the art simulation techniques, from QM to MD, we determine the different impact of introducing selenoguanine in DNA, finding that, despite the extreme difference between O and Se, the modified nucleobase can be easily incorporated into DNA, creating oligos with improved conductimetric possibilities. The results have been submitted to a top-class journal in the field. Ignacio was the main responsible of all the calculations shown in the paper.

6

Summary (Spanish)

El DNA genómico define gran parte de la expresión génica y el fenotipo de un individuo. Sin embargo, resulta cada vez más claro la influencia que tiene una capa reguladora que se encuentra por encima de la propia secuencia del genoma y que hoy llamamos control epigenético (Feinberg 2007). Dentro de estos mecanismos de regulación se encuentran las modificaciones post-transcripcionales del DNA y de las histonas, así como las cortas y largas secuencias de RNA no codificantes. Todos estos mecanismos de manera conjunta regulan la expresión génica mediante cambios en la organización de la cromatina y de la accesibilidad del DNA genómico (Mattick 2007).

Posiblemente, una de las las mayores consecuencias del Proyecto Genoma Humano fue la gran cantidad de estudios de asociación a gran escala de mutaciones de variaciones a nivel genómico con las enfermedades. El DNA eucariótico está presente dentro de la célula tanto en su forma cromosómica como en su forma extracromosómica. Mientras que el DNA extracromosómico, principalmente de origen mitocondrial, se encuentra de manera relativamente desorganizada y circular, el DNA cromosómico se encuentra empaquetado en diferentes niveles jerárquicos, estructuras repetitivas llamadas cromatina. En su forma elemental, la cromatina comprende una vuelta de 147 pares de bases de ácido nucleico que se enrollan un vuelta y media alrededor de un nucleosoma. Sin embargo, por cuestiones de eficiencia reguladora, el procesamiento de los alrededor de 23.000 genes se realiza de manera compartimentalizada dentro del genoma. Por una parte, se encuentran hay áreas más activas del genoma que reciben el nombre de eucromatina que están empaquetadas de manera un poco más dispersa y que presentan una mayor accesibilidad por parte de los factores de regulación. Por otro lado, existen otras regiones más inactivas desde el punto de vista genómico que se encuentran fuertemente empaquetadas que reciben el nombre de heterocromatina. Dentro de las regiones de heterocromatina, éstas se pueden clasificar en dos tipos: heterocromatina constitutiva,

que comprende regiones que están silenciadas en todos los tipos de células, y la heterocromatina facultativa, que son regiones que no necesitan estar activas en todas las células de manera que están en unas pero no en otras. Básicamente, el grado de compactación del cromatina está basado en modificaciones del DNA, en modificaciones de las histonas y por la regulación del RNA no codificantes.

Dentro de las modificaciones del DNA genómico, la más común de todas es sin lugar a dudas la metilación de la citosina. Aproximadamente existe un 1% de todas las citosinas que se encuentran metiladas, de las cuales un 70% se encuentran en los pasos CpG (no confundir con el apareamiento CG) (Lister et al. 2009). Actualmente, se conocen otros tipos de modificaciones de estos pasos incluyendo hidroximetilaciones, formilaciones y carboxilaciones (Ito et al. 2010). Hacia 1975, la metilación del DNA se creía ligada únicamente a la inactivación de cromosoma X pero posteriormente se identificó como una marca específica en el genoma en diferentes tejidos celulares. De esta manera se demostró que los genes específicos de tejido se encuentran inframetilados y los genes expresados constitutivamente están controlados por promotores con islas CpG no metiladas (Razin & Szyf 1984). El análisis de hipersensibilidades por DNasaI (Crawford et al. 2006) y MNasa (Weiner et al. 2010) permiten la identificación de regiones de cromatina que se encuentran más accesibles.

Un elemento clave en todos estos procesos reguladores es la flexibilidad inherente del DNA que permite la interacciones de largo alcance entre elementos distantes a través la formación de bucles de DNA (Zhao et al. 2006). Esta versatilidad estructural viene definida principalmente por su naturaleza química y por tanto, por su secuencia. Existen otros factores medioambientales que inciden en las características de flexibilidad de las cadenas de ácidos nucleicos. Por ejemplo, la proximidad de iones divalentes así como las interacciones con las moléculas de agua alteran esta variabilidad estructural en los ácidos nucleicos. De hecho, los diferentes patrones de deformación del DNA son utilizados como mecanismo de reconocimiento por las proteínas, aunque actualmente los mecanismos con que las proteínas escanean el genoma para encontrar la secuencia a la que unirse a una velocidad mucho mayor de la que se puede esperar por difusión en tres dimensiones se desconocen (Araúzo-Bravo et al. 2005; Oguey et al. 2010).

Normalmente, el DNA aparece en la forma B de doble cadena de manera que dos hebras complementarias entre sí interaccionan mediante pares de bases Watson-Crick. Aunque la doble hélice es termodinámicamente estable en condiciones fisiológicas, se encuentra en un equilibrio conformacional que genera cambios en los patrones de los apareamientos así como cambios dependientes de la secuencia. Esta variabilidad conformacional se basa principalmente en las interacciones de van der Waals y en interacciones electrostáticas entre los pares de bases adyacentes.

6.1 Estudio teórico de nucleobases modificadas

Las modificaciones químicas de ácidos nucleicos tienen una amplia variedad de aplicaciones tanto en clínica, utilizadas como agentes antisentido (Kurreck 2003), así como en el estudio de las estructuras de los propios ácidos nucleicos, las interacciones proteína DNA o en la catalisis de ácidos nucleicos por poner algunos ejemplos (Blackburn et al. 2006).

Actualmente se sintetizan cientos de análogos de nucleósidos estándar en laboratorios farmacéuticos, algunos de ellos como, por ejemplo, los análogos de la arabinosa de adenosina y citosina, son útiles como fármacos antivirales o anticancerígenos (Abel et al. 1975). La mayoría de los fármacos anticancerígenos actúan inhibiendo la síntesis de DNA de alguna manera si bien es cierto que tienen bastantes efectos secundarios. Sin embargo, la combinación de varios de estos fármacos reduce la resistencia y minimizan la toxicidad. Ejemplos de este tipo de análogos son los antimetabolitos que al ser estructuralmente muy similares a los nucleósidos estándar son capaces de interferir en la producción de los ácidos nucleicos inhibiendo las rutas de síntesis de ácidos nucleicos o incorporándose dentro de la estructura del DNA y el RNA (Kinsella 1997). Una vez que se incorporan al DNA, estos compuestos provocan la terminación de la cadena durante la replicación del DNA y en consecuencia la muerte celular.

Existen modificaciones del esqueleto azúcar-fosfato de los oligonucleótidos que mejoran la eficacia en la síntesis de siRNAs controlando la expresión génica en unos casos al bloquear estéricamente la actividad enzimática de las nucleasas, en otros inhibiendo la hibridación del mRNA con la hebra antisentido o incluso favoreciendo la formación de triplexes con secuencias específicas de DNA. Un ejemplo de ello son los fosforotioatos que poseen un átomo de oxígeno no enlazante sustituido por azufre en el enlace fosfodiéster (Eckstein 1985). Éste tipo de modificaciones han demostrado ser resistentes tanto para exo- como endonucleasas aumentando la resistencia en cultivos celulares y en suero.

6.2 Estudio comparativo de la flexibilidad DNA versus RNA de doble cadena

En las dos últimas décadas, los estudios físicos sobre el DNA combinados con las técnicas de modelado molecular y las simulaciones, han permitido confirmar que la doble hélice tiene una flexibilidad notable en respuesta a la aplicación de fuerzas externas. Esta marcada flexibilidad es utilizada por las proteínas para interaccionar con secuencias

específicas. En los datos cristalográficos existen varios ejemplos de cómo las proteínas pueden inducir cambios locales en la estructura del DNA.

Como se ha dicho anteriormente, los puentes de hidrógeno y las interacciones por apilamiento generan de alguna manera una restricción en la flexibilidad geométrica del DNA. Las pequeñas modificaciones en la estructura helicoidal en respuesta a cambios en la secuencia influyen en la anchura de los surcos, en la torsión helicoidal, la curvatura, la rigidez mecánica y la resistencia a ser dobladas. Todas estas características intrínsecas de la secuencia son reconocidas por las proteínas para interaccionar con las secuencias específicas. Es por ello que se puede clasificar las secuencias del genoma en regiones que son más flexibles o más rígidas. Un ejemplo típico son los segmentos de adeninas consecutivas (tres o más) que generan cierta curvatura pero a la vez una rigidez intrínseca en la estructura. Se ha sugerido que el origen de dicha rigidez es debida a una red de interacciones de puentes de hidrógeno en el surco mayor que proporcionan unas regiones especialmente curvadas de los cromosomas.

6.3 Objetivos de la presente tesis

En esta tesis se han utilizado las técnicas de mecánica cuántica y mecánica clásica para estudiar las propiedades tautoméricas de formas modificadas de las bases estándar con posibles aplicaciones en el campo de la terapia clínica y de la biotecnología. Por otro lado también se ha realizado un estudio sobre la flexibilidad de los ácidos nucleicos acentuando las diferencias entre RNA y DNA de doble hebra y sus consecuencias biológicas.

6.3.1 Estudio teórico de timinas modificadas con azufre.

1. Estudio de las preferencias tautoméricas de la 2- y 4-tiocetotiminas tanto en fase gas como en atmósfera de hidratación.

2. Determinar la influencia estructural de las formas más estables de la 2- y 4-tiocetotiminas en un DNA de doble hebra teniendo como base complementaria tanto adenina como guanina.

3. Determinar las preferencias energéticas de las mutaciones de T $\rightarrow$ S (S = 2- ó 4-tiocetotimina) cambiando en función de la base complementaria.

4. Determinar el papel de las formas tautoméricas minoritarias en el contexto de un apareamiento no estándar con guanina.

6.3.2 Estudio teórico de guanina modificadas con selenio.

1. Estudio de las preferencias tautoméricas de la 6-selenoguanina (6SeG) tanto en fase gas como en medio acuoso.

2. Determinar la influencia de la 6-selenoguanina en el DNA de doble, triple y estructuras G-cuádruplex.

3. Determinar la estabilidad relativa de la mutación G $\rightarrow$ 6SeG en los diferentes sistemas de DNA y comparar los resultados teóricos con los datos experimentales disponibles hasta la fecha.

4. Explorar las propiedades electrónicas de la 6-selenoguanina en su uso potencial en aplicaciones biotecnológicas.

6.3.3 Inhibición de 3'-exonucleasas con nucleósidos diméricos modificados con N-etil-N.

1. Inhibición de la familia de 3'-exonucleasas y mejora de la efectividad de siRNAs en estrategias antisentido mediante diseño racional de nucleósidos modificados.
2. Validación teórica y experimental del los nucleósidos modificados propuestos con 3'-exonucleasas y ensayos con células.

6.3.4 Estudio de la flexibilidad del RNA de doble hebra.

1. Determinar el grado de convergencia de los resultados derivados de los campos de fuerzas más utilizados para ácidos nucleicos.
2. Caracterizar las propiedades dependientes de secuencia para las moléculas de RNA de doble hebra.
3. Evaluar las propiedades mecánicas de moléculas de RNA de doble hebra y la comparación con los análogos de DNA.

6.4 Resumen de las publicaciones

6.4.1 *Unique Tautomeric and Recognition Properties of Thioketothymines?*

Ignacio Faustino, Anna Aviñó, Iván Marchán, F. Javier Luque, Ramón Eritja y Modesto Orozco.
J Am Chem Soc. 2009, 131(35), pp.12845–12853.

Las bases sintéticas, una vez introducidas en el DNA, son capaces de inducir cambios conformaciones y alterar la estabilidad induciendo incluso mutaciones o modificando los patrones de reconocimiento. La forma tautomérica que suele suponerse para análogos de nucleobases estándar es la forma ceto/amino. Sin embargo, la influencia del entorno puede en algunos casos favorecer la existencia de formas teóricamente minoritarias. En este estudio se describen las propiedades tautoméricas de la timina, y la 2- y 4-tiocetotiminas mediante métodos *ab initio* y la combinación de simulaciones de dinámica molecular y cálculos de energía libre. Los patrones de reconocimiento así como las estabilidades relativas de los tautómeros mayoritarios en diversos entornos fueron identificados. Los resultados teóricos fueron apoyados por los resultados experimentales aportados por el laboratorio del Prof. Dr. Ramón Eritja, y permiten describir las características tautoméricas y de reconocimiento de las tiocetotiminas.

6.4.2 *The DNA-forming properties of 6-selenoguanine.*

Ignacio Faustino, Carles Curutchet, F. Javier Luque y Modesto Orozco.
Enviado.

Los análogos de purinas 6-mercaptopurina y 6-tioguanina se utilizan como agentes inhibidores del metabolismo de los ácidos nucleicos en diversos tipos de leucemias. En este trabajo presentamos una caracterización exhaustiva de la estructura y propiedades de interacción de la 6-selenoguanina (6SeG), un isóstero de la guanina que introducido en el DNA puede afectar diferencialmente en estructuras de mayor orden del DNA. Mediante cálculos exactos de mecánica cuántica y la combinación de simulaciones de dinámica molecular con métodos de energía libre, hemos evaluado la influencia de la sustitución de la guanina estándar por la 6-selenoguanina en sistemas de doble, triple hebra de DNA así como en la estabilidad de estructuras G-cuádruplex. Finalmente, se aborda el estudio de las propiedades electrónicas de la 6-selenoguanina ya que se

ha observado que ciertas modificaciones introducidas en el DNA podrían aumentar la capacidad conductimétrica de estos sistemas.

6.4.3 *Functionalization of the 3'-Ends of DNA and RNA Strands with N-ethyl-N-coupled Nucleosides: A Promising Approach To Avoid 3'-exonuclease-Catalyzed Hydrolysis of Therapeutic Oligonucleotides.*

Montserrat Terrazas, Adele Alagia, **Ignacio Faustino**, Modesto Orozco y Ramón Eritja.
ChemBioChem. 2013, 14(4), pp.510–520.

Existen ciertas limitaciones en el desarrollo de ácidos nucleicos modificados con aplicaciones clínicas y es que, la vulnerabilidad a la acción de nucleasas, tanto exo- como endonucleasas, limita la vida media de estos oligos en suero. En la actualidad, los fosforotioatos son las modificaciones que generan mayor resistencia aunque presentan la limitación de que un elevado número de los fosfatos del oligo han de ser modificados. En este trabajo presentamos una nueva modificación en el extremo 3', en la que dos nucleósidos se unen por el lado de la base en lugar de por enlace fosfodiéster que es eliminado en el diseño. Simulaciones de dinámica molecular del complejo molecular con el fragmento Klenow de la polimerasa I de DNA de *E. coli* y el oligo modificado permiten demostrar la base del diseño racional del dímero modificado. Los estudios estructurales se complementaron con los experimentos realizados por la Dr. Montserrat Terrazas y Adele Alagia del grupo del Prof. Dr. Ramón Eritja. Éstos incluían estudios de digestión por 3'-exonucleasas (fragmento Klenow de la polimerasa I de DNA de *E. coli* y la fosfodiesterasa de veneno de serpiente) así como experimentos celulares de RNA de interferencia.

6.4.4 *Toward a consensus view of duplex RNA flexibility.*

Ignacio Faustino, Alberto Pérez y Modesto Orozco.
Biophysical Journal. 2010, 99(6), pp.1876–1885.

El estudio de la flexibilidad de los ácidos nucleicos ha sido y sigue siendo un campo actual en efervescencia gracias a la aportación de las técnicas de moléculas individuales y a los métodos computacionales como la dinámica molecular. En este estudio se aborda la influencia de la secuencia en la estructura, estabilidad y flexibilidad de moléculas de

RNA de doble cadena mediante simulaciones de dinámica molecular. Otro de los objetivos del estudio era comparar los dos campos de fuerza que son mayoritariamente utilizados en el campo de los ácidos nucleicos, AMBER (parm99bsc0) y CHARMM (CHARMM27). Los resultados obtenidos demuestran que la flexibilidad es dependiente de la secuencia pudiendo agrupar los diferentes dinucleótidos en su mayor o menor capacidad para ser deformados. Medidas experimentales como la longitud de la persistencia o la torsión de las dobles hélices de RNA son parámetros moleculares globales que definen la flexibilidad en una escala mayor. Mediante métodos de dinámica esencial, se calcularon los promedios de las entropías intramoleculares de las cuatro secuencias consideradas así como las estimaciones teóricas de la longitud de la persistencia, de la torsión helicoidal y del estiramiento molecular. Además, los resultados del estudio aportan más datos acerca de la diferente flexibilidad que existe entre el DNA y el RNA de doble cadena.

6.5 Conclusiones

6.5.1 Estudio teórico de nucleobases modificadas

El análisis de las preferencias tautoméricas de los derivados de timina, *2- y 4-tiocetotiminas*, mostraron que en fase gas y en disolvente acuoso las formas correspondientes más estables corresponden a la forma tioceto de manera que las el resto de formas posibles, en especial las formas tiol, tendrían mucha menor importancia. Estos estudios se trasladaron al ambiente del DNA donde las interacciones con diferentes bases complementarias demostraron que las tiocetotiminas inducen pequeñas distorsiones siempre de manera localizada. Sin embargo, estos cambios son, según los cálculos de energía libre, los causantes de la desestabilización de la mutación timina → tiocetotimina. Dicha desestabilización aparece sin importar la identidad de la base complementaria y, aunque la guanina pudiera estabilizar alguna de las formas tiol, este efecto no sería suficiente para invertir las preferencias tautoméricas anteriormente comentadas en contraposición a resultados previos (Sintim & Kool 2006). Estos estudios confirman que la sustitución de timina por tiocetotiminas en sistemas de doble hebra de DNA conservan la estructura global, relativa estabilidad y especialmente la fidelidad de las interacciones abriendo las posibilidades farmacéuticas y de reactividad de estos compuestos y del azufre en DNA.

En el estudio de las propiedades de la *6-selenoguanina*, los estudios con mecánica cuántica mostraron que, aunque en fase gas la preferencia tautomérica se decanta por las formas selenol, la influencia del disolvente acuoso favorece la forma tautomérica selona resultando decisiva en el equilibro tautomérico. La introducción del selenoderivado en la estructura en el DNA se probó en sistemas de doble, triple hélice y en G-cuádruplexes. El análisis de las simulaciones de dinámica molecular demostraron que la peor interacción de puentes de hidrógeno en los que se encuentra involucrado el átomo de selenio producen una importante desestabilización de la estructura confirmada por los cálculos de energía libre y los resultados experimentales publicados anteriormente Salon:2008kl. Uno de los resultados más interesantes de estos cálculos fue la menor desestabilidad de la mutación cuando ésta se introduce en la hebra formadora del triplex (TFO en inglés) en comparación con la producida cuando una tétrada de guaninas de un G-cuádruplex. Dicha menor desestabilización sugiere por tanto que, un oligo rico en guaninas en la que se introdujera esta mutación, formaría preferentemente un triplex antiparalelo con un dúplex de DNA complementario en lugar de formar un estructura de G-cuádruplex. Este efecto sería de suma utilidad puesto que uno de los mayores problemas en el estudio de formación de triplex antiparalelos desde el punto de vista experimental es precisamente la tendencia que tienen las secuencias ricas en guaninas a formar estructuras de G-cuádruplex.

Finalmente, el estudio de las propiedades electrónicas en biotecnología se realizó mediante cálculos a nivel cuántico mostrando una reducción de la diferencia entre los orbitales atómicos HOMO (del inglés, mayor orbital molecular ocupado) y LUMO (del inglés, menor orbital molecular desocupado) en comparación con la guanina. Es precisamente entre estos orbitales donde se produce la excitación de los electrones y la posterior transferencia entre orbitales de moléculas suficientemente cercanas. Por tanto, en teoría, cuanto menor sea la diferencia entre ambos orbitales mejor será la capacidad para generar dicha transferencia. La posibilidad de utilizar biotecnológicamente la 6-selenoguanina como conductor eléctrico en sistemas de DNA se realiza por primera vez en este trabajo.

El estudio sobre la resistencia a la acción de 3'-exonucleasas presenta una modificación de *nucleósidos diméricos* con modificación N-etil-N que producen el bloqueo de dichas enzimas. La modificación se caracteriza además por la ausencia del grupo fosfato entre ambos nucleósidos que, junto al puente N-etil-N impiden la correcta posición del sustrato en el sitio activo de las 3'-exonucleasas consideradas y lo bloquean sin permitir la hidrólisis del enlace fosfodiéster más cercano. Los ensayos de digestión con 3'-exonucleasas confirmaron la resistencia de dicha modificación produce en los oligos al incorporar una única de estas modificaciones en el extremo 3'. Además, las pruebas realizadas con los oligos modificados fosforotioatos demostraron tener una resistencia alta pero menor que la provista por el dímero modificado en 3'. Por último, los estudios con RNAs de interferencia con las modificaciones en cultivos celulares demostraron su capacidad para ser reconocidos por los mecanismos celulares de degradación de mRNA con una eficacia similar a la obtenida con la forma no modificada. Con estos experimentos, la modificación presentada en este trabajo confiere una resistencia excelente contra la acción de 3'-exonucleasas. De hecho, al necesitarse únicamente una única modificación en el oligo, ésta puede combinarse con otras modificaciones para mejorar la resistencia contra otras nucleasas y para facilitar su introducción en la célula como agente antisentido.

6.5.2 Estudio comparativo de la flexibilidad DNA versus RNA de doble cadena

La flexibilidad de los ácidos nucleicos puede darse en un contexto local a nivel del par de bases o bien si se consideran las moléculas como gomas más o menos elásticas. El análisis de las estructuras de RNA de doble hélice a nivel del par de bases muestran la diversidad que existe y, por tanto, la dependencia de la secuencia. El cálculo de las constantes de fuerza permitieron agrupar los diferentes dinucleótidos en función de

su naturaleza flexible o rígida. Así, los resultados obtenidos revelan que la tendencia mostrada para RNA se cumple también para DNA con una excepción, el paso r(ApU) resulta ser más flexible que el d(ApT). Esta diferencia adquiere mayor interés si se piensa en la metilación análoga de los pasos d(CpG) que adquieren una mayor rigidez y están involucrados en la mecanismos epigenéticos de la expresión génica.

Por otra parte, se consideró los ácidos nucleicos con un tamaño mayor, es decir, como moléculas de cierta longitud con propiedades mecánicas dependientes de su secuencia. Tanto los valores de entropía intramolecular como otros parámetros globales calculados para las secuencias consideradas mostraron que las dobles hélices de DNA son más flexibles que los de RNA. Los resultados obtenidos están en excelente concordancia con los datos experimentales y prueban la capacidad de las técnicas de dinámica molecular en el estudio de propiedades mesoscópicas. La dependencia de secuencia para deformaciones a nivel de par de bases o dinucleótidos permitió su aplicación en el diseño y predicción de complejos de RNA de interferencia con el complejo RISC, esencial en la maquinaria de degradación de mRNA dentro de la célula.

Finalmente, mencionar que las deficiencias observadas en las trayectorias con el campo de fuerzas CHARMM27 en sistemas de RNA obligaron a centrarse en los resultados obtenidos sólo con el campo de fuerzas de AMBER que demostraron ser más acordes a los datos experimentales. Las distorsiones observadas pudieron ser debidas a problemas con la estabilidad de los puentes de hidrógeno y fueron parcialmente solventadas con una parametrización posterior a la publicación de este trabajo.

6.6 References

Abel, R., Kaufman, H.E. & Sugar, J., 1975. Intravenous adenine arabinoside against herpes simplex keratouveitis in humans. American journal of ophthalmology, 79(4), pp.659–664.

Araúzo-Bravo, M.J. et al., 2005. Sequence-Dependent Conformational Energy of DNA Derived from Molecular Dynamics Simulations: Toward Understanding the Indirect Readout Mechanism in Protein-DNA Recognition. Journal of the American Chemical Society, 127(46), pp.16074–16089.

Blackburn, G.M. et al., 2006. Nucleic acids in chemistry and biology Royal Society of Chemistry, Royal Society of Chemistry.

Crawford, G.E. et al., 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Research, 16(1), pp.123–131.

Eckstein, F., 1985. Nucleoside Phosphorothioates. Annual review of biochemistry, 54(1), pp.367–402.

Feinberg, A.P., 2007. Phenotypic plasticity and the epigenetics of human disease. Nature, 447(7143), pp.433–440.

Ito, S. et al., 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature, 466(7310), pp.1129–1133.

Kinsella, A.R., Smith, D.& Pickard, M., 1997. Resistance to chemotherapeutic antimetabolites: a function of salvage pathway involvement and cellular response to DNA damage. British Journal of Cancer, 75(7), p.935.

Kurreck, J., 2003. Antisense technologies. Improvement through novel chemical modifications. European journal of biochemistry / FEBS, 270(8), pp.1628–1644.

Lister, R. et al., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271), pp.315–322.

Mattick, J.S., 2007. A new paradigm for developmental biology. The Journal of experimental biology, 210(Pt 9), pp.1526–1547.

Oguey, C., Foloppe, N. & Hartmann, B., 2010. Understanding the sequence-dependence of DNA groove dimensions: implications for DNA interactions. PLoS ONE, 5(12), p.e15931.

Razin, A. & Szyf, M., 1984. DNA methylation patterns. Formation and function. Biochimica et biophysica acta, 782(4), pp.331–342.

Sintim, H.O. & Kool, E.T., 2006. Enhanced Base Pairing and Replication Efficiency of Thiothymidines, Expanded-size Variants of Thymidine. Journal of the American Chemical Society, 128(2), pp.396–397.

Weiner, A. et al., 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Research, 20(1), pp.90–100.

Zhao, Z. et al., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics, 38(11), pp.1341–1347.

7

Other publications

Dans, P.D., Pérez, A., **Faustino, I.**, Lavery, R., and Orozco, M., 2012. Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Research*, 40(21), pp.10668–10678.

Sciabola, S., Cao, Q., Orozco, M., **Faustino, I.**, and Stanton, R.V., 2012. Improved nucleic acid descriptors for siRNA efficacy prediction. *Nucleic Acids Research*, 41(3), pp.1383–1394.

Hospital, A., **Faustino, I.**, Collepardo-Guevara, R., González, C., Gelpí, J.L., and Orozco, M., 2013. NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Research*. Epub ahead of publication.

Ackowledgments

To Modesto, because you have always trusted on me, and to all group members who I have shared these years with.

And to all the people who has supported me, my family, my friends and of course, to you, Eva.

Thanks a lot to all of you.

A

Appendix: supplementary material from the publications

A.1 Unique tautomeric and recognition properties of thioketothymines?

Ignacio Faustino, Anna Aviñó, Iván Marchán, F. Javier Luque, Ramón Eritja and Modesto Orozco.

Journal of the American Chemical Society. 2009, 131(35), pp.12845–12853.

Figure S1

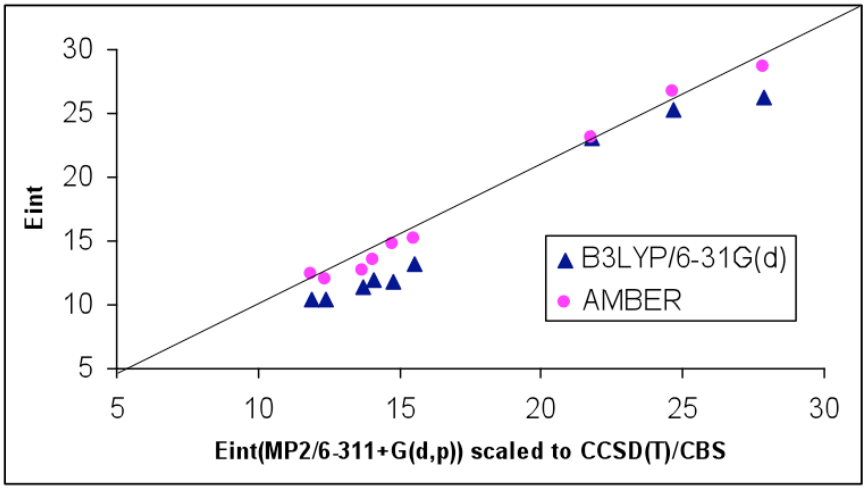


Figure S1. Correlation between AMBER (and AMBER-adapted) force-field calculations and QM reference values (MP2/6-311++G(d,p) scaled to reproduce CCSD(T)/CBS results). Correlations obtained from B3LYP/6-31G(d) calculations are shown as reference. See text for details.

Figure S2

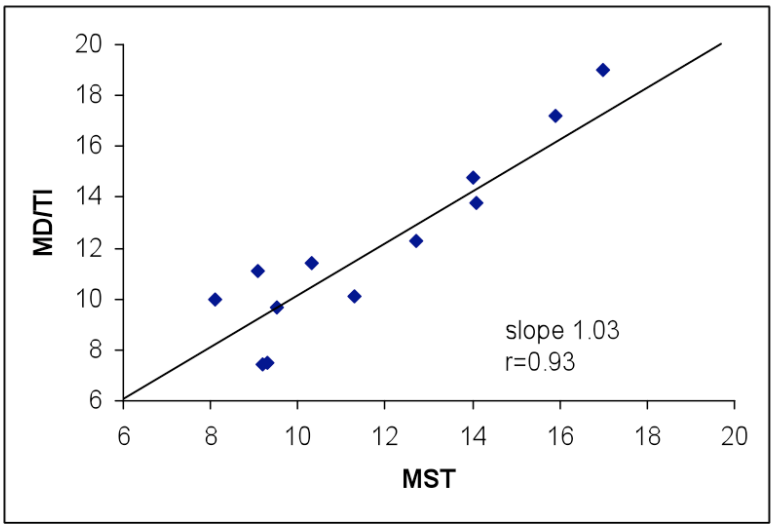


Figure S2. Correlation between MST and MD/TI estimates of the relative solvation free energy for the different species considered in the study. See text for details.

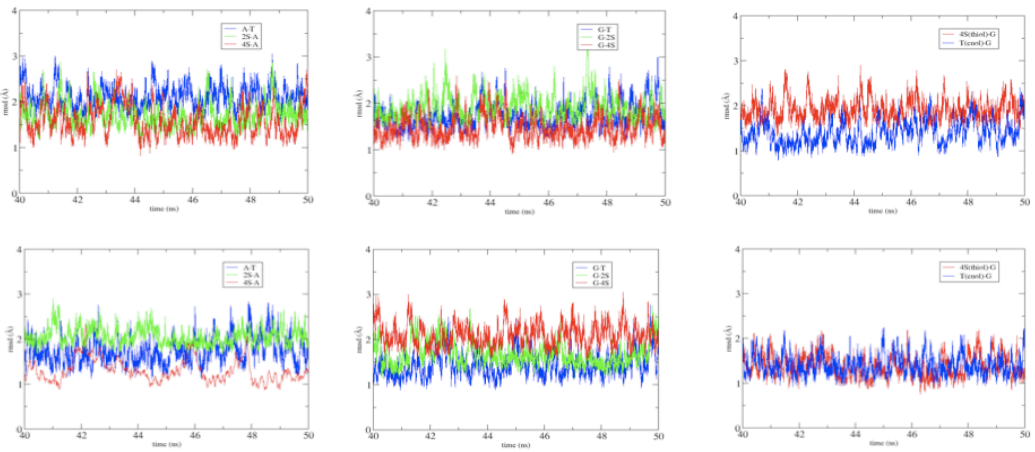


Figure S3. RMS deviations during the last 10 ns of the trajectories of d(CGCGA**X**TACGCG) duplexes displayed in the upper part and d(CGCGA**X**GACGCG) duplexes in the bottom. Values for duplexes containing T·A, ^{2}S·A and ^{4}S·A on the left; G·T, ^{2}S·G and ^{4}S·G in the center and ^{4}S(thiol)·G and T(enol)·G on the right. Average structure has been taken as reference structure in all cases.

Figure S4

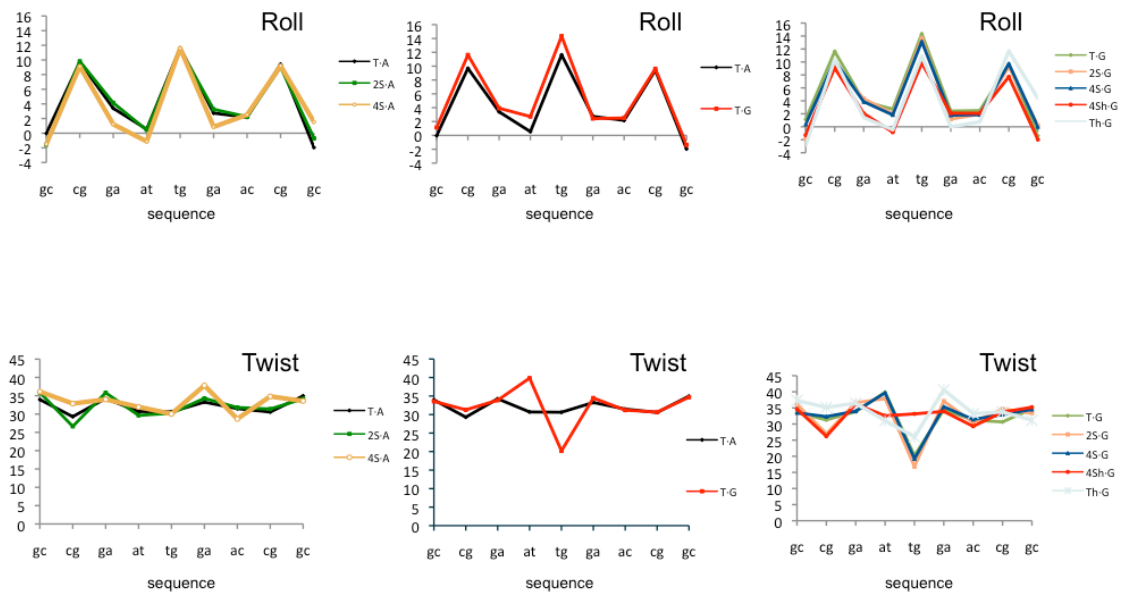


Figure S4. Change in selected helical parameters along the sequences (MD-averaged values) for different duplexes containing different pairings in the central triad d(AXG)·d(CYT); with A=X, ^{2}S or ^{4}S and Y=T or G (similar profiles are obtained for the other central triad). The label 4Sh refers to a ^{4}S in the 4-enol tautomeric form. All values are in degrees.

Figure S5

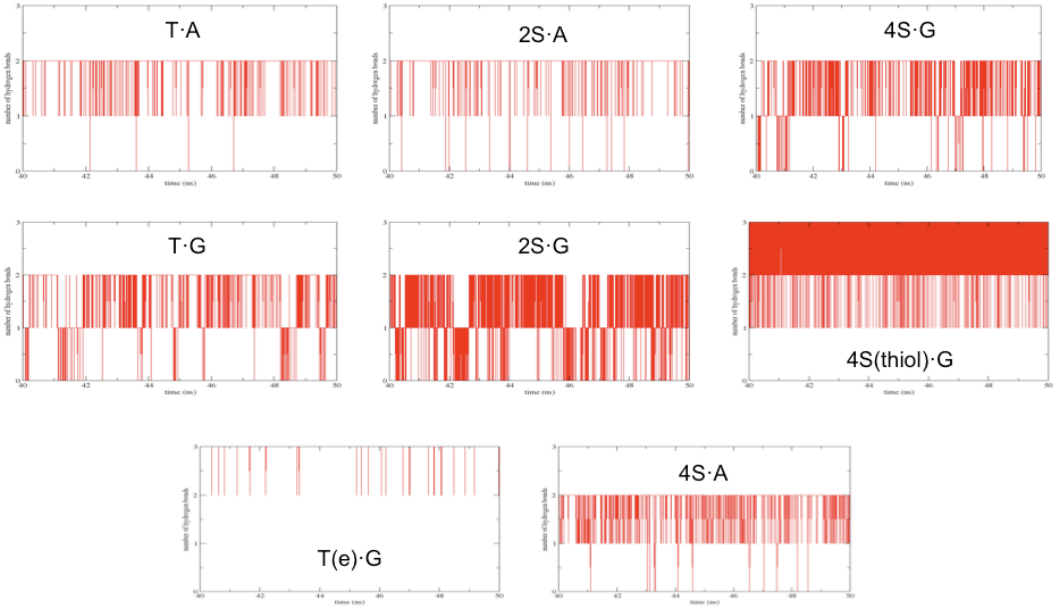


Figure S5. Variation along the last 10 ns of the number of hydrogen bonds at the mutation site for different pairings.

Figure S6

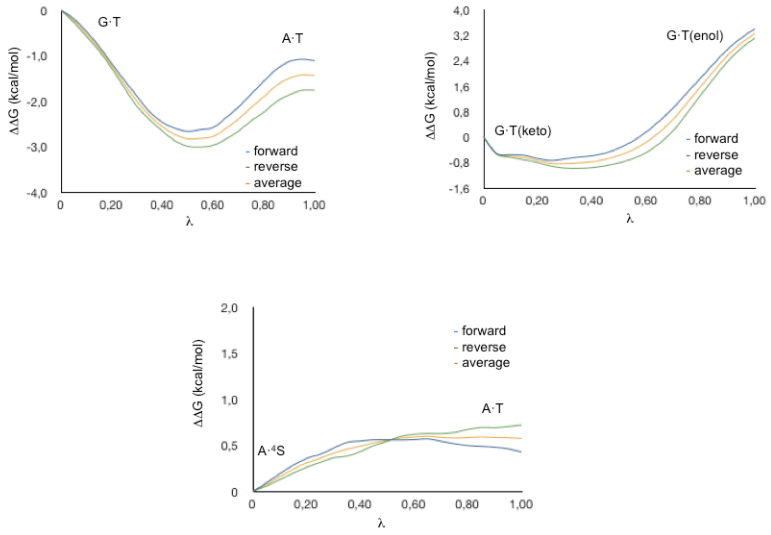


Figure S6. Randomly selected examples of individual mutations performed in this study. The smoothness of the free energies curves and the similarity between "forward" and "reverse" pathways demonstrate the lack of hysteresis effects.

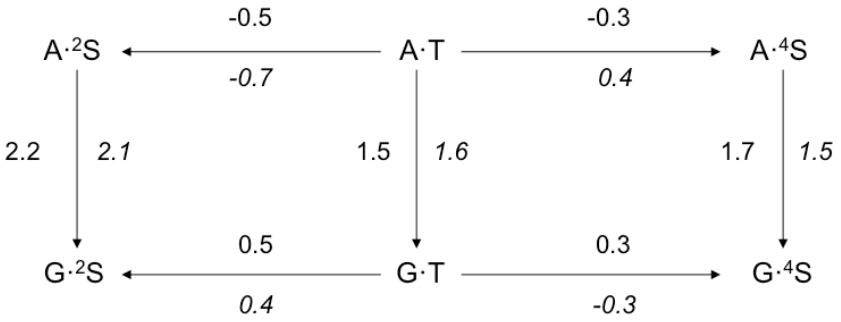


Figure S7. Futile cycles that can be defined from average free energy estimates displayed in Tables 3 and 5. Note that up to 6 futile cycles can be closed with reduced errors.

Force field parameters:

```
MASS
SS      32.064
SH      32.064

NONB
SS       2.0      0.5
SH       2.45     0.25

BOND
C -SS            570.00          1.66
C -SH            300.00          1.78

ANGL
NA-C –SS         80.0            122.00
CM-C -SS         80.0            122.48
N*-C -SS         80.0            124.23
NC-C -SH         80.0            119.14
N*-C -SH         80.0            115.91
C -SH-HS         35.0            91.06
CM-C -SH         80.0            117.70

DIHE
X   -C    -SH-X          2       126.60        180.0       2.0

IMPROPER
X -X -C -SS    10.5              180.0          2.0
X -X -C -SH    10.5              180.0          2.0
```

Table S1. Force-field parameters for thioketothymines used for MD simulations and MD/TI calculations.

(A)

Parameter	T·A	T·G	^{2}S·A	^{2}S·G	^{4}S·G	4S$_{thiol}$·G
Rise	3.44 3.41	3.42 3.38	3.37 3.38	3.38 3.42	3.34 3.43	3.45 3.47
Roll	5.22 5.05	5.77 4.52	4.68 4.62	5.10 4.97	5.13 5.17	3.64 4.15
Twist	32.04 32.13	32.33 31.95	32.25 32.42	31.59 32.47	32.57 31.96	32.67 32.22
Inclination	9.66 9.54	11.40 9.39	8.91 8.64	11.09 9.41	10.31 10.06	7.04 8.06
Opening	0.04 0.26	1.21 0.47	-0.49 -0.40	-0.19 -0.90	0.95 0.66	0.58 0.73
Prop twist	-11.15 -11.53	-11.82 -11.28	-11.82 -11.25	-11.06 -12.00	-11.69 -11.33	-10.94 -11.26
Phase	129.20 129.45	130.36 129.67	127.07 129.49	131.80 128.82	128.25 129.91	133.62 130.36
mG width	12.94 12.81	13.11 12.74	13.19 12.82	13.34 13.28	13.20 12.74	12.36 12.59
MG width	19.48 19.25	19.06 19.54	19.13 19.10	19.33 19.16	19.02 19.81	19.01 20.01

(B)

Parameter	T·A	T·G	2S·A	2S·G	4S·G	4S(t)·G
Rise	3.41 3.32	3.15 3.26	3.31 3.28	3.08 3.07	2.98 3.24	3.63 3.50
Roll	11.65 3.12	14.34 6.77	11.40 2.94	13.63 6.00	13.15 6.65	9.80 2.75
Twist	30.57 33.67	20.19 23.07	30.23 33.73	16.77 21.31	19.19 22.17	33.15 28.94
Inclination	21.03 5.55	36.30 16.57	21.00 5.18	37.14 15.93	34.68 16.64	16.83 5.91
Opening	0.26 0.67	12.82 3.54	-2.14 -1.12	2.83 -5.89	10.02 4.39	7.98 7.24
Prop twist	-12.08 -14.40	-15.66 -15.72	-11.66 -15.77	-12.59 -15.32	-12.77 -14.78	-18.19 -16.84
Phase	108.36 119.38	84.83 89.06	97.71 106.25	84.65 84.50	78.69 86.68	109.51 96.06
mG width	13.00 12.60	13.20 12.33	13.47 12.56	13.89 13.37	13.65 12.37	11.61 12.05
MG width	19.29 19.49	18.83 19.26	18.78 19.10	19.27 19.45	18.98 19.80	18.92 20.34

Table S2. DNA descriptors for the two sequences considered here with the different pairings. In each cell top values correspond to the duplex with the central A**X**G triad and down values to those for the duplex with the central A**X**T triad. (A) Values for the entire DNA (excluding ends). (B) values for the central triads. Angular parameters are in degree, translational parameters are in Å.

A.2 The DNA-forming properties of 6-selenoguanine.

Ignacio Faustino, Carles Curutchet, F. Javier Luque and Modesto Orozco.
In press.

Molecules	Name	Sequences
Duplexes	duplex/Se-duplex	5’-GCGCA**GXA**GCGC-3’ 3’-CGCGTCCTCGCG-5’
Triplexes	rH-triplex/rH-WCtriplex/ rH-Htriplex/H-triplex/H-WCtriplex	5’-CTCTCTCTCTCT-3’ 3’-GAGAGAXAGAGA-5’ (WC) 5’-GAGAGAXAGAGA-3’ (rH) or 3’ CTCTCTCTCTCT-5’ (H)
Aptamer	aptamer	5’-GXTTXGTGTXGTTXG-3’

Table 1. List of oligonucleotides studied and their corresponding sequences. X (= guanine/6-selenoguanine) stands for the mutation. H- and rH- stand for parallel (Hoogsteen interaction) and antiparallel (reverse Hoogsteen interaction) and WC and H stand for mutated strand.

	interaction pattern	BHandHLYP/cc-pVTZ	AMBER
G·C	WC-like	-25.6	-25.3
1H-6SeG·C	WC-like	-23.3	-22.1
6c-6SeG·C	WC-like	-10.7	
6t-6SeG·C	WC-like	-11.7	
G#G	rH-like	-17.2	-17.6
6SeG#G	rH-like	-17.8	-17.3
G#6SeG	rH-like	-18.6	-17.8
G-tetrad	H-like	-69.7	-67.0
6SeG-tetrad	H-like	-65.1	-63.2
G·C//G·C	Stacking	-14.8[a]	-14.5
G·C//6SeG·C	Stacking	-14.3[a]	-14.3

Table 2. Hydrogen bonding and stacking interaction energies (kcal/mol) computed at BHandHLYP/cc-pVTZ level of theory and corrected by the counterpoise method. For the rH-like interactions, the first nucleobase interacts through the Hoogsteen edge while the second one through its WC edge. G-tetrad was calculated and compared with the G-tetrad with one 6SeG mutation. For comparison, dimerization energies calculated by AMBER MD package for selected interactions are shown. Results shown here agree with previous studies and 6-thioguanine {Sponer:2004jx}. [a]Stacking energies calculated at M06-2X/6-31G** level.

	MP2/6-31G(d,p)	MP2/cc-pVDZ	MP2/cc-pVTZ	MP2/cc-pVQZ	CCSD(T)/CBS
1H-2c6c-6SeG	22.8	21.5	21.3	21.2	21.4
1H-2c6t-6SeG	22.8	21.5	21.4	21.2	21.4
1H-2t6c-6SeG	17.2	16.6	16.6	16.6	16.7
1H-2t6t-6SeG	20.0	19.2	19.0	18.9	18.9
1H-6SeG	0.0	0.0	0.0	0.0	0.0
3H-2c6c-6SeG	20.3	19.8	19.9	19.9	19.9
3H-2c6t-6SeG	20.3	19.5	19.3	19.1	19.2
3H-2t6c-6SeG	23.7	22.6	22.9	22.9	23.1
3H-2t6t-6SeG	27.8	26.7	26.4	26.2	26.3
3H-6SeG	17.7	17.4	17.3	17.2	17.4
6c-6SeG	-7.0	-7.6	-6.5	-6.2	-6.1
6t-6SeG	-7.4	-8.3	-7.5	-7.3	-7.2

Table 3. Tautomerization free energies (kcal/mol) relative to the 1H-6SeG tautomer at different levels of theory.

	$\Delta\Delta G_{B3LYP,\ hyd}$	$\Delta G_{taut,\ B3LYP}$	MD/TI
1H-2c6c-6SeG	-0.2	21.2	
1H-2c6t-6SeG	3.4	24.8	
1H-2t6c-6SeG	3.4	20.1	
1H-2t6t-6SeG	7.4	26.4	
1H-6SeG	0.0	0.0	
3H-2c6c-6SeG	4.9	24.8	
3H-2c6t-6SeG	5.3	24.4	
3H-2t6c-6SeG	5.4	28.5	
3H-2t6t-6SeG	0.1	26.4	
3H-6SeG	0.2	17.5	
6c-6SeG	9.9	3.9	*3.1*
6t-6SeG	10.4	3.2	*3.7*

Table 4. Hydration and tautomerization free energies (in kcal/mol) for the different 6-selenoguanine tautomers relative to the 1H-6SeG tautomer. MD/TI values (in kcal/mol) for the most stable tautomers are shown.

	ΔΔG (kcal/mol)
Se-duplex	5.2 ± 0.9
WC-Se-triplex	6.6 ± 2.0
rH-Se-triplex	4.2 ± 2.8
Se-aptamer	5.6 ± 1.0

Table 5. Differences in the Free Energy Changes (ΔΔG) associated to the G → 6SeG mutation.

	duplex	Se-duplex	triplex	WC-Se-triplex	rH-Se-triplex	aptamer	Se-aptamer
H-bonding	-27.3 ± 1.7	-21.5 ± 1.6	-45.1 ± 1.5 (-27.4, -17.7)	-39.4 ± 1.5 (-22.6, -16.8)	-42.9 ± 1.6 (-27.2, -15.7)	-98.1 ± 1.8	-80.1 ± 1.5
Stacking	-24.0 ± 1.0	-25.9 ± 1.0	-49.0 ± 1.1	-47.7 ± 1.0	-47.1 ± 1.2	-19.2 ± 1.5	-17.2 ± 1.3
K^+-tetrads						-147.8 ± 1.8	-150.5 ± 1.7
Total	-51.3 ± 1.2	-47.4 ± 1.1	-94.1 ± 1.1	-87.1 ± 1.1	-90.0 ± 1.2	-265.1 ± 1.3	-247.8 ± 1.2

Table 6. Interaction energies (H-bond, stacking) and standard deviations for duplex, triplex and aptamer structures around the mutation site and K^+-nucleobases interaction energies for the two tetrads of the aptamer and the modified aptamer with single mutation. Energies are in kcal/mol. H-bonding energies correspond to base pairs at the mutation site while stacking energies correspond to the central three base pairs around the mutation site. Values in parentheses correspond to Watson-Crick and reverse-Hoogsteen hydrogen bond interactions.

		HOMO	LUMO	gap
	G	-5.00	-0.94	4.05
	6SeG	-4.26	-1.84	2.42
H-bonded	G·C	-4.34	-1.95	2.39
	6SeG·C	-4.15	-1.98	2.17
	C·G#G	-4.00	-2.31	1.69
	C·6SeG#G	-4.11	-2.30	1.81
	C·G#6SeG	-3.58	-2.34	1.24
	G4	-4.55	-0.85	3.70
	G3-6SeG	-4.08	-1.61	2.47

Stacked	$(G\text{-}C)_2$	-4.44	-2.53	1.91
	(G-C/6SeG-C)	-3.96	-2.35	1.61
	$(6SeG\text{-}C)_2$	-4.11	-2.46	1.65

Table 7. Energies of HOMO and LUMO orbitals, and HOMO-LUMO gap for isolated bases and base pairs calculated with M062X/6-31++G(d,p) (in eV).

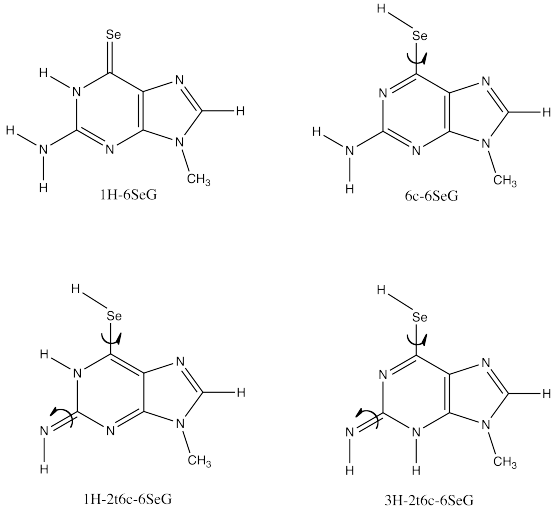


Figure 1. Considered tautomers in our QM calculations.

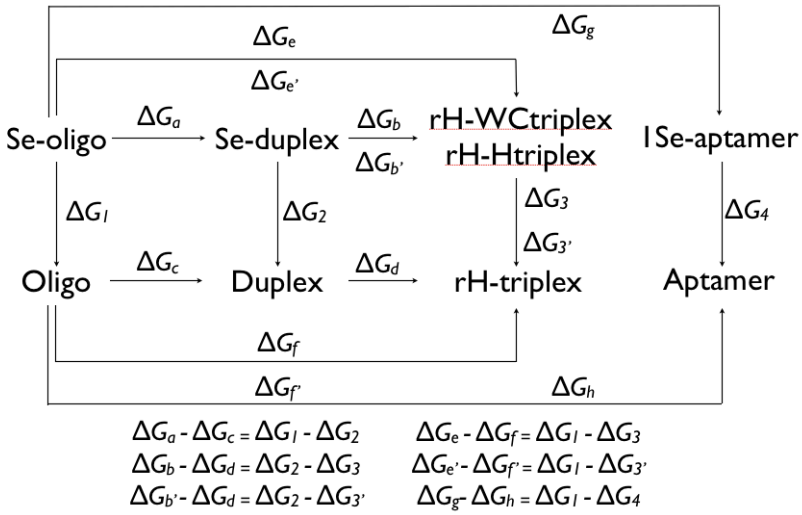


Figure 2. Thermodynamic cycles used to estimate the free energy change in stability for the 6SeG → G mutation when incorporating the 6-selenoguanine in single, duplex and triplex structures.

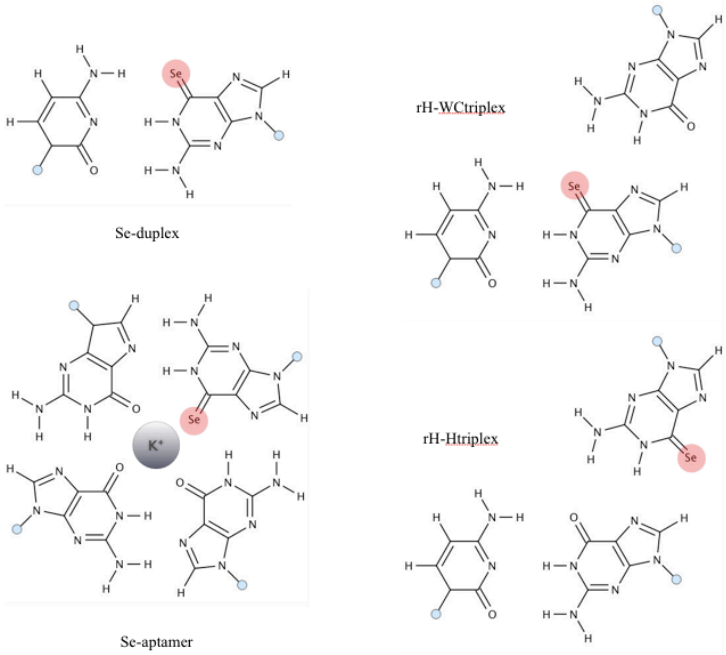


Figure 3. Structures of duplex, triplexes and G-quadruplex tetrad containing 6-selenoguanine.

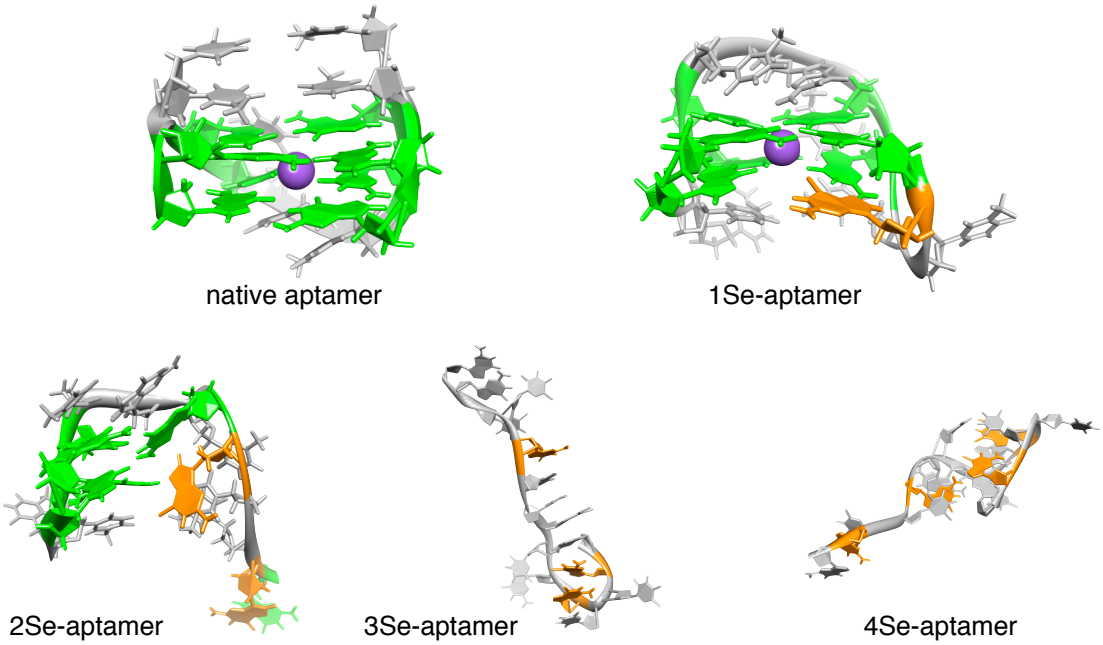


Figure 4. Final structures of the MD simulations of the non-modified (a) and modified aptamers with one (b), two (c), three (d), and four 6SeG (e) residues in the same tetrad. Central K^+ ion is depicted as a purple sphere.

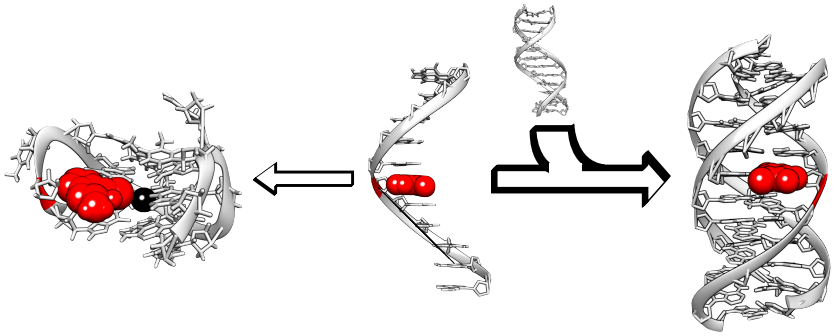


Figure 5. Schematic representation of the thermodynamical preferences of a modified single stranded oligonucleotide to form either triplex or G-quadruplex structures. The 6-selenoguanine is represented in red.

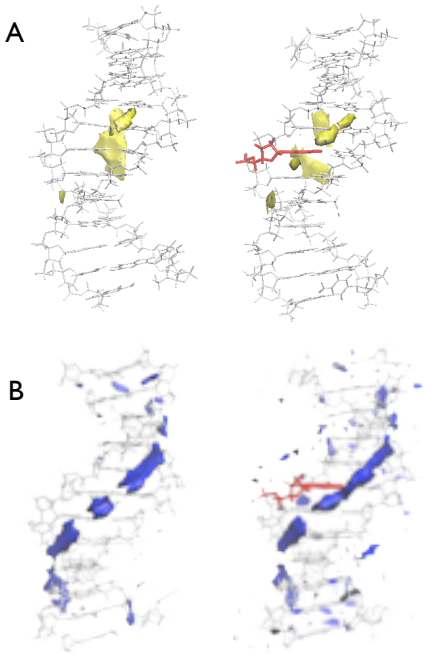


Figure S1. MIP profiles and solvation maps for duplex (left) and Se-duplex (right) molecules. (A) The probe molecule for MIP calculations was Na^+ and the depicted contour corresponds to -4.5 kcal/mol. (B) Solvation maps depicted with contours at 2.0 g/mL density. The 6-selenoguanine residue is depicted in red for the MIP plots and in yellow for the water density profiles.

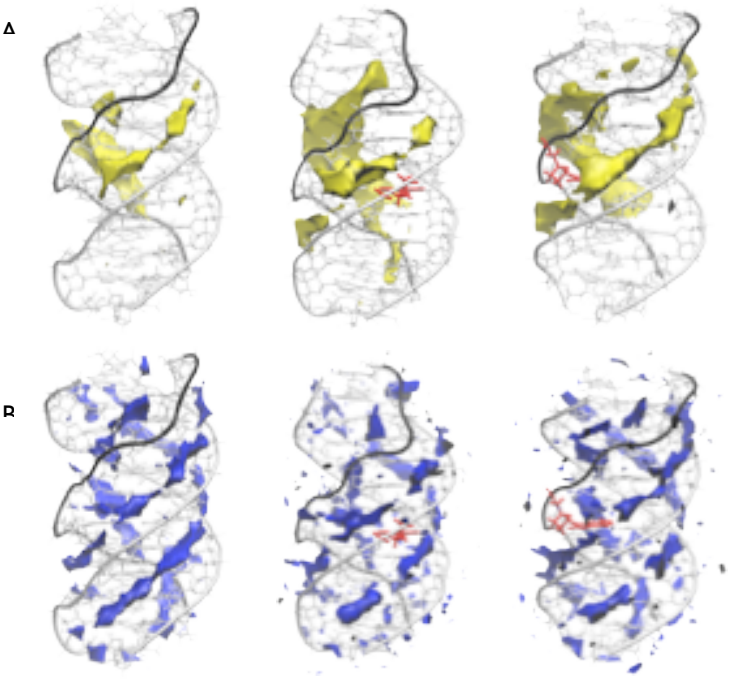


Figure S2. MIP profiles (A) and solvation maps (B) for canonical (left), WC-modified (center) and H-modified (right) triplexes. (A) The probe molecule for MIP calculations was Na^+ and the depicted contour corresponds to -4 kcal/mol. (B) Solvation maps depicted with contours at 2.0 g/mL density. The Hoogsteen strand is depicted with black backbone trace while the 6-selenoguanine residue is depicted here in red.

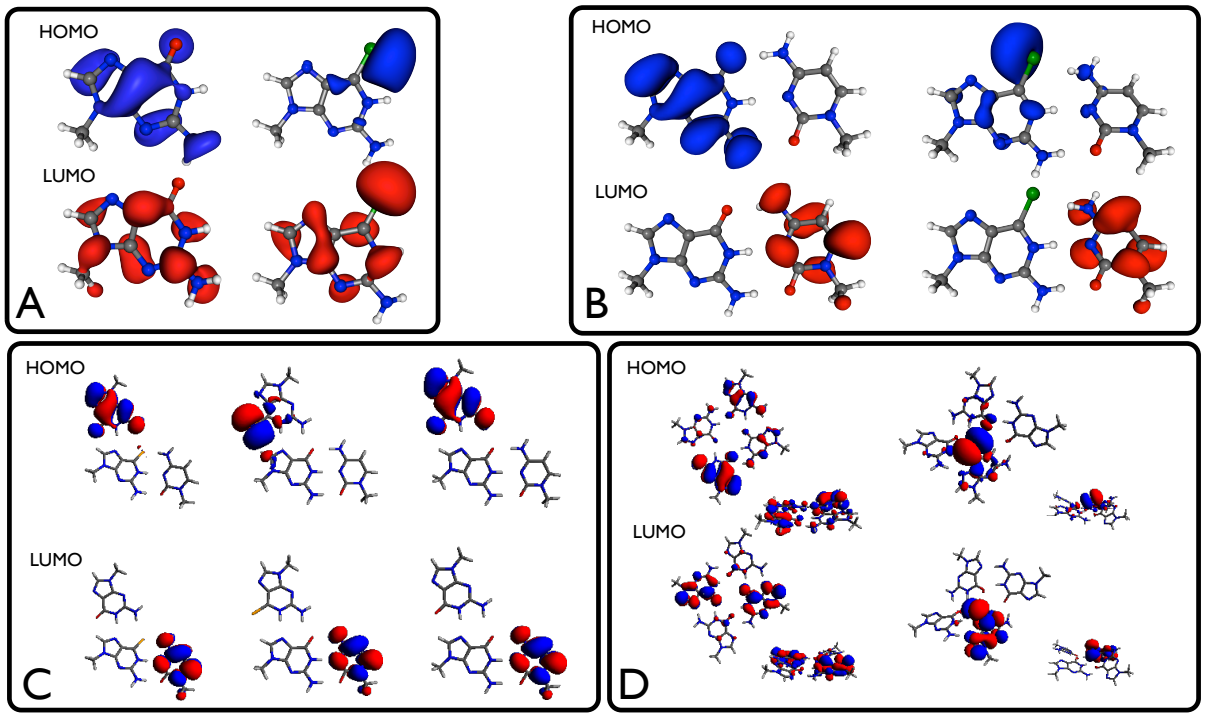


Figure S3. M06-2X/6-31G** contour plot of electron density (isodensity at 0.02 au) of HOMO (blue) and LUMO (red) for monomers (A), H-bonded dimers (B, G·C dimer on the left and 6SeG·C dimer on the right), H-bonded triads (C, C·6SeG#G triad on the left, C·G#6SeG triad in the center and C·G#G on the right) and G-quartets (D, G4 on the left and G4 with one 6SeG on the right).

	μvac	Δμ
1H-2c6c-6SeG	4.7	1.5
1H-2c6t-6SeG	4.7	1.5
1H-2t6c-6SeG	3.6	0.5
1H-2t6t-6SeG	4.9	1.4
1H-6SeG	9.1	3.2
3H-2c6c-6SeG	6.3	2.1
3H-2c6t-6SeG	7.4	2.7
3H-2t6c-6SeG	8.8	3.2
3H-2t6t-6SeG	9.8	3.7

	μvac	Δμ
3H-6SeG	13.0	6.2
6c-6SeG	4.0	1.2
6t-6SeG	4.5	1.5

Table S1. Gas phase and difference between water-induced and gas phase dipole moments (in Debye) for the 6-SeG tautomers. The values in gas phase and in water solution were determined at the MP2/cc-pVQZ and the B3LYP/cc-pVQZ levels of theory respectively.

	Slide		Rise		Twist	
	Duplex	Se-duplex	Duplex	Se-duplex	Duplex	Se-duplex
6-GX/CC	-1.1 ± 0.8	-1.0 ± 0.7	3.6 ± 0.4	3.7 ± 0.4	30.9 ± 5.2	32.3 ± 5.1
7-XA/TC	-0.5 ± 0.7	-0.7 ± 0.6	3.3 ± 0.3	3.3 ± 0.3	33.9 ± 5.2	32.7 ± 5.2
	Stretch		Buckle		Opening	
	Duplex	Se-duplex	Duplex	Se-duplex	Duplex	Se-duplex
6-GC	0.02 ± 0.1	0.02 ± 0.1	5.6 ± 9.8	4.2 ± 9.9	1.0 ± 3.2	1.2 ± 3.2
7-XC	0.03 ± 0.1	0.24 ± 0.1	-4.9 ± 12.4	-6.7 ± 12.1	1.6 ± 3.7	6.2 ± 3.8
8-AT	0.04 ± 0.1	0.04 ± 0.1	-1.4 ± 12.5	-6.0 ± 11.5	3.2 ± 5.9	3.3 ± 6.1

Table S2. Averaged values and standard errors of selected helical parameters calculated for duplex structures (X stands for G or SeG) and adjacent base pairs using the Curves+ program (see Methods).

WC	Rise			Roll			Twist		
	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex
6-CT/AX	3.4 ± 0.3	3.7 ± 0.4	3.4 ± 0.4	-0.7 ± 4.4	-1.2 ± 5.3	0.6 ± 4.7	28.0 ± 3.5	24.4 ± 3.2	25.2 ± 3.7
7-TC/XA	3.3 ± 0.3	3.4 ± 0.3	3.2 ± 0.3	-0.4 ± 5.0	-2.8 ± 5.2	-2.3 ± 5.6	33.9 ± 3.9	33.1 ± 4.1	32.6 ± 3.8

WC	Stretch			Buckle			Opening		
	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex
6-TA	0.04 ± 0.1	0.04 ± 0.2	0.04 ± 0.3	-5.5 ± 9.3	-1.7 ± 8.9	-9.0 ± 7.1	1.2 ± 5.7	4.2 ± 7.1	1.8 ± 10.3
7-CX	0.03 ± 0.1	0.31 ± 0.1	0.05 ± 0.1	2.9 ± 9.1	9.3 ± 9.0	0.6 ± 8.7	1.3 ± 3.1	7.8 ± 3.4	1.9 ± 3.2
8-TA	0.03 ± 0.1	0.07 ± 0.2	0.04 ± 0.2	0.2 ± 8.6	-1.3 ± 8.7	0.2 ± 9.8	5.6 ± 5.1	8.0 ± 5.9	6.7 ± 5.5

Table S3. Averaged values and standard errors of selected helical parameters for the WC duplex calculated for triplex structures (X stands for G or SeG) and adjacent base pairs using the Curves+ program (see Methods).

rH	Rise			Roll			Twist		
	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex
6-AX/XA	3.1 ± 0.2	3.1 ± 0.3	3.4 ± 0.4	-3.7 ± 6.1	-1.6 ± 6.4	-2.6 ± 7.3	27.9 ± 4.6	28.5 ± 5.0	22.8 ± 14.3
7-XA/AX	3.3 ± 0.3	3.3 ± 0.3	3.3 ± 0.3	13.1 ± 8.4	10.5 ± 8.6	11.3 ± 7.8	26.4 ± 3.5	27.4 ± 4.0	23.6 ± 3.2
rH	Stretch			Buckle			Opening		
	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex	triplex	WC-Se-triplex	rH-Se-triplex
6-AA	0.82 ± 0.3	1.00 ± 0.4	0.84 ± 2.8	-18.2 ± 9.6	-18.9 ± 12.9	-13.9 ± 12.5	-99.7 ± 4.9	-102.2 ± 5.3	-99.2 ± 28.6
7-XX	-0.89 ± 0.4	-0.79 ± 0.5	-1.21 ± 0.3	9.4 ± 11.0	7.0 ± 11.9	8.9 ± 13.8	-105.5 ± 5.3	-109.9 ± 5.8	-103.7 ± 7.0
8-AA	1.29 ± 0.4	1.19 ± 0.5	1.31 ± 0.4	-20.8 ± 9.3	-23.3 ± 10.3	-20.8 ± 10.9	-105.5 ± 6.2	-105.4 ± 7.5	-103.2 ± 5.7

Table S4. Averaged values and standard errors (in parenthesis) of selected helical parameters for the rH duplex calculated for triplex structures (X stands for G or SeG) and adjacent base pairs using the Curves+ program (see Methods).

A.3 Functionalization of the 3'-Ends of DNA and RNA Strands with N-ethyl-N-coupled Nucleosides: A Promising Approach To Avoid 3'-exonuclease-Catalyzed Hydrolysis of Therapeutic Oligonucleotides.

Montserrat Terrazas, Adele Alagia, **Ignacio Faustino**, Modesto Orozco and Ramón Eritja.

ChemBioChem. 2013, 14(4), pp.510–520.

CHEMBIOCHEM

Supporting Information

© Copyright Wiley-VCH Verlag GmbH & Co. KGaA, 69451 Weinheim, 2013

Functionalization of the 3′-Ends of DNA and RNA Strands with N-ethyl-N-coupled Nucleosides: A Promising Approach To Avoid 3′-Exonuclease-Catalyzed Hydrolysis of Therapeutic Oligonucleotides

Montserrat Terrazas,*[a] Adele Alagia,[a] Ignacio Faustino,[b, c] Modesto Orozco,[b, c] and Ramon Eritja*[a]

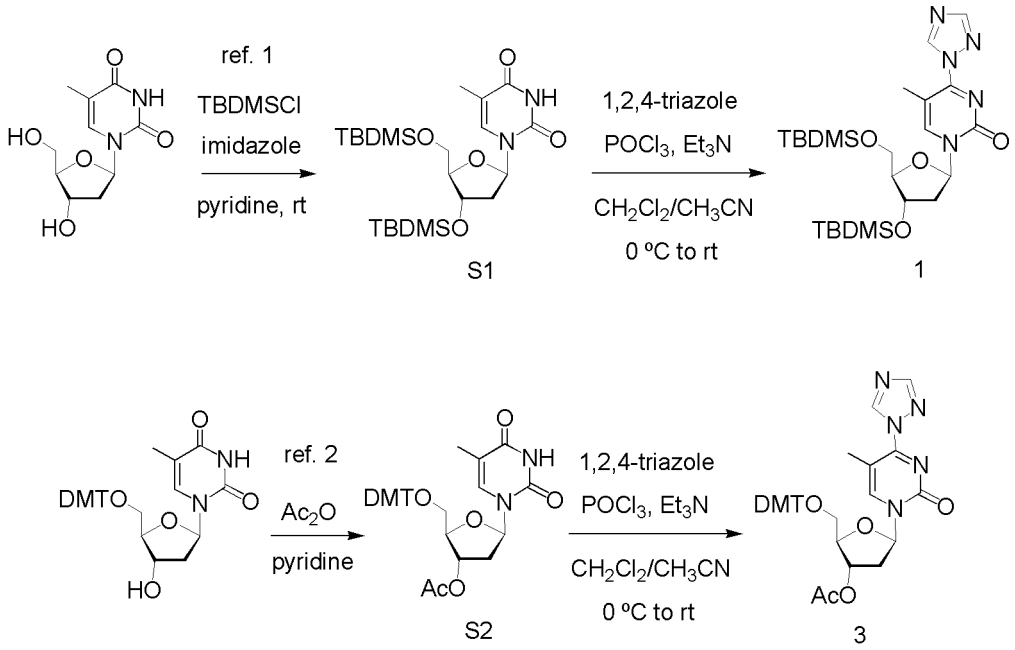
cbic_201200611_sm_miscellaneous_information.pdf

General experimental methods

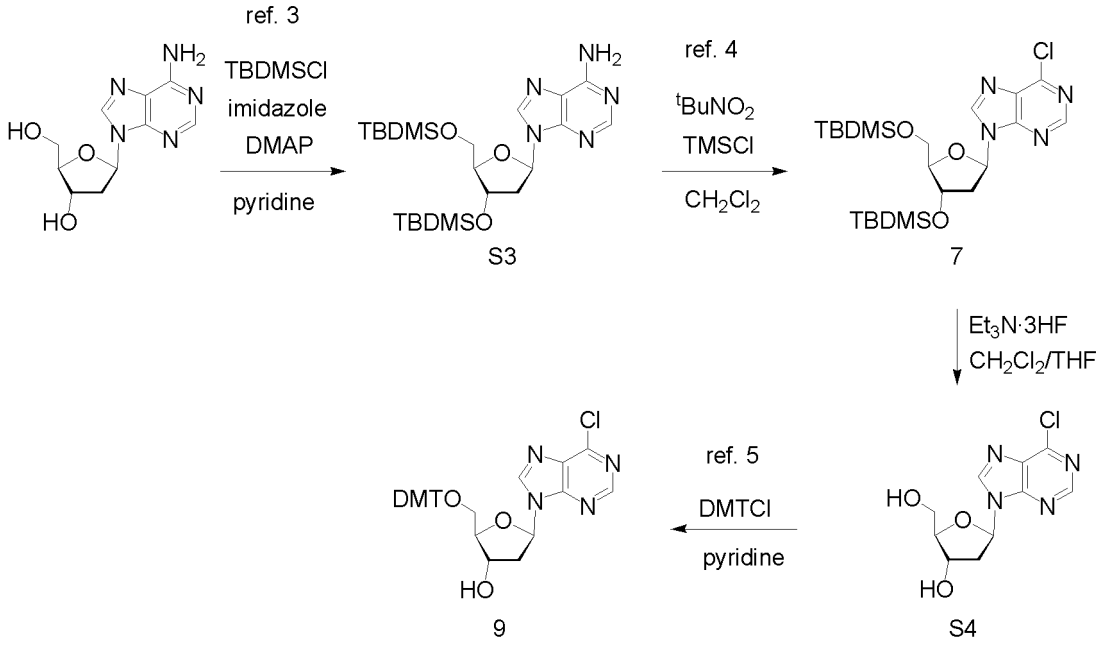
Common chemicals and solvents in addition to 2-cyanoethyl diisopropyl-phosphoramidochloridite were purchased from commercial sources and used without further purification. Anhydrous solvents and deuterated solvents ($CDCl_3$ and DMSO-d_6) were obtained from reputable sources and used as received.
Reagents for oligonucleotide synthesis including 2'-*O*-TBDMS-protected phosphoramidite monomers of A^{Bz}, C^{Ac}, G^{dmf} and U, the 5'-deblocking solution (3% TCA in CH_2Cl_2), activator solution (0.4 M 1*H*-tetrazole in CH_3CN), CAP A solution (acetic anhydride/pyridine/THF), oxidizing solution (0.02 M iodine in tetrahydrofuran/pyridine/water (7:2:1), sulfurizing reagent (0.49 M tetraethylthiuram disulfide, TETD, in CH_3CN), succinyl polystyrene functionalized with 5'-*O*-DMT-thymidine and with 5'-*O*-DMT-2'-deoxyadenosine, and LCAA-CPG were purchased from commercial sources and used as received.
Phosphodiesterase I from *Crotalus adamanteus* venom (SNVPD) and Large (Klenow) Fragment of *E. coli* DNA polymerase I were commercially available (Sigma-Aldrich and Invitrogen, respectively) and used without further purification.
All reactions were carried out under argon atmosphere in oven-dried glassware. Thin-layer chromatography was carried out on aluminium-backed Silica-Gel 60 F_{254} plates. Column chromatography was performed using Silica Gel (60 Å , 230 x 400 mesh). NMR spectra were measured on Varian Mercury-400, Varian-500 or Varian-500 instruments. Chemical shifts are given in parts per million (ppm); *J* values are given in hertz (Hz). All spectra were internally referenced to the appropriate residual undeuterated solvent.

RP-HPLC purifications were performed using a Nucleosil 120-10 C18 column (250x4 mm).

HRMS and ESI spectra were performed on a LC/MSD-TOF (Agilent technologies) mass spectrometer.
MALDI-TOF spectra were performed using a Perspective Voyager DETMRP mass spectrometer, equipped with nitrogen laser at 337 nm using a 3ns pulse. The matrix used contained 2,4,6-trihydroxyacetophenone (THAP, 10 mg/mL in CH_3CN/water 1:1) and ammonium citrate (50 mg/mL in water).



Scheme S1. Synthesis of O^4-triazolyl intermediates



Scheme S2. Synthesis of 6-chloropurine intermediates

3',5'-Di-*O-tert*-butyldimethylsilylthymidine (S1):[1] Thymidine (1.50 g, 6.19 mmol) and imidazole (2.11 g, 30.95 mmol) were dissolved in DMF (15 mL) and stirred for 5 min at rt. TBDMSCl (2.24 g, 14.86 mmol) was added and the reaction stirred for 48 h at rt. The reaction mixture was diluted with EtOAc and washed with water. The aqueous layer was extracted with EtOAc. The organic layers were combined, dried over Na_2SO_4, and evaporated under reduced pressure. The residue that was obtained was purified by column chromatography eluting with CH_2Cl_2/MeOH (95:5) to give **S1** as a colorless liquid (2.86 g, 98%). ^{1}H NMR ($CDCl_3$, 500 MHz) δ 8.69 (bs, 1H), 7.48 (s, 1H), 6.34 (dd, *J* = 6.0 Hz, *J* = 8.0 Hz, 1H), 4.41 (m, 1H), 3.94 (m, 1H), 3.87 (d, *J* = 2.5 Hz, *J* = 11.0 Hz, 1H), 3.76 (dd, *J* = 2.0 Hz, *J* = 11.0 Hz, 1H), 2.25 (ddd, *J* = 2.5 Hz, *J* = 6.0 Hz, *J* = 13.0 Hz, 1H), 2.01 (ddd, *J* = 6.5 Hz, *J* = 8.0 Hz, *J* = 13.0 Hz, 1H), 1.92 (s, 3H), 0.93 (s, 9H), 0.90 (s, 9H), 0.12 (s, 3H), 0.11 (s, 3H), 0.09 (s, 3H), 0.08 (s, 3H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 164.1, 150.4, 135.4, 110.8, 87.7, 84.7, 72.1, 62.9, 41.3, 25.9, 25.7, 18.3, 17.9, 12.5, -4.7, -4.9, -5.4, -5.5.

3',5'-Di-*O-tert*-butyldimethylsilyl-4-(*N*-1-triazolyl)thymidine (1): A suspension of 1,2,4-triazole (3.15 g, 45.52 mmol) in a CH_2Cl_2/CH_3CN mixture (1:1, 50 mL) was treated with Et_3N (9.12 mL, 65.44 mmol) and the mixture stirred at 0 ºC for 5 min. Phosphorous oxychloride (663 µL, 7.11 mmol) was slowly added. After stirring at 0 ºC for 30 min, a solution of **S1** (1.34 g, 2.85 mmol) in CH_2Cl_2 (3.3 mL) was added. After stirring at room temperature for 50 min, the starting material was completely converted to the O^4-triazolyl intermediate **1** as evidenced by TLC. The reaction mixture was diluted with CH_2Cl_2 and washed with 5% $NaHCO_3$ followed by saturated NaCl. The aqueous layers were extracted with CH_2Cl_2. The organic layers were combined, dried over $MgSO_4$ and evaporated under reduced pressure to give the O^4-triazolyl intermediate **1** as a yellow foam (1.63 g), which was used without further purification. ^{1}H NMR ($CDCl_3$, 500 MHz) δ 9.29 (s, 1H), 8.26 (s, 1H), 8.11 (s, 1H), 6.30 (dd, *J* = 6.5 Hz, *J* = 6.0 Hz, 1H), 4.40 (m, 1H), 4.07 (m, 1H), 3.97 (dd, *J* = 2.5 Hz, *J* = 11.5 Hz, 1H), 3.80 (dd, *J* = 2.0 Hz, *J* = 11.5 Hz, 1H), 2.65 (ddd, *J* = 4.0 Hz, *J* = 6.0 Hz, *J* = 13.5 Hz, 1H), 2.45 (s, 3H), 2.08 (ddd, *J* = 6.5 Hz, *J* = 7.0 Hz, *J* = 13.5 Hz, 1H), 0.92 (s, 9H), 0.90 (s, 9H), 0.13 (s, 3H), 0.11 (s, 3H), 0.09 (s, 3H), 0.08 (s, 3H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 158.0, 153.9, 153.3, 146.6, 145.0, 105.2, 88.7, 87.7, 71.6, 62.5, 42.6, 25.9, 25.7, 18.4, 17.9, 8.6, -4.6, -4.9, -5.4, -5.4. HRMS (ES+): calculated for $C_{24}H_{44}N_5O_4Si_2$ $[M+H]^+$ 522.2926, found 522.2921.

3',5'-Di-*O-tert*-butyldimethylsilyl-N^4-(2-aminoethyl)-2'-deoxycytidine (2): A solution of O^4-triazolyl intermediate **1** (1.1 g, 1.9 mmol) in anhydrous pyridine (84 mL) was treated dropwise with ethylenediamine (5.2 mL, 78.1 mmol). After stirring for 15 h at rt, the solvents were evaporated. Residual pyridine was removed by co-evaporation with toluene followed by 95% ethanol. The crude product was

purified by silica gel chromatography with 20% MeOH and 4% Et_3N in CH_2Cl_2 to give **2** as a yellow oil (830 mg, 89%). ^{1}H NMR ($CDCl_3$, 400 MHz) δ 7.45 (s, 1H), 6.36 (dd, *J* = 6.6 Hz, *J* = 6.3 Hz, 1H), 5.79 (bs, 1H), 4.34 (m, 1H), 3.87-3.83 (m, 2H), 3.75 (m, 1H), 3.58 (bs, 2H), 2.93 (t, *J* = 5.7 Hz, 2H), 2.33 (m, 1H), 2.10 (bs, 2H), 1.95 (m, 1H), 1.91 (s, 3H), 0.90 (s, 9H), 0.86 (s, 9H), 0.09 (s, 3H), 0.08 (s, 3H), 0.04 (s, 3H), 0.03 (s, 3H). ^{13}C NMR ($CDCl_3$, 75 MHz) δ 163.2, 156.3, 136.7, 113.2, 101.9, 87.3, 85.5, 71.6, 62.6, 42.6, 41.9, 40.6, 25.9, 25.7, 18.3, 17.9, 13.1, -4.6, -4.9, -5.4, -5.5. HRMS (ES+): calculated for $C_{24}H_{49}N_4O_4Si_2$ $[M+H]^+$ 513.3287, found 513.3281.

3'-*O*-Acetyl-5'-*O*-(4,4'-dimethoxytrityl)thymidine (S2):[2] 5'-*O*-(4,4'-Dimethoxytrityl)thymidine (Peninsula Laboratories, Inc.) (700 mg, 1.29 mmol) and DMAP (16 mg, 0.13 mmol) were dissolved in pyridine (5 mL). Acetic anhydride (304 μL, 3.22 mmol) was added and the reaction mixture was stirred at room temperature. After 15 h, the reaction was quenched by addition of MeOH and the solvent was evaporated. The crude product was purified by silica gel column chromatography with 5% MeOH in CH_2Cl_2 to give **S2** (756 mg, 99%) as a white foam. ^{1}H NMR ($CDCl_3$, 400 MHz) δ 9.52 (bs, NH), 7.63 (m, 1H), 7.42-7.27 (m, 9H), 6.86 (d, *J* = 8.8 Hz, 4H), 6.46 (m, 1H), 5.46 (m, 1H), 4.15 (m, 1H), 3.81 (s, 6H), 3.51 (dd, *J* = 2.8 Hz, *J* = 10.8 Hz, 1H), 3.47 (dd, *J* = 2.4 Hz, *J* = 10.8 Hz, 1H), 2.52-2.41 (m, 2H), 2.11 (s, 3H), 1.41 (s, 3H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 170.4, 163.6, 158.7, 150.4, 144.2, 135.3, 135.1, 130.1, 130.0, 128.1, 128.0, 127.2, 113.3, 111.7, 87.1, 84.3, 84.0, 75.4, 63.7, 55.2, 21.0, 11.3. HRMS (ES+): calculated for $C_{33}H_{34}N_2NaO_8$ $[M+Na]^+$ 609.2207, found 609.2207.

3'-*O*-Acetyl-5'-*O*-(4,4'-dimethoxytrityl)-4-(*N*-1-triazolyl)thymidine (3): A suspension of 1,2,4-triazole (990 mg, 14.33 mmol) in a CH_2Cl_2/CH_3CN mixture (1:1, 15 mL) was treated with Et_3N (2.87 mL, 20.61 mmol) and the mixture stirred at 0 ºC for 5 min. Phosphorous oxychloride (209 μL, 2.24 mmol) was slowly added. After stirring at 0 ºC for 30 min, a solution of **S2** (525 mg, 0.90 mmol) in CH_2Cl_2 (2 mL) was added. After stirring at rt for 50 min, the starting material was completely converted to the O^4-triazolyl intermediate **3** as evidenced by TLC. The reaction mixture was diluted with CH_2Cl_2 and washed with 5% $NaHCO_3$ followed by saturated NaCl. The aqueous layers were extracted with CH_2Cl_2. The organic layers were combined, dried over $MgSO_4$ and evaporated under reduced pressure to give the O^4-triazolyl intermediate **3** as a yellow foam (558 mg), which was used without further purification. ^{1}H NMR ($CDCl_3$, 300 MHz) δ 9.28 (s, 1H), 8.30 (s, 1H), 8.08 (s, 1H), 7.38-7.20 (m, 9H), 6.82 (d, *J* = 9.0 Hz, 4H), 6.40 (dd, *J* = 5.7 Hz, *J* = 8.1 Hz, 1H), 5.42 (m, 1H), 4.28 (m, 1H), 3.77 (s, 6H), 3.52 (dd, *J* = 3.0 Hz, *J* = 10.8 Hz, 1H), 3.44 (dd, *J* = 3.0 Hz, *J* = 10.8 Hz, 1H), 2.86 (ddd, *J* = 1.8 Hz, *J* = 5.7 Hz, *J* = 14.4 Hz, 1H), 2.40 (ddd, *J* = 6.3 Hz, *J* = 8.1 Hz, *J* = 14.4 Hz, 1H), 2.09 (s, 3H), 1.99 (s, 3H). ^{13}C NMR ($CDCl_3$, 75 MHz) δ 170.3, 158.7, 158.2, 153.4, 146.5, 145.0, 144.0, 135.1, 135.0, 129.9, 128.0, 127.9, 127.2, 113.3,

106.1, 87.3, 87.2, 85.0, 74.9, 63.3, 55.2, 20.9, 8.6. HRMS (ES+): calculated for $C_{35}H_{36}N_5O_7$ $[M+H]^+$ 638.2609, found 638.2613.

1-{*N*^{4}-[3'-*O*-Acetyl-5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[*N*^{4}-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]ethane (4): To a solution of 3',5'-di-*O*-*tert*-butyldimethylsilyl-*N*^{4}-(2-aminoethyl)-2'-deoxycytidine (**2**, 857 mg, 1.67 mmol) in pyridine (18 mL) was added Et_3N (1.3 mL, 9.54 mmol), followed by 3'-*O*-acetyl-5'-*O*-(4,4'-dimethoxytrityl)-4-(*N*-1-triazolyl)thymidine (**3**, 427 mg, 0.67 mmol). The reaction mixture was allowed to stir at rt for 15 h. The solution was evaporated to dryness. Silica gel column chromatography using 5% MeOH in CH_2Cl_2 yielded **4** (666 mg, 92%) as a white foam. ^{1}H NMR ($CDCl_3$, 300 MHz) δ 7.56 (bs, 1H), 7.45 (s, 1H), 7.41-7.23 (m, 9H), 6.85 (d, *J* = 8.9 Hz, 4H), 6.52 (dd, *J* = 5.1 Hz, *J* = 8.7 Hz, 1H), 6.39 (dd, *J* = 6.3 Hz, *J* = 7.2 Hz, 1H), 5.40 (m, 1H), 4.40 (m, 1H), 4.15 (m, 1H), 3.93-3.75 (m, 13H), 3.48 (dd, *J* = 3.0 Hz, *J* = 10.8 Hz, 1H), 3.41 (dd, *J* = 3.3 Hz, *J* = 10.8 Hz, 1H), 2.55 (m, 1H), 2.37-2.25 (m, 2H), 2.07 (s, 3H), 2.00-1.91 (m, 4H), 1.61 (s, 3H), 0.94 (s, 9H), 0.90 (s, 9H), 0.12 (s, 6H), 0.07 (s, 6H). ^{13}C NMR ($CDCl_3$, 75 MHz) δ 170.4, 146.2, 163.9, 158.6, 156.2, 156.1, 144.4, 136.2, 135.3, 135.2, 130.0, 129.9, 128.0, 127.9, 127.0, 113.2, 103.7, 103.0, 87.4, 86.8, 85.4, 85.2, 83.5, 75.5, 72.0, 63.7, 62.8, 55.2, 43.1, 41.8, 38.6, 25.9, 25.7, 21.0, 18.3, 17.9, 13.5, 12.8, -4.7, -4.9, -5.4, -5.5. HRMS (ES+): calculated for $C_{57}H_{81}N_6O_{11}Si_2$ $[M+H]^+$ 1081.5496, found 1081.5466.

1-{*N*^{4}-[5'-*O*-(4,4'-Dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[*N*^{4}-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]ethane (5): Compound **4** (69 mg, 0.064 mmol) was dissolved in saturated methanolic ammonia (2 mL) and was stirred for 15 h at rt. After removal of the solvents under reduced pressure, compound **5** was obtained in 86% yield (57 mg). ^{1}H NMR ($CDCl_3$, 300 MHz) δ 7.54 (s, 1H), 7.42-7.16 (m, 10H), 6.80 (d, *J* = 9.0 Hz, 4H), 6.44 (t, *J* = 6.6 Hz, 1H), 6.35 (t, *J* = 6.6 Hz, 1H), 4.53 (m, 1H), 4.36 (m, 1H), 4.09 (m, 1H), 3.89-3.66 (m, 13H), 3.42 (dd, *J* = 3.3 Hz, *J* = 10.5 Hz, 1H), 3.34 (dd, *J* = 3.0 Hz, *J* = 10.5 Hz, 1H), 2.55 (m, 1H), 2.30 (ddd, *J* = 3.3 Hz, *J* = 6.0 Hz, *J* = 13.2 Hz, 1H), 2.19 (m, 1H), 1.97-1.88 (m, 4H), 1.57 (s, 3H), 0.90 (s, 9H), 0.87 (s, 9H), 0.08 (s, 6H), 0.05 (s, 6H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 164.1, 163.9, 158.5, 156.2, 144.5, 136.6, 136.3, 135.5, 135.5, 130.0, 128.0, 127.9, 126.9, 113.2, 103.3, 102.9, 87.4, 86.6, 85.7, 85.6, 85.4, 72.2, 71.9, 63.7, 62.8, 55.2, 42.6, 41.8, 25.9, 25.7, 18.3, 17.9, 13.4, 12.8, -4.6, -4.9, -5.4, -5.5. HRMS (ES+): calculated for $C_{55}H_{79}N_6O_{10}Si_2$ $[M+H]^+$ 1039.5391, found 1039.5421.

3',5'-Di-*O*-*tert*-butyldimethylsilyl-2'-deoxyadenosine (S3):[3] *tert*-Butyldimethylsilyl chloride (1.80 g, 11.95 mmol) was added to a solution of 2'-deoxyadenosine (1.2 g, 4.78 mmol), DMAP (88 mg, 0.72

mmol) and imidazole (1.95 g, 28.7 mmol) in anhydrous DMF (12 mL). After the reaction mixture was stirred for 48 h at room temperature, it was quenched with 5% $NaHCO_3$ and extracted with CH_2Cl_2. The combined organic layers were dried over Na_2SO_4 and concentrated. Silica gel column chromatography using 5% MeOH in CH_2Cl_2 yielded **S3** (2.23 g, 97%) as a yellow oil. ^{1}H NMR ($CDCl_3$, 500 MHz) δ 8.36 (s, 1H), 8.15 (s, 1H), 6.46 (t, *J* = 6.5 Hz, 1H), 5.70 (bs, 2H), 4.62 (m, 1H), 4.02 (m, 1H), 3.88 (dd, *J* = 4.2 Hz, *J* = 11.2 Hz, 1H), 3.78 (dd, *J* = 3.0 Hz, *J* = 11.2 Hz, 1H), 2.64 (ddd, *J* = 6.4 Hz, *J* = 6.5 Hz, *J* = 13.0 Hz, 1H), 2.44 (ddd, *J* = 4.0 Hz, *J* = 6.0 Hz, *J* = 13.0 Hz, 1H), 0.92 (s, 9H), 0.91 (s, 9H), 0.11 (s, 6H), 0.10 (s, 6H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 155.4, 152.9, 149.6, 139.0, 120.0, 87.9, 84.3, 71.8, 62.7, 41.3, 25.9, 25.7, 18.4, 18.0, -4.7, -4.8, -5.4, -5.5.

6-Chloro-9-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxy-β-D-ribofuranosyl)-9*H*-purine (7):[4] To a solution of **S3** (1.59 g, 3.23 mmol) in CH_2Cl_2 (33 mL) at 0 ºC was added dropwise trimethylsilyl chloride (837 μL, 6.60 mmol) followed by *tert*-butyl nitrite (2.18 mL, 16.5 mmol). The solution was stirred at 0 ºC for 4 h and the reaction was quenched by addition of a saturated solution of $NaHCO_3$. The aqueous layer was extracted with CH_2Cl_2. The combined organic layers were washed with water and dried over Na_2SO_4. The solvent was removed under vacuum an the residue was purified by silica gel column chromatography eluted with hexanes/EtOAc (4:1) to provide **7** (956 mg, 58%). ^{1}H NMR ($CDCl_3$, 400 MHz) δ 8.72 (s, 1H), 8.47 (s, 1H), 6.51 (t, *J* = 6.4 Hz, 1H), 4.62 (m, 1H), 4.04 (m, 1H), 3.88 (dd, *J* = 3.6 Hz, *J* = 11.2 Hz, 1H), 3.77 (dd, *J* = 2.8 Hz, *J* = 11.2 Hz, 1H), 2.63 (ddd, *J* = 6.4 Hz, *J* = 6.5 Hz, *J* = 12.8 Hz, 1H), 2.48 (ddd, *J* = 4.0 Hz, *J* = 5.6 Hz, *J* = 12.8 Hz, 1H), 0.91 (s, 9H), 0.89 (s, 9H), 0.10 (s, 6H), 0.08 (s, 6H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 151.8, 151.1, 150.9, 143.8, 132.1, 88.2, 84.9, 71.8, 62.6, 41.5, 25.9, 25.7, 18.4, 17.9, -4.7, -4.9, -5.4, -5.5. HRMS (ES+): calculated for $C_{22}H_{40}ClN_4O_3Si_2$ $[M+H]^+$ 499.2322, found 499.2325.

3',5'-Di-*O*-*tert*-butyldimethylsilyl-N^6-(2-aminoethyl)-2'-deoxyadenosine (8): A solution of 6-chloropurine nucleoside **7** (59 mg, 0.12 mmol) in pyridine (5.5 mL) was treated with ethylenediamine (339 μL, 5.1 mmol). After 15 h at rt, the solvent was evaporated and the residual pyridine was removed with toluene (x 3) followed by EtOH (x 2). The residue that was obtained was purified by silica gel chromatography eluting with CH_2Cl_2/MeOH 90:10 followed by CH_2Cl_2/MeOH 80:20 + 4% Et_3N to give compound **8** (64 mg, 99%) as a yellow foam. ^{1}H NMR ($CDCl_3$, 500 MHz) δ 8.31 (s, 1H), 8.05 (s, 1H), 6.59 (bs, 1H), 6.39 (dd, *J* = 6.8 Hz, *J* = 6.4 Hz, 1H), 4.59 (m, 1H), 4.03 (bs, 2H), 3.84 (dd, *J* = 4.8 Hz, *J* = 11.2 Hz, 1H), 3.79 (bs, 2H), 3.75 (dd, *J* = 3.2 Hz, *J* = 11.2 Hz, 1H), 3.11 (t, *J* = 5.6 Hz, 2H), 2.64 (ddd, *J* = 6.4 Hz, *J* = 6.8 Hz, *J* = 13.2 Hz, 1H), 2.39 (ddd, *J* = 3.6 Hz, *J* = 6.0 Hz, *J* = 13.2 Hz, 1H), 0.90 (s, 9H), 0.89 (s, 9H), 0.09 (s, 6H), 0.07 (s, 6H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 158.8, 152.9, 148.8, 138.4, 120.1,

87.8, 84.3, 71.9, 62.8, 42.2, 41.1, 35.7, 25.9, 18.4, 18.0, -4.7, -4.8, -5.4, -5.5. HRMS (ES+): calculated for $C_{24}H_{47}N_6O_3Si_2$ $[M+H]^+$ 523.3243, found 523.3245.

6-Chloro-9-(2'-deoxy-β-D-ribofuranosyl)-9*H*-purine (S4): A solution of **7** (515 mg, 1.03 mmol) in CH_2Cl_2/THF (1:1; 28 mL), Et_3N (1.45 mL) and $Et_3N{\cdot}3HF$ (4.2 mL) were successively added. After 18 h at rt, the solvents were removed under vacuum. Silica gel chromatography using a gradient of CH_2Cl_2/MeOH (from 95:5 to 90:10) yielded compound **S4** (253 mg, 91%) as a white foam. ^{1}H NMR (DMSO-d_6, 400 MHz) δ 8.87 (s, 1H), 8.78 (s, 1H), 6.45 (dd, *J* = 6.8 Hz, *J* = 6.4 Hz, 1H), 5.34 (d, *J* = 4.4 Hz, 1H), 4.94 (dd, *J* = 5.6 Hz, *J* = 5.2 Hz, 1H), 4.43 (m, 1H), 3.88 (m, 1H), 3.61 (ddd, *J* = 4.8 Hz, *J* = 5.2 Hz, *J* = 11.6 Hz, 1H), 3.51 (ddd, *J* = 4.8 Hz, *J* = 5.6 Hz, *J* = 11.6 Hz, 1H), 2.75 (ddd, *J* = 6.0 Hz, *J* = 6.8 Hz, *J* = 13.2 Hz, 1H), 2.36 (ddd, *J* = 3.6 Hz, *J* = 6.0 Hz, *J* = 13.2 Hz, 1H). ^{13}C NMR (DMSO-d_6, 100 MHz) δ152.3, 152.0, 149.9, 146.4, 132.1, 88.8, 84.9, 71.1, 62.0, 40.1. HRMS (ES+): calculated for $C_{10}H_{12}ClN_4O_3$ $[M+H]^+$ 271.0592, found 271.0594.

6-Chloro-9-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-β-D-ribofuranosyl]-9*H*-purine (9):[5] Compound **S4** (215 mg, 0.79 mmol) was co-evaporated with anhydrous pyridine and the residue that was obtained was dissolved in anhydrous pyridine (4 mL). The resulting solution was cooled down to 0 ºC and ^{i}Pr$_2$NEt (207 μL, 1.19 mmol) and 4,4'-dimethoxytrityl chloride (323 mg, 0.95 mmol) were successively added. The reaction mixture was allowed to stir at 0 ºC for 10 min and then, it was allowed to warm up to rt. After 90 min, the reaction was quenched with 5% $NaHCO_3$ and extracted with CH_2Cl_2. The combined organic layers were dried with $MgSO_4$ and concentrated under vacuum. The residue that was obtained was purified by silica gel chromatography eluting with CH_2Cl_2/MeOH (98:2) to give compound **9** (300 mg, 66%) as a yellow foam. ^{1}H NMR ($CDCl_3$, 400 MHz) δ 8.66 (s, 1H), 8.27 (s, 1H), 7.39-7.18 (m, 9H), 6.80 (d, *J* = 8.8 Hz, 4H), 6.49 (t, *J* = 6.4 Hz, 1H), 4.71 (m, 1H), 4.18 (m, 1H), 3.78 (s, 6H), 3.44 (dd, *J* = 4.8 Hz, *J* = 10.4 Hz, 1H), 3.38 (dd, *J* = 5.2 Hz, *J* = 10.4 Hz, 1H), 2.87 (ddd, *J* = 6.4 Hz, *J* = 7.2 Hz, *J* = 13.6 Hz, 1H), 2.59 (ddd, *J* = 4.4 Hz, *J* = 6.4 Hz, *J* = 13.6 Hz, 1H).^{13}C NMR ($CDCl_3$, 100 MHz) δ 158.6, 151.9, 151.1, 144.3, 143.7, 135.5, 135.4, 132.2, 130.0, 129.9, 129.1, 128.0, 127.9, 127.0, 113.2, 86.7, 86.4, 84.9, 72.6, 63.6, 55.2, 40.3. HRMS (ES+): calculated for $C_{31}H_{30}ClN_4O_5$ $[M+H]^+$ 573.1899, found 573.1900.

1-{N^6-[5'-*O*-(4,4'-Dimethoxytrityl)-2'-deoxyadenosylyl]-2-[N^6-(3',5'-di-*O*-*tert*-butyl-dimethylsilyl-2'-deoxyadenosylyl)]ethane (10) To a solution of 3',5'-di-*O*-*tert*-butyldimethylsilyl-N^6-(2-aminoethyl)-2'-deoxyadenosine (**8**, 45 mg, 0.09 mmol) in CH_3CN/CH_2Cl_2 (1:1, 3 mL) was added Et_3N (68 μL, 0.49 mmol), followed by 6-chloro-9-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-β-D-ribofuranosyl]-9*H*-purine (**9**, 20 mg, 0.03 mmol). The reaction mixture was allowed to stir at room temperature for 3 days. The solution

was evaporated to dryness. Silica gel column chromatography (CH_2Cl_2-MeOH from 98:2 to 95:5 followed by CH_2Cl_2-MeOH 80:20 + 4% Et_3N) yielded **10** (19 mg, 54%) as a white foam. ^{1}H NMR ($CDCl_3$, 400 MHz) δ 8.21 (s, 1H), 8.26 (s, 1H), 7.83 (s, 1H), 7.82 (s, 1H), 7.43-7.16 (m, 9H), 6.83 (d, *J* = 9.2 Hz, 2H), 6.82 (d, *J* = 9.2 Hz, 2H), 6.39 (dd, *J* = 5.6 Hz, *J* = 6.0 Hz, 1H), 6.34 (dd, *J* = 6.0 Hz, *J* = 6.4 Hz, 1H), 4.63 (m, 1H), 4.54 (m, 1H), 4.13 (m, 1H), 4.02 (m, 1H), 3.89 (dd, *J* = 4.4 Hz, *J* = 11.2 Hz, 1H), 3.82-3.70 (m, 9H), 3.51 (m, 1H), 3.43 (dd, *J* = 3.6 Hz, *J* = 10.0 Hz, 1H), 2.63-2.44 (m, 4H), 0.93 (s, 9H), 0.89 (s, 9H), 0.11 (s, 3H), 0.10 (s, 3H), 0.09 (s, 3H), 0.08 (s, 3H). ^{13}C NMR ($CDCl_3$, 100 MHz) δ 158.5, 154.3, 154.2, 153.0, 152.9, 144.2, 138.0, 137.8, 135.6, 135.5, 130.0, 128.3, 127.9, 127.0, 119.6, 113.2, 87.6, 86.7, 85.6, 84.9, 84.8, 84.4, 77.2, 71.5, 62.9, 62.6, 55.1, 41.4, 38.8, 29.7, 25.9, 25.9, 25.8, 18.4, 18.0, -4.6, -4.8, -5.4, -5.5. HRMS (ES+): calculated for $C_{55}H_{75}N_{10}O_8Si_2$ $[M+H]^+$ 1059.5302, found 1059.5284.

Solid support functionalization: The polymer support was functionalized with 1-{N^4-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl]}-2-[N^4-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]ethane (**5**) or 1-{N^6-[5'-*O*-(4,4'-dimethoxytrityl)-2'-deoxyadenosylyl]-2-[N^6-(3',5'-di-*O*-*tert*-butyldimethylsilyl-2'-deoxyadenosylyl)]ethane (**10**) as described for natural nucleosides.[6]

Step I. Preparation of 1-{N^4-[5'-O-(4,4'-dimethoxytrityl)-2'-deoxy-5-methylcytidylyl-3'-O-succinyl]}-2-[N^4-(3',5'-di-O-tert-butyldimethylsilyl-2'-deoxy-5-methylcytidylyl)]ethane and 1-{N^6-[5'-O-(4,4'-dimethoxytrityl)-2'-deoxyadenosylyl-3'-O-succinyl]-2-[N^6-(3',5'-di-O-tert-butyldimethylsilyl-2'-deoxyadenosylyl)]ethane: Succinic anhydride (1.3 mmol), $^iPr_2NEt$ (1.4 mmol), and DMAP (0.1 equiv) were added to a solution of 5'-*O*-DMT-3',5'-di-*O*-TBDMS-protected dimeric nucleoside **5** or **10** (1 mmol) in CH_2Cl_2 (0.2 M). The resulting solution was stirred for 24 h at rt and then washed with sodium dihydrogen phosphate (1%). The aqueous layer was extracted with CH_2Cl_2 and the organic layer was dried with $MgSO_4$ and concentrated.

Step II: 2,2'-Dithio-bis-(5-nitropyridine) (0.1 mmol), dissolved in an acetonitrile/dichloroethane mixture (1:3, 400 µL), was mixed with a solution of 5'-*O*-DMT-3',5'-di-*O*-TBDMS-protected 3'-*O*-succinyl dimeric nucleoside (0.1 mmol) and DMAP (0.1 mmol) in CH_3CN (500 µL). The resulting clear solution was added at rt to a solution of triphenylphosphine (0.1 mmol) in CH_3CN (200 µL). The mixture was vortexed for few seconds and then added to a vial containing CPG (500 Å, 500 mg) and allowed to react for 30 min at rt. MeOH (500 µL) was then added and the support was recovered on a sintered glass funnel, followed by washings with MeOH (2 x 5 mL) and Et_2O (2 x 5 mL). The support was air-dried and then placed under high vacuum. The support was subjected to capping by the standard protocol.[7] The dimeric nucleoside loading on the derivatized support was determined by the acid treatment method.[7]

RNA synthesis

Unmodified and PS-modified RNAs were synthesized on the 0.2 µmol scale and 3'-B^C/B^A-modified RNA oligonucleotides on the 1 µmol scale, respectively, with an Applied Biosystems 394 synthesizer. PS-modified oligonucleotides were synthesized following previously described methods.[8] 2'-*O*-TBDMS-5'-*O*-DMT-protected phosphoramidites (A^{Bz}, G^{dmf}, C^{Ac} and U) were used. Acetonitrile (synthesis grade) and the 2'-*O*-TBDMS-protected phosphoramidite monomers of A, C, G, and U were from commercial suppliers. For the synthesis of unmodified or 5'-B^C-modified RNA strands, commercially available succinyl polystyrene functionalized with 5'-*O*-DMT-thymidine was used as the solid support (0.2 µmol scale). For the synthesis of 3'-B^C- and B^A-modified RNA strands, CPG functionalized with B^C and B^A

units were used as the solid supports. The following solutions were used: 0.4 M 1*H*-tetrazole in acetonitrile (activation); 3% trichloroacetic acid in dichloromethane (detritylation), acetic anhydride/pyridine/tetrahydrofuran (1:1:8) (capping A), 10% *N*-methylimidazole in tetrahydrofuran (capping B), 0.02 M iodine in tetrahydrofuran/pyridine/water (7:2:1) (P (III) to P(V) oxidation), sulfurizing reagent (0.49 M tetraethylthiuram disulfide, TETD, in CH_3CN). The coupling time was 15 min. The coupling yields of natural and modified phosphoramidites were around 95%. Incorporation of the dimeric nucleoside modification did not have a negative effect in the yield. All oligonucleotides were synthesized in DMT-ON mode.

DNA synthesis

Unmodified and 3'-B^C/B^A-modified DNA oligonucleotides were synthesized on the 0.2 µmol and 1 µmol scale, respectively, with an Applied Biosystems 394 synthesizer and use of 5'-*O*-DMT-protected phosphoramidites of natural 2'-deoxynucleotides (dA^{Bz}, dG^{dmf}, dC^{Bz} and T). DNA synthesis was performed under the same conditions (activation, detritylation, capping and oxidation) as those used for RNA synthesis except for the coupling time, which was 1 minute.

Deprotection and purification of unmodified and modified RNA and DNA oligonucleotides

After the solid-phase synthesis, the solid support was transferred to a screw-cap vial and incubated at 55 ºC for 1 h with 1.5 mL of NH_3 solution (33%) and 0.5 mL of ethanol. The vial was then cooled on ice and the supernatant was transferred into a 2 mL eppendorf tube. The solid support and vial were rinsed with 50% ethanol (2 x 0.25 mL). The combined solutions were evaporated to dryness using an evaporating centrifuge. The residue that was obtained was dissolved in 1 M TBAF in THF (unmodified and modified RNA oligonucleotides: 85 µL per 0.2 µmol resin, 330 µL per 1 µmol resin; modified DNA oligonucleotides: 85 µL per 1 µmol resin) and incubated at room temperature for 15 h. Then, 1 M triethylammonium acetate (TEEA) and water were added to the solution (0.2 µmol RNA synthesis: 85 µL TEEA and 330 µL water; 1 µmol RNA synthesis: 330 µL TEEA and 330 µL water, respectively; 1 µmol modified-DNA synthesis: 85 µL TEEA and 830 µL water, respectively). The oligonucleotides were desalted on NAP-5 or NAP-10 columns (0.2 µmol synthesis and 1 µmol synthesis, respectively) using water as the eluent and evaporated to dryness. The oligonucleotides were purified by HPLC (DMT-ON). Column: Nucleosil 120-10 C_{18} (250 x 4 mm); 20 min linear gradient from 15% to 80% B and 5 min 80% B, flow rate 3 mL/min; solution A was 5% ACN in 0.1 M aqueous TEAA and B 70% ACN in 0.1 M

aqueous TEAA. The pure fractions were combined and evaporated to dryness. The residue that was obtained was treated with 1 mL of 80% AcOH solution and incubated at room temperature for 30 min. The deprotected oligonucleotide was desalted on a NAP-10 column using water as the eluent. All oligonucleotides were quantified by absorption at 260 nm and confirmed by MALDI and ESI mass spectrometry. SiRNAs were prepared by annealing equimolar quantities of complementary oligonucleotides in siRNA buffer (100 mM KOAc, 30 mM HEPES-KOH, 2 mM $MgCl_2$, pH 7.4) by slowly cooling from 96 ºC to r.t.

Table S1. Mass spectrometry analysis of synthesized oligonucleotides

ON	Sequence	MW calcd.	MW found
12	5'- CGTTTCCTTTGTTCTGGA -3'	5468.6 (+Na)	5461.1 (+Na) [a]
13	5'- CGTTTCCTTTGTTCTGGA**B**^{C} -3'	6037.9 (+Na)	6033.3 (+Na) [b]
14	5'- CGTTTCCTTTGTTCTGGA**B**^{A} -3'	6035.1	6033.1 [b]
15	3'- GCAAAGGAAACAAGACCT-5'	5517.6	5520.1 [a]
16	3'-TTAAAAAGAGGAAGAAGUCUA-5'	6790.0	6787.6 [a]
17	3'-**B**^{C}AAAAAGAGGAAGAAGUCUA-5'	6775.0 (+Na)	6774.1 (+Na) [a]
18	3'-**B**^{A}AAAAAGAGGAAGAAGUCUA-5'	6795.4 (+Na)	6792.7 (+Na) [a]
19	5'-UUUUUCUCCUUCUUCAGAUTT-3'	6423.0	6426.4 [a]
20	5'-UUUUUCUCCUUCUUCAGAU**B**^{C}-3'	6407.1 (+Na)	6406.1 (+Na) [a]
21	5'-UUUUUCUCCUUCUUCAGAU**B**^{A}-3'	6427.3 (+Na)	6424.9 (+Na) [a]
22	3'-TTAGGAAAGAAAGAAAGCUAU	6812.5 (+Na)	6812.1 (+Na) [a]
23	5'-UCCUUUCUUUCUUUCGAUATT	6445.8 (+Na)	6441.1 (+Na) [a]
24	3'-TTGAAGUAGUGAUAGAGGGCC	6814.4 (+Na)	6808.6 (+Na) [a]
25	3'-**B**^{C}GAAGUAGUGAUAGAGGGCC	6777.2 (+Na)	6771.0 (+Na) [a]
26	5'-CUUCAUCACUAUCUCCCGGTT	6505.2 (+Na)	6500.2 (+Na) [a]
27	3'-TTUGCACUGUGCAAGCCUCUU	6585.4 (+Na)	6579.7 (+Na) [a]
28	5'-ACGUGACACGUUCGGAGAATT	6734.2 (+Na)	6728.3 (+Na) [a]
29	5'- CGTTTCCTTTGTTCTGGAT$_{S}$T -3'	6090.8 (+Na)	6087.2 (+Na) [b]
30	5'- C$_{S}$G$_{S}$T$_{S}$T$_{S}$T$_{S}$C$_{S}$C$_{S}$T$_{S}$T$_{S}$T$_{S}$G$_{S}$T$_{S}$T$_{S}$C$_{S}$T$_{S}$G$_{S}$G$_{S}$A -3'	5741.5	5736.1 [b]
31	3'-T$_{S}$TAAAAAGAGGAAGAAGUCUA-5'	6846.2 (+Na)	6849.4 (+Na) [b]
32	3'-T$_{S}$T$_{S}$A$_{S}$A$_{S}$AAAGAGGAAGAAGUCUA-5'	6917.3 (+2 Na)	6920.3 (+2 Na) [b]
33	3'-T$_{S}$T$_{S}$A$_{S}$A$_{S}$A$_{S}$A$_{S}$A$_{S}$G$_{S}$A$_{S}$G$_{S}$G$_{S}$A$_{S}$A$_{S}$G$_{S}$A$_{S}$A$_{S}$G$_{S}$U$_{S}$C$_{S}$U$_{S}$A-5'	7160.5	7166.0 [b]

[a] Oligonucleotides **12** and **15-28** were confirmed by MALDI-TOF mass spectrometry.

[b] Oligonucleotides **13**, **14** and **29-33** were confirmed by ESI mass spectrometry.

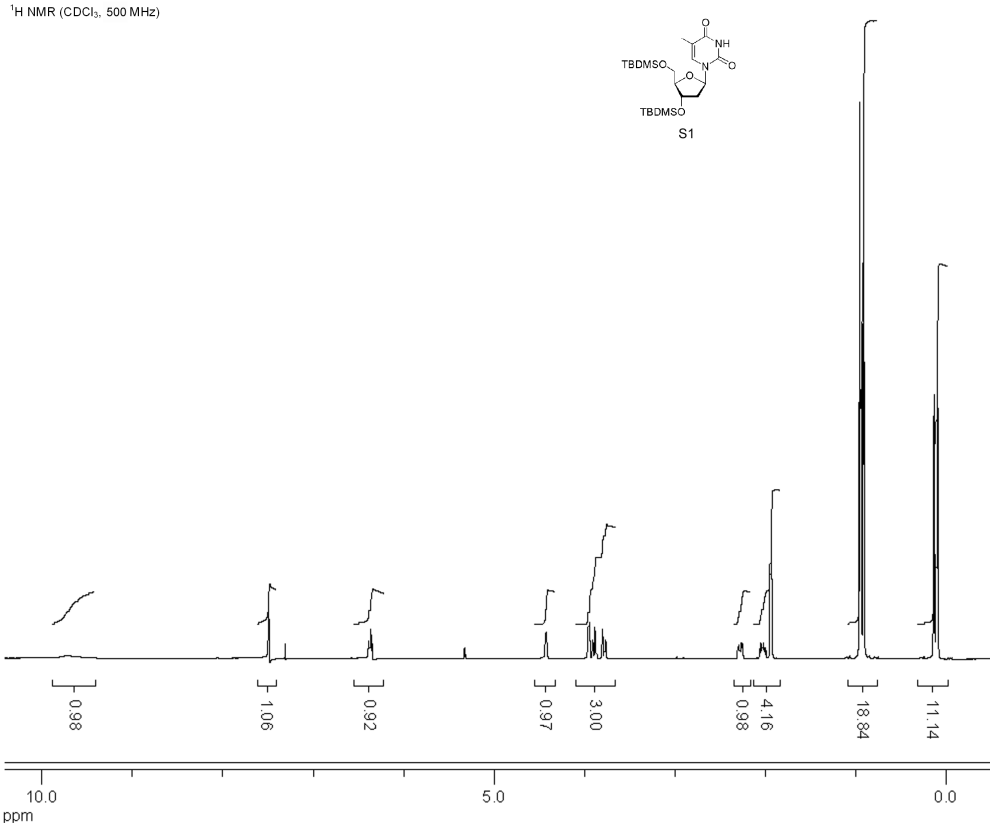


Figure S1. ^{1}H NMR spectrum of compound **S1**

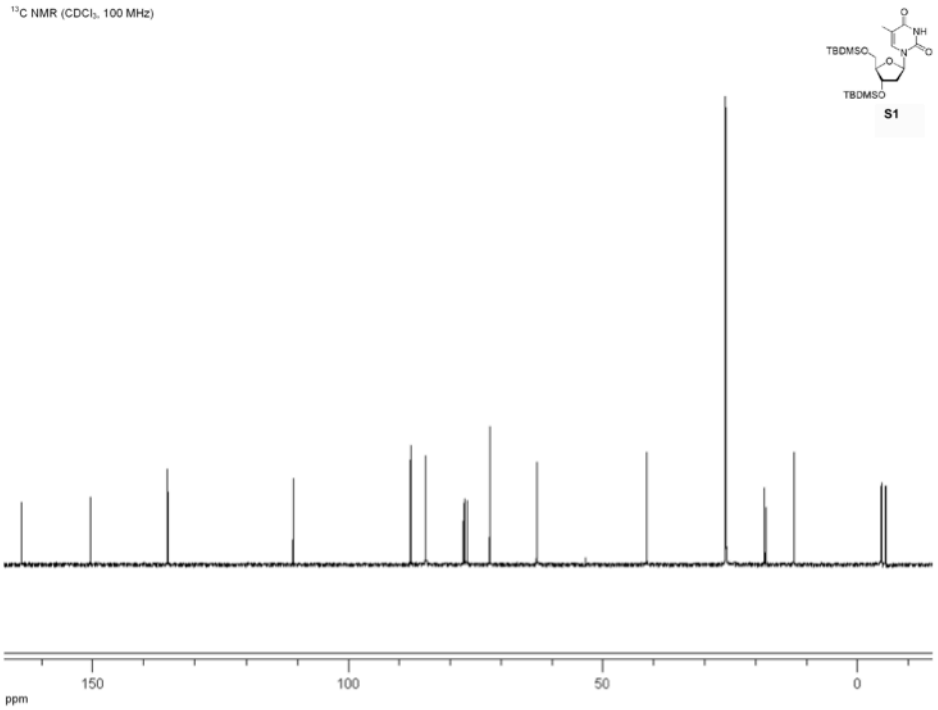


Figure S2. ^{13}C NMR spectrum of compound **S1**

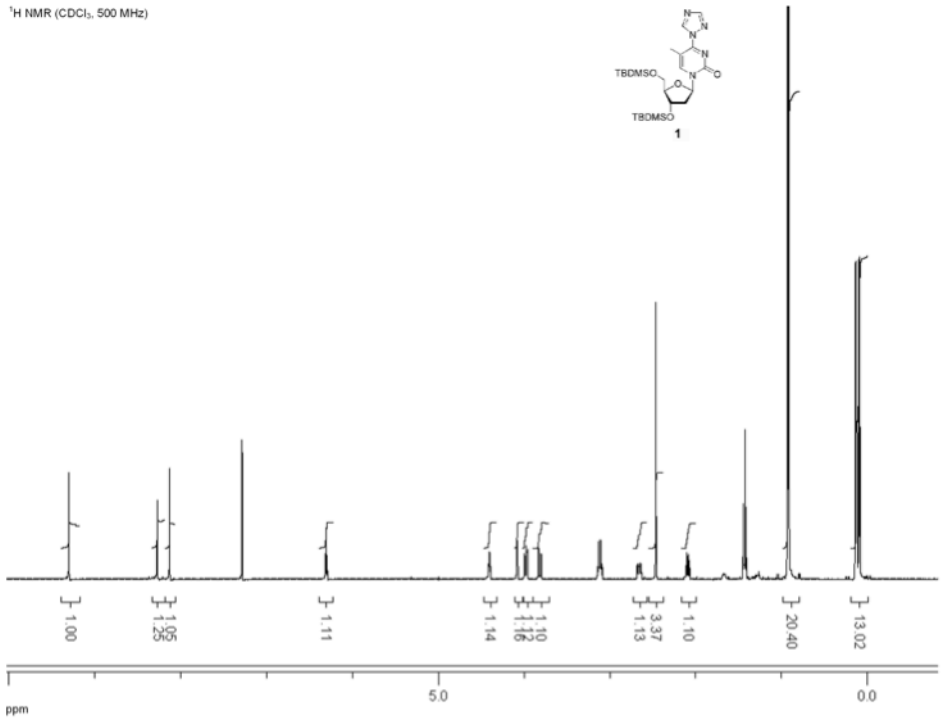


Figure S3. ^{1}H NMR spectrum of compound **1**

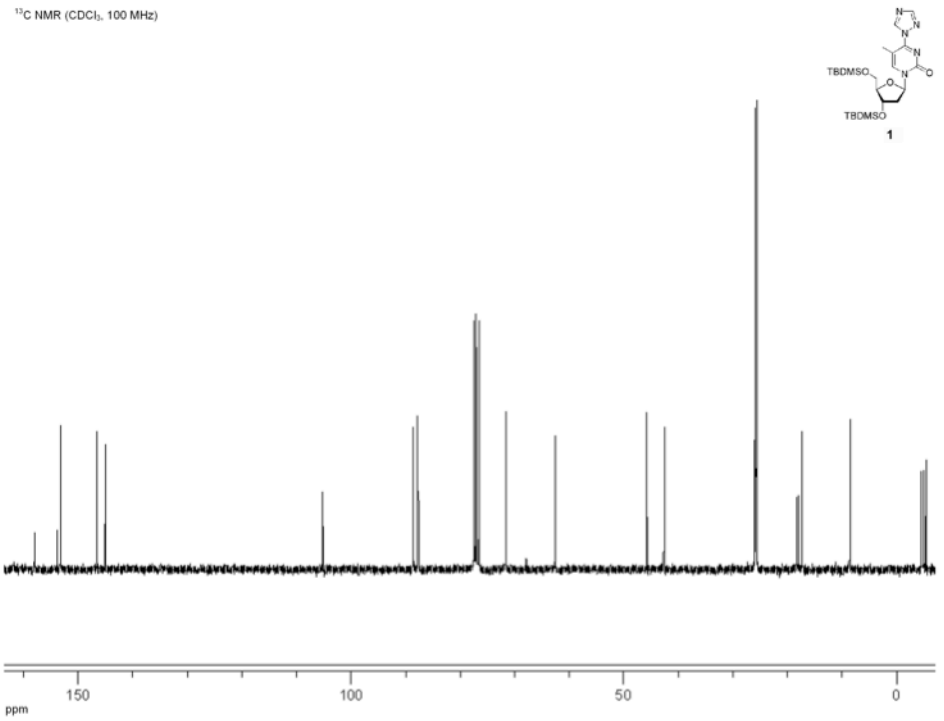


Figure S4. ^{13}C NMR spectrum of compound **1**

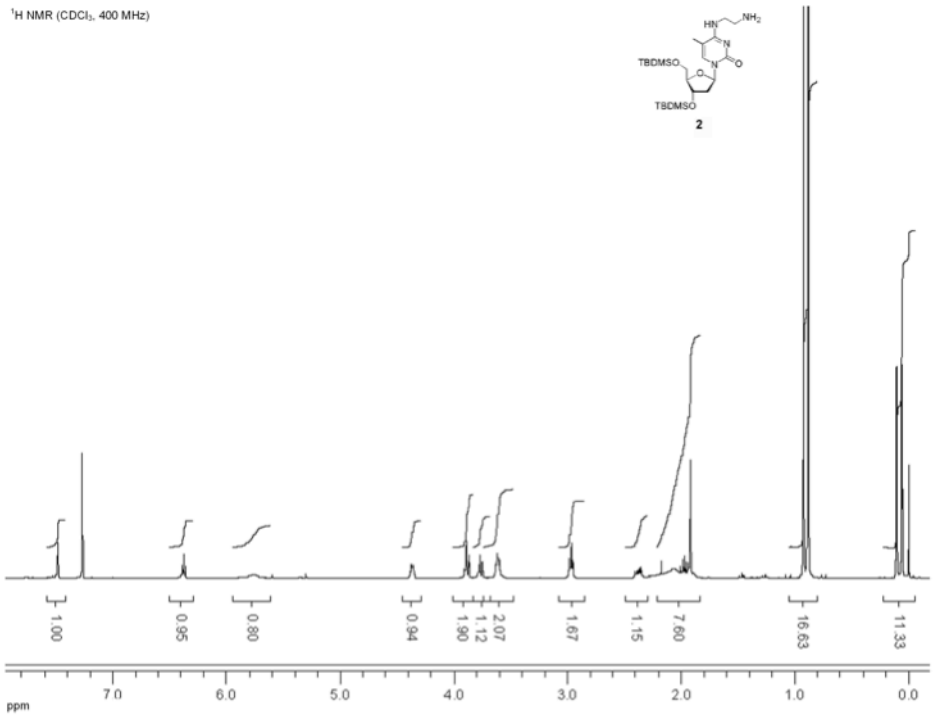


Figure S5. ^{1}H NMR spectrum of compound **2**

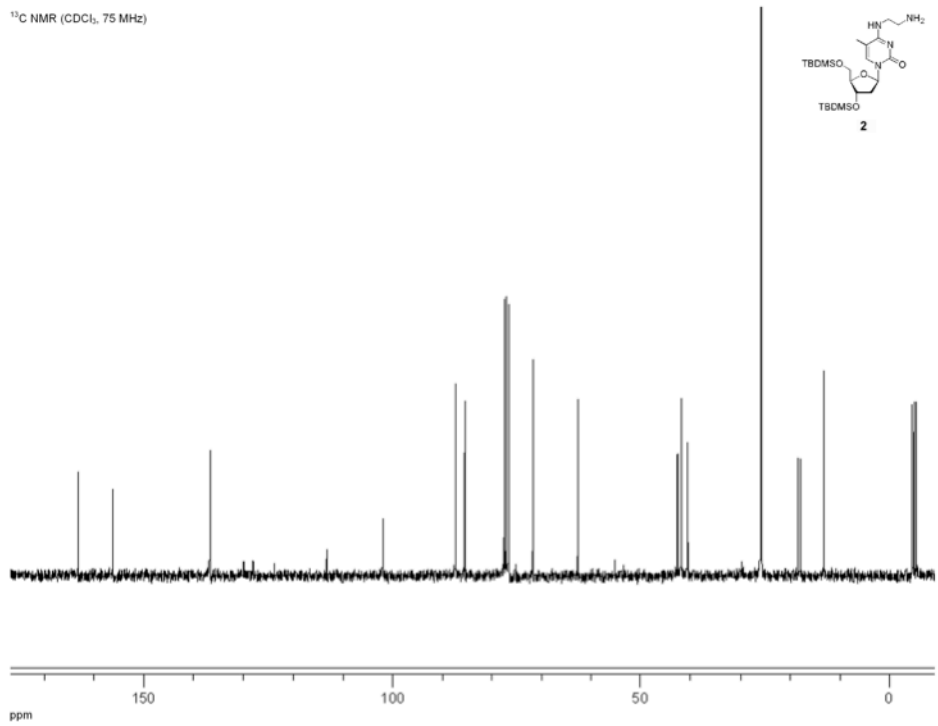


Figure S6. ^{13}C NMR spectrum of compound **2**

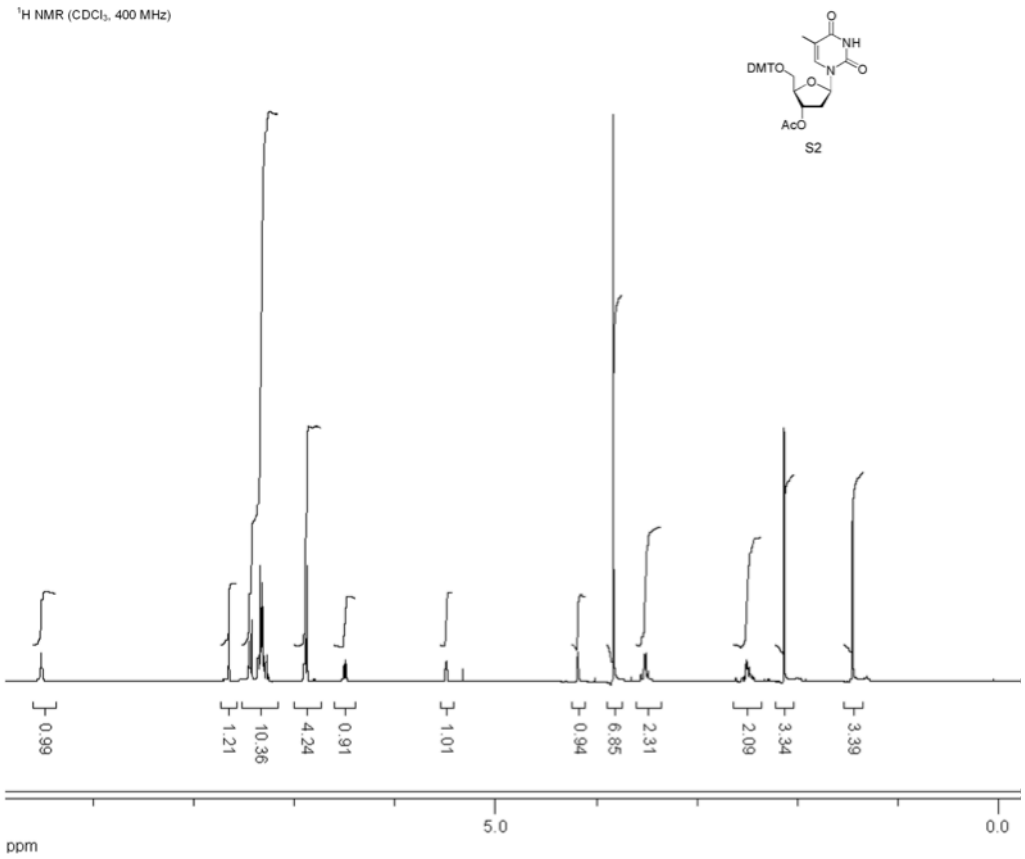


Figure S7. ^{1}H NMR spectrum of compound **S2**

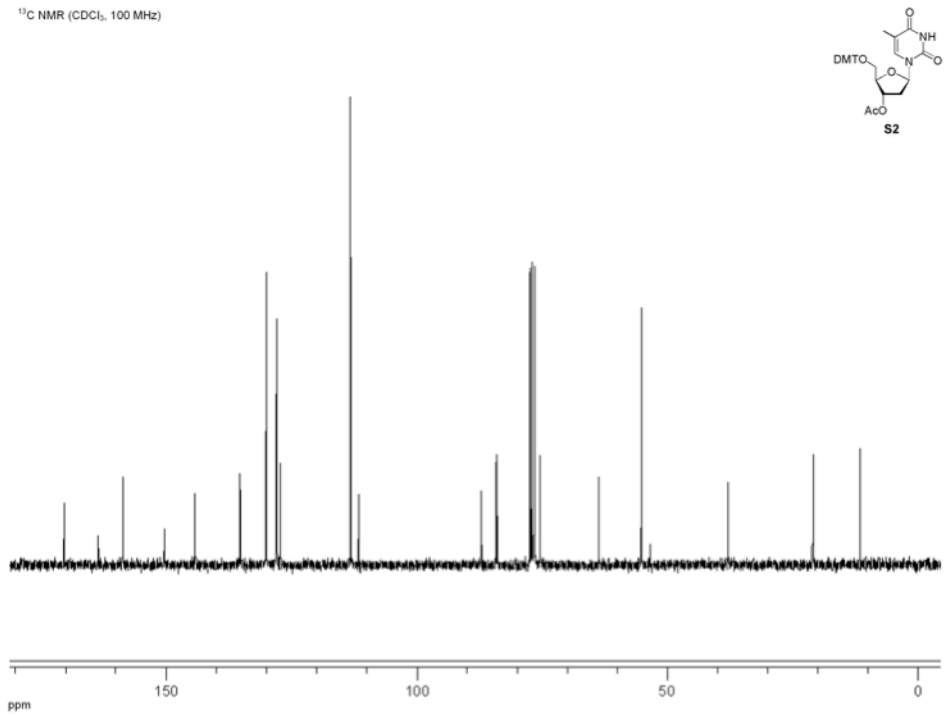


Figure S8. ^{13}C NMR spectrum of compound **S2**

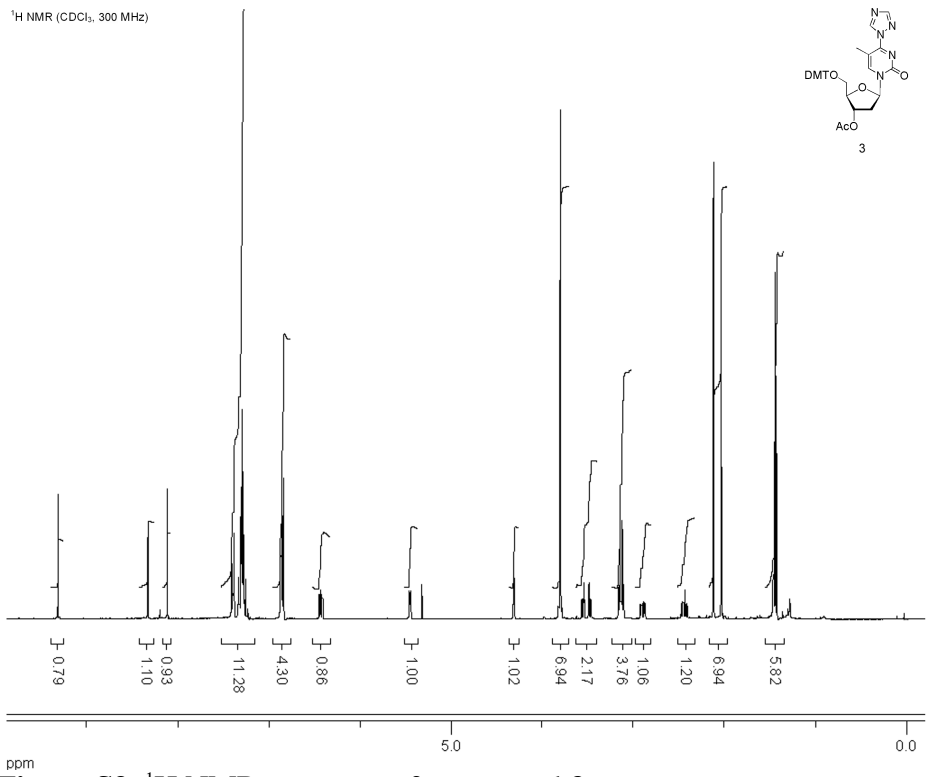


Figure S9. ^{1}H NMR spectrum of compound **3**

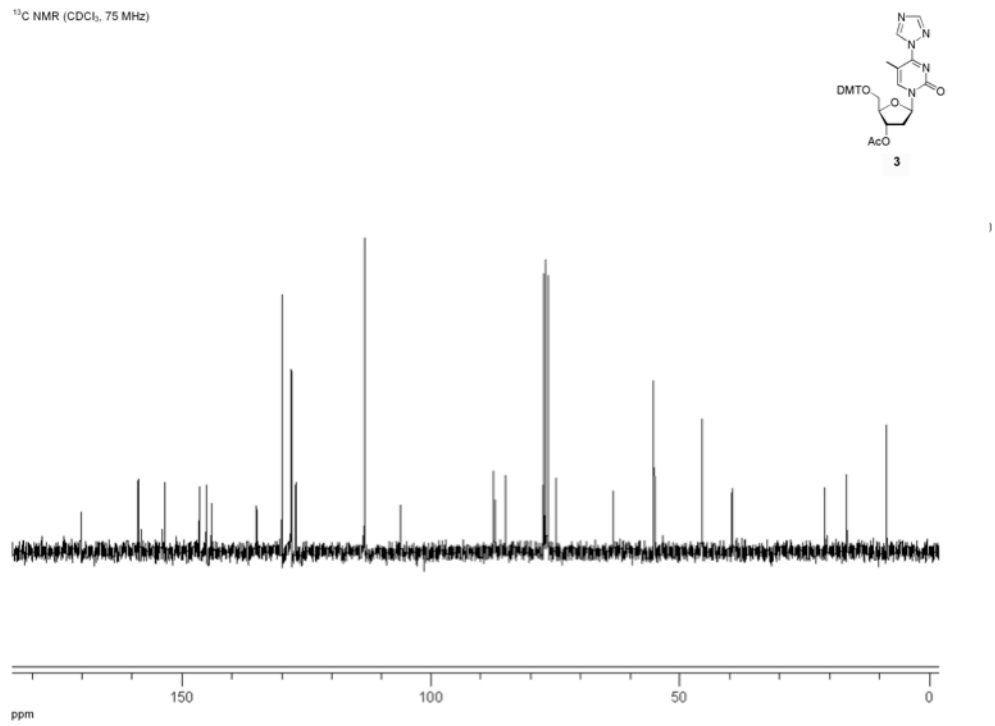


Figure S10. ^{13}C NMR spectrum of compound **3**

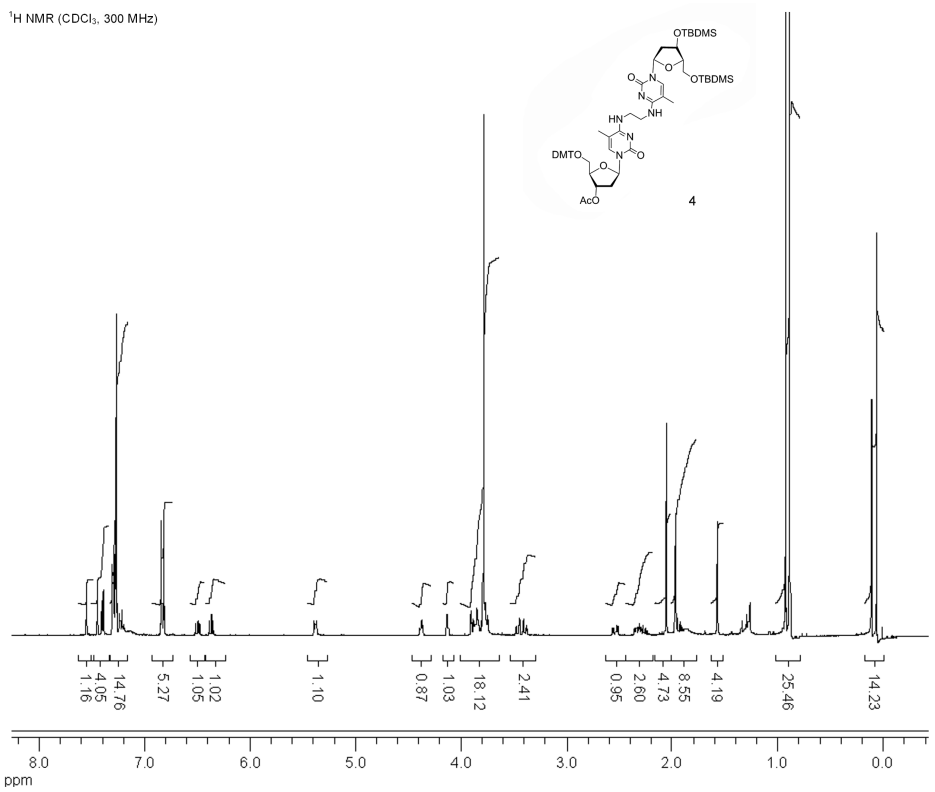


Figure S11. [1]H NMR spectrum of compound **4**

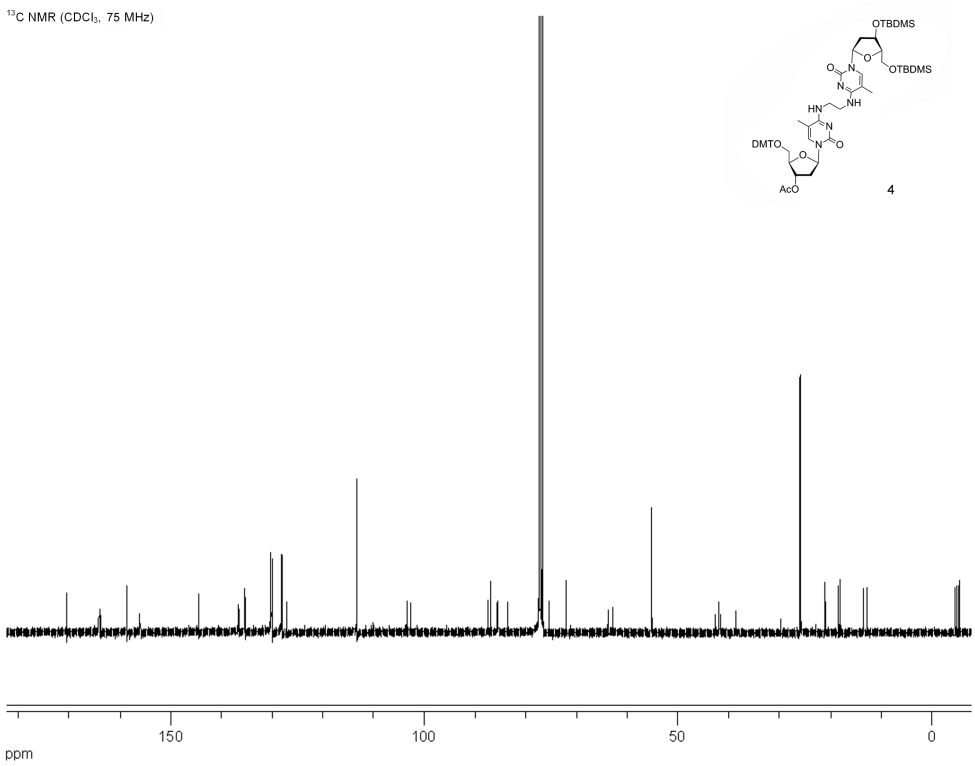


Figure S12. [13]C NMR spectrum of compound **4**

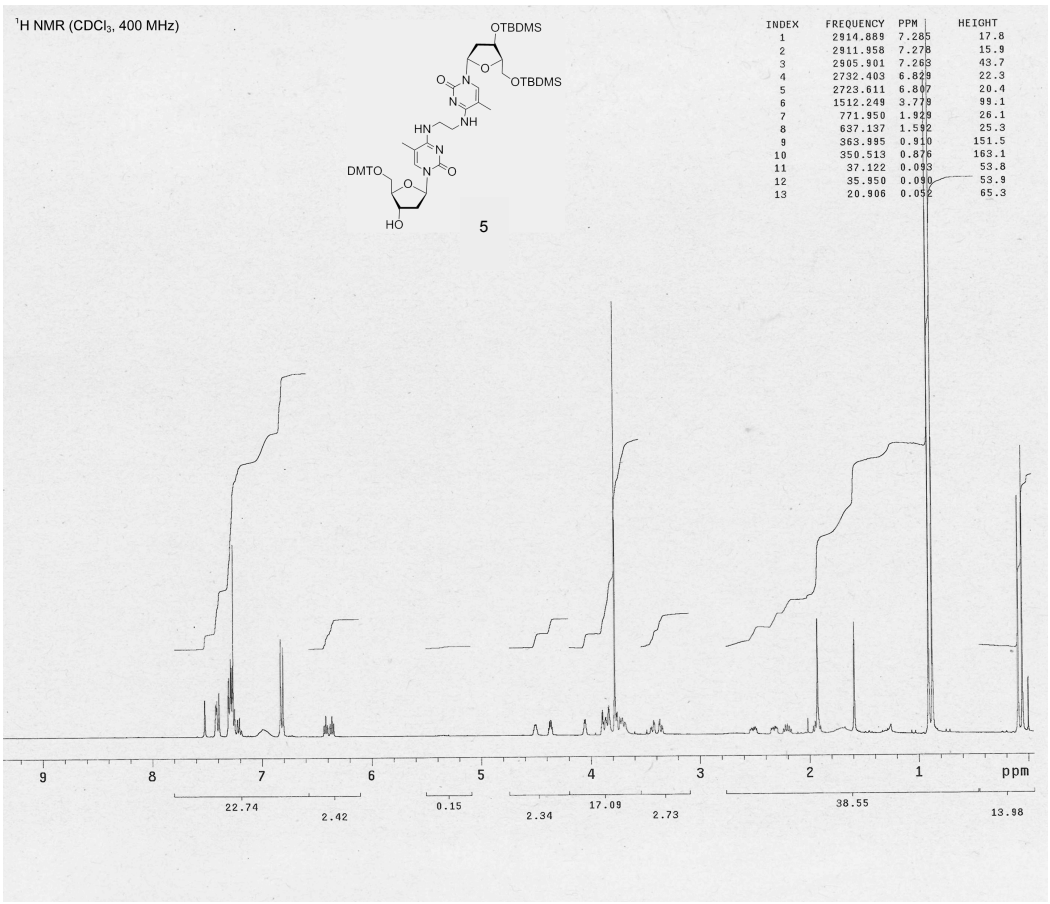


Figure S13. ^{1}H NMR spectrum of compound **5**

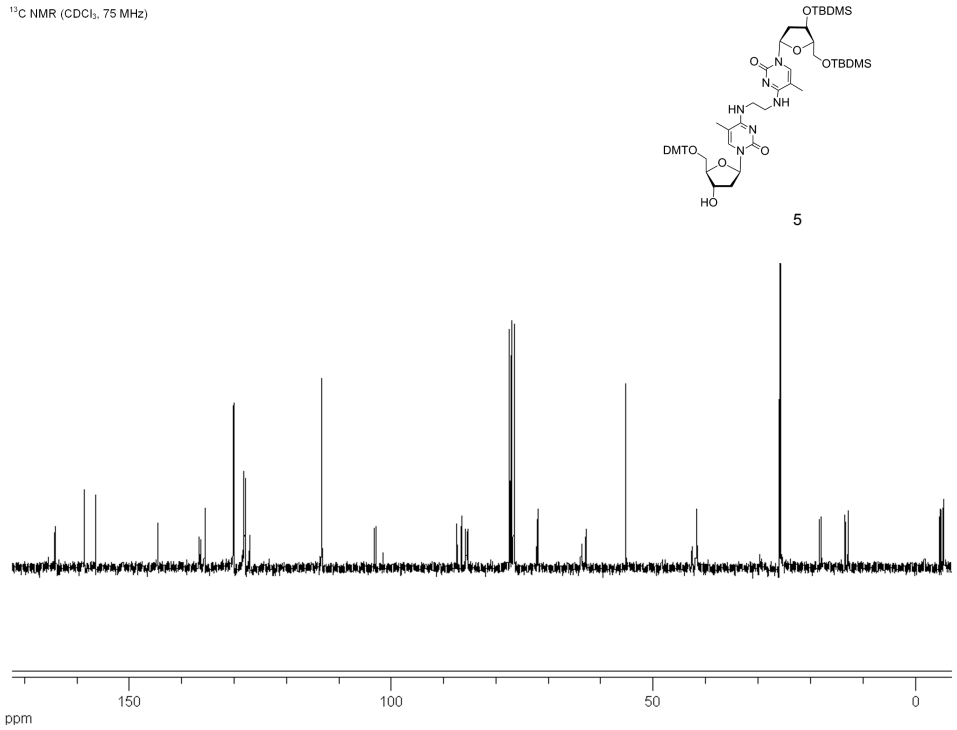


Figure S14. ^{13}C NMR spectrum of compound **5**

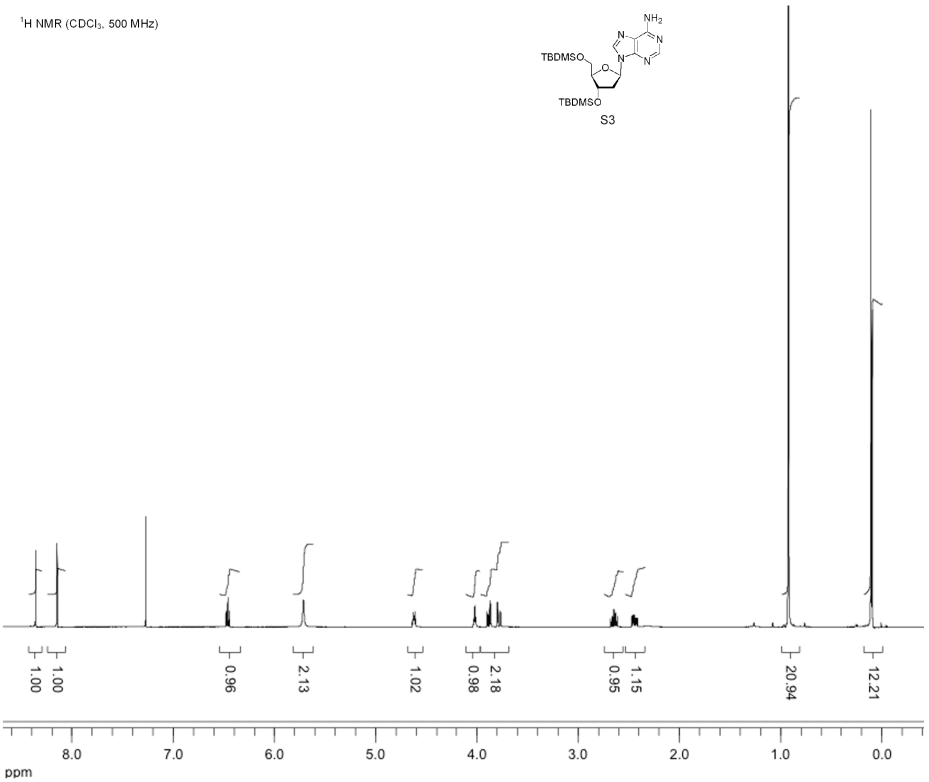


Figure S15. ^{1}H NMR spectrum of compound **S3**

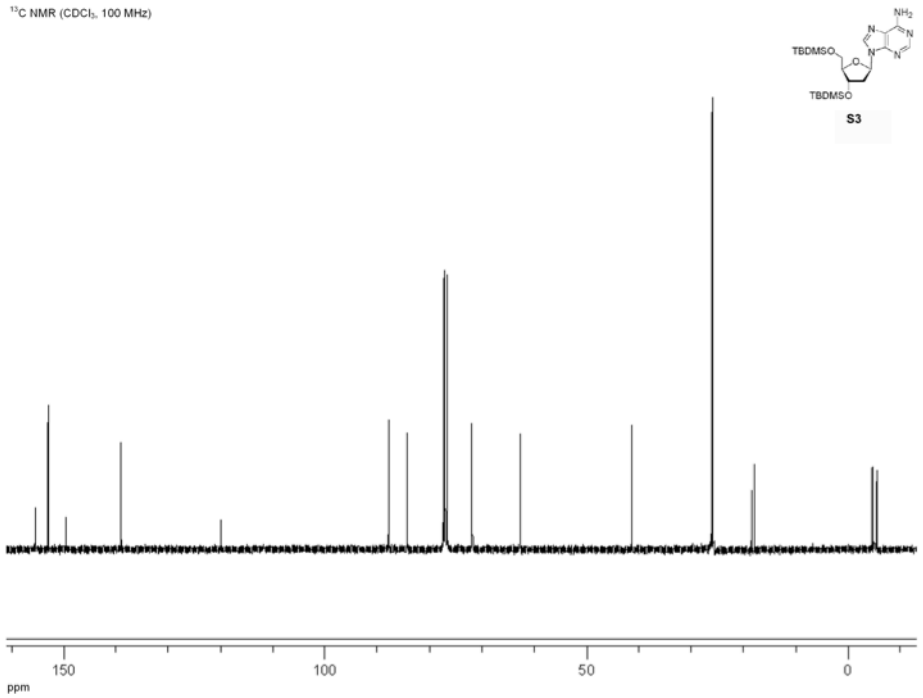


Figure S16. ^{13}C NMR spectrum of compound **S3**

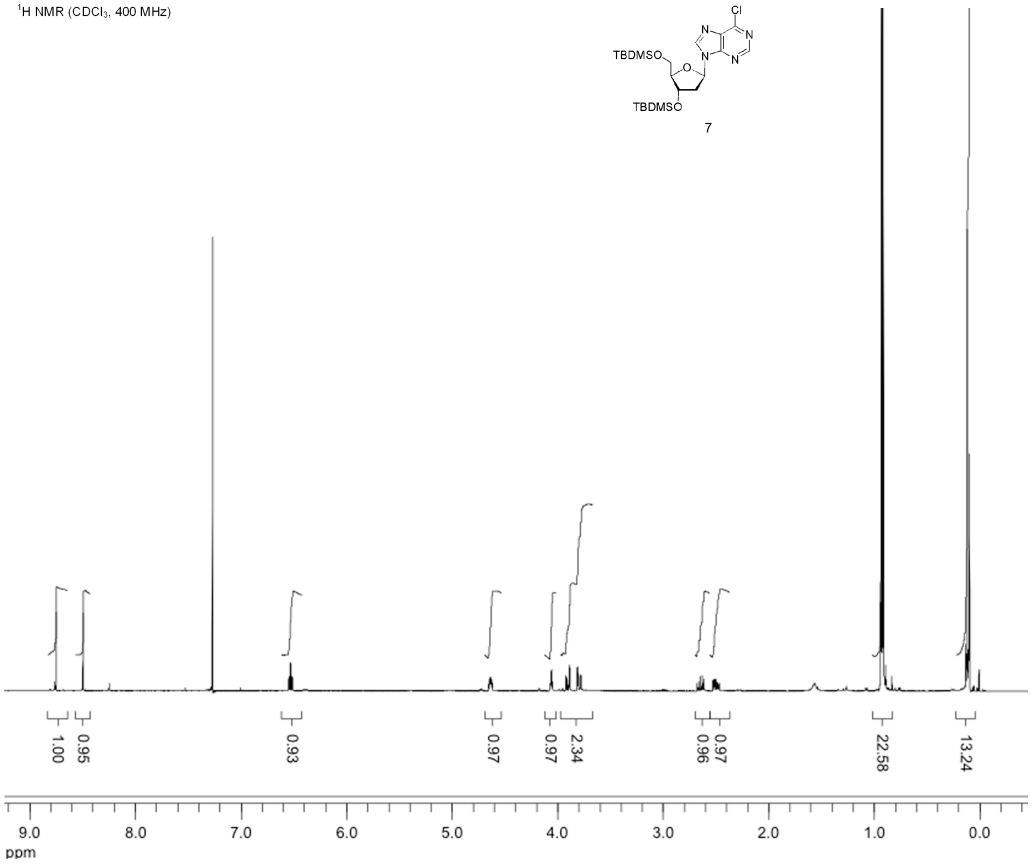


Figure S17. ^{1}H NMR spectrum of compound **7**

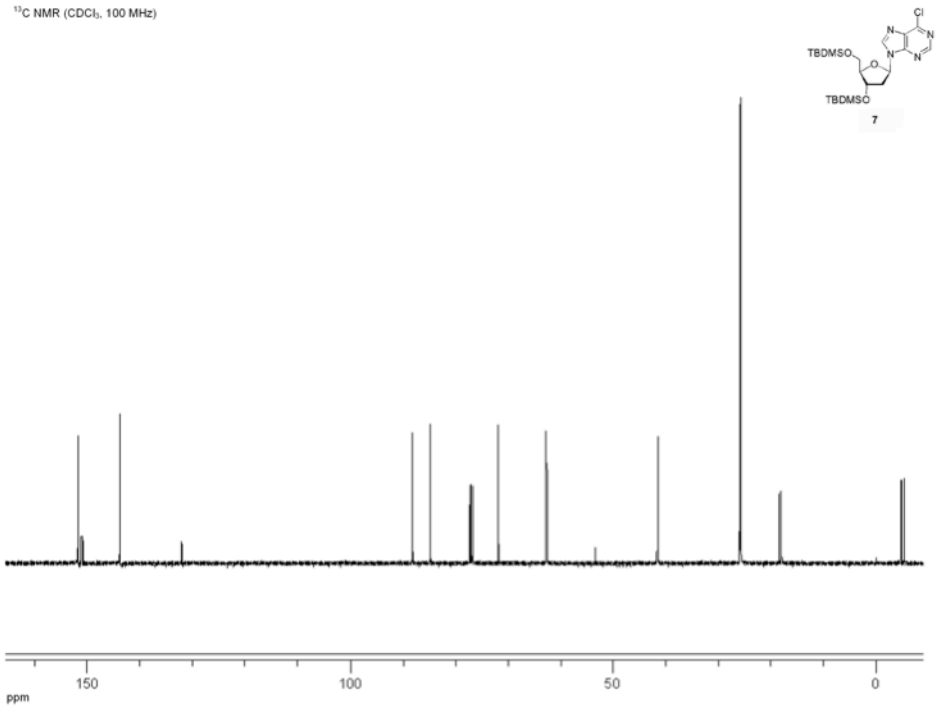


Figure S18. ^{13}C NMR spectrum of compound **7**

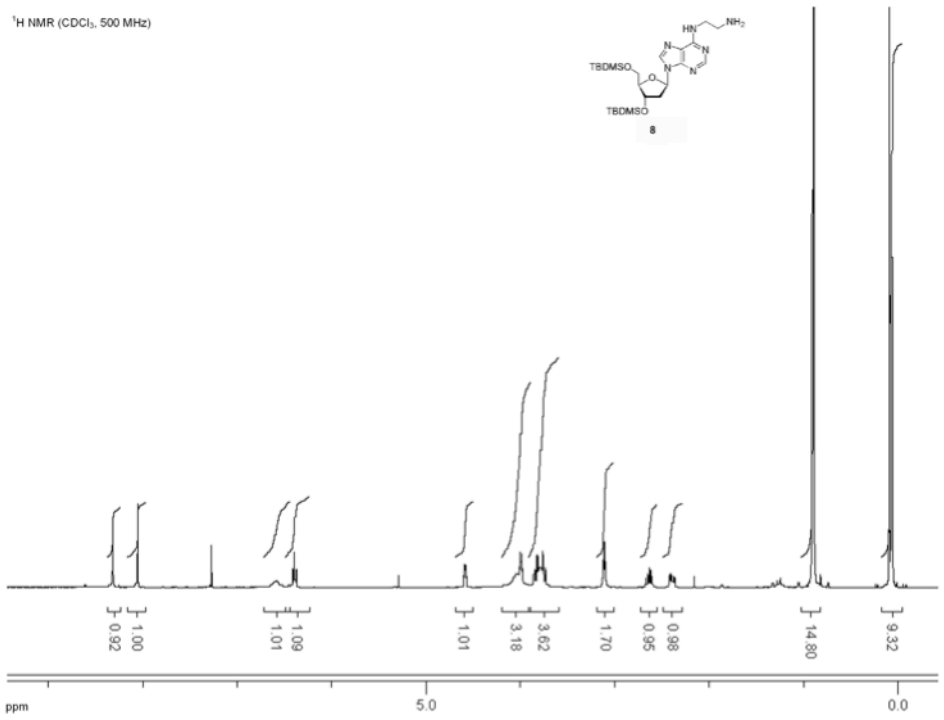


Figure S19. ^{1}H NMR spectrum of compound **8**

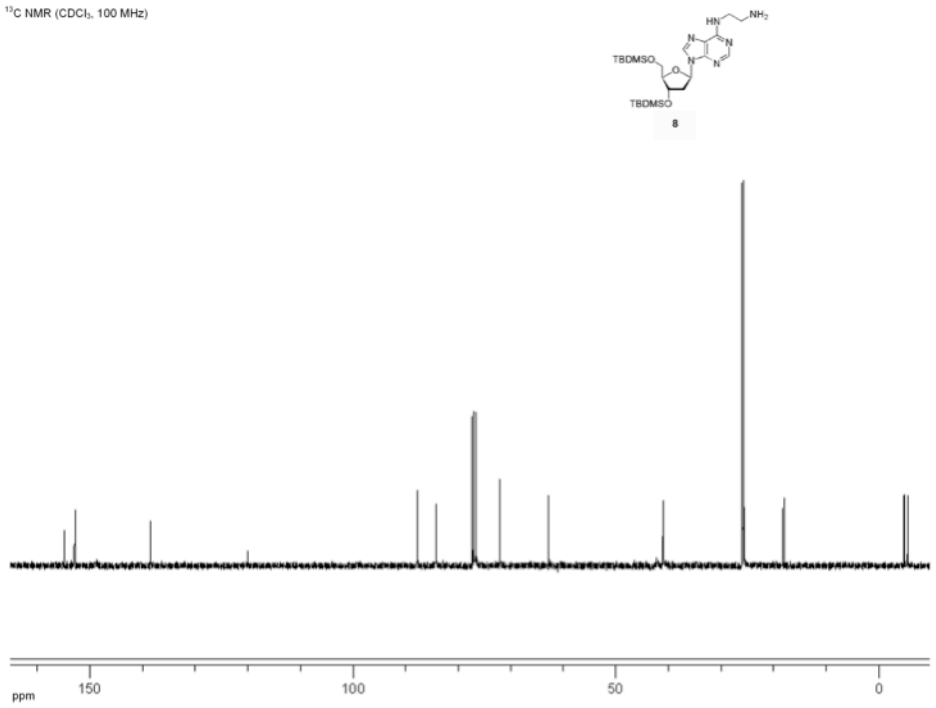


Figure S20. ^{13}C NMR spectrum of compound **8**

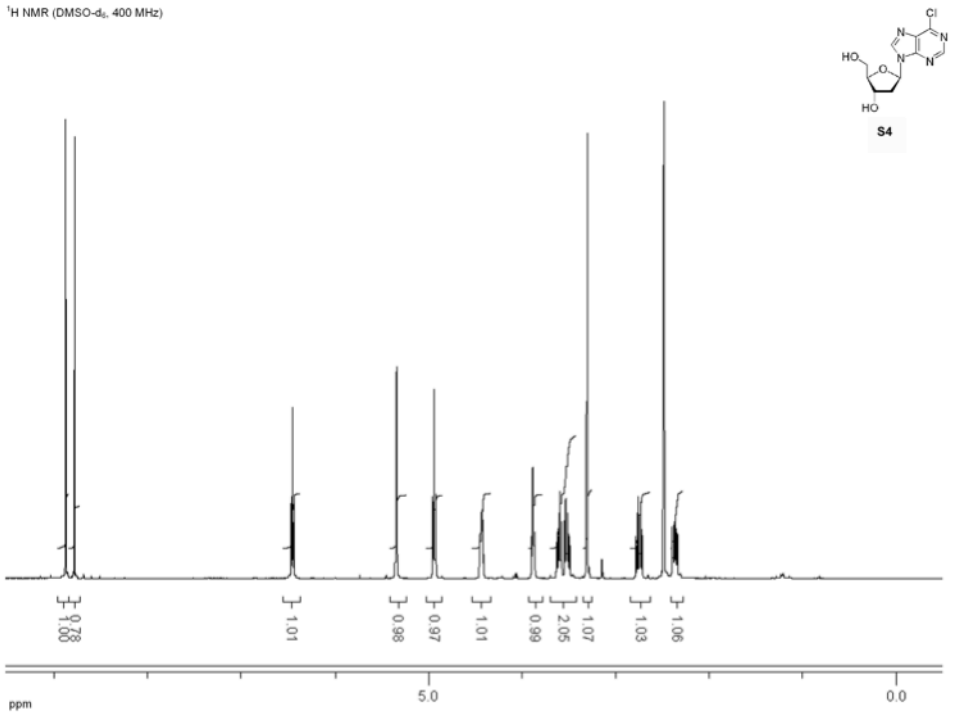


Figure S21. ^{1}H NMR spectrum of compound **S4**

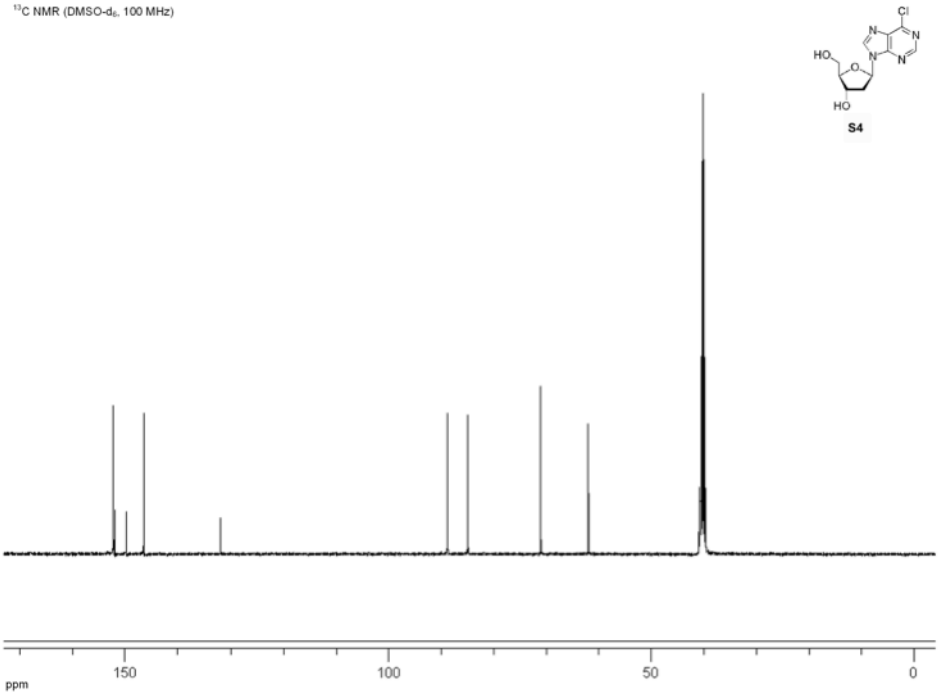


Figure S22. ^{13}C NMR spectrum of compound **S4**

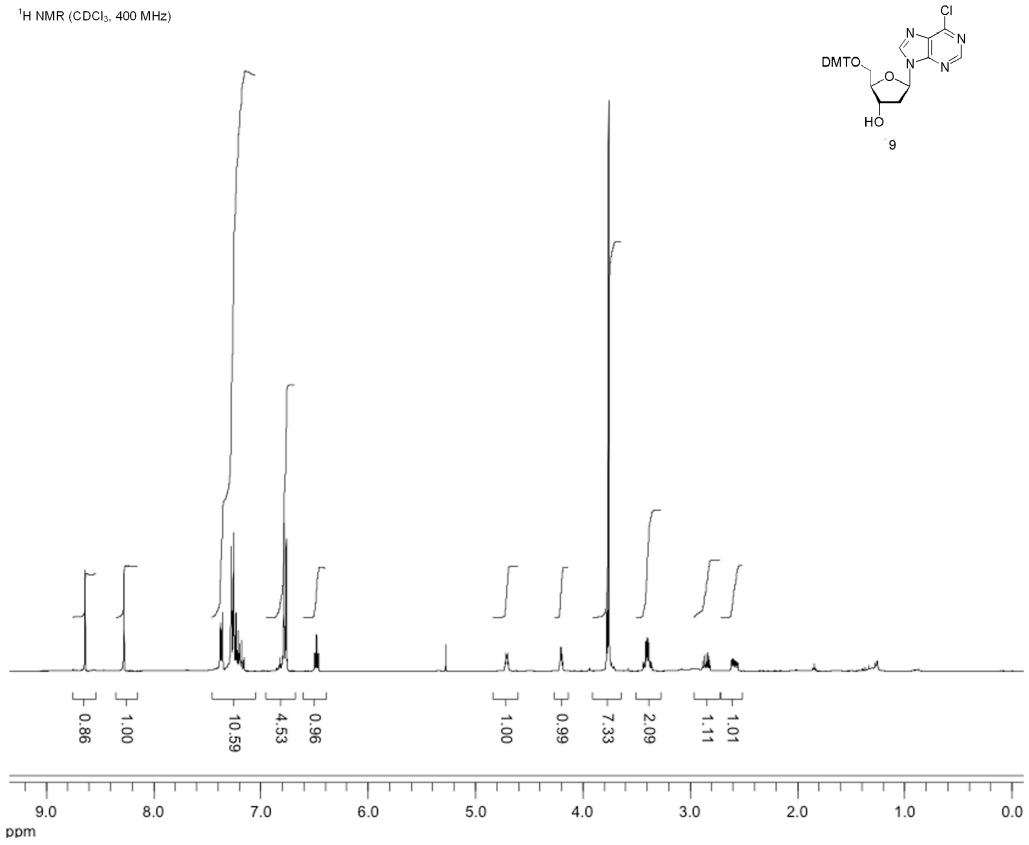


Figure S23. ^{1}H NMR spectrum of compound **9**

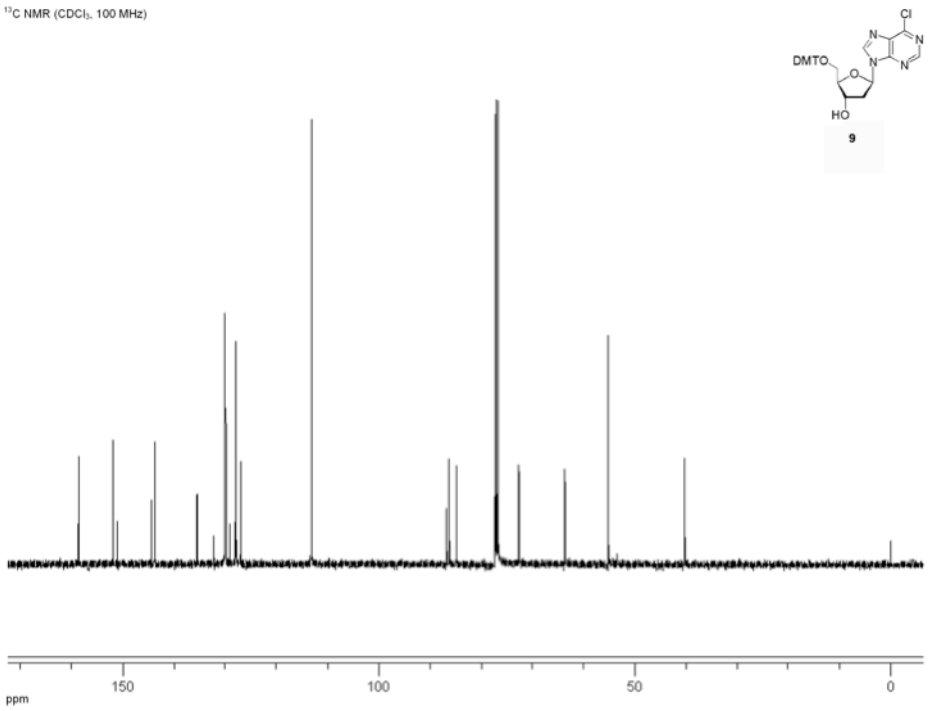


Figure S24. ^{13}C NMR spectrum of compound **9**

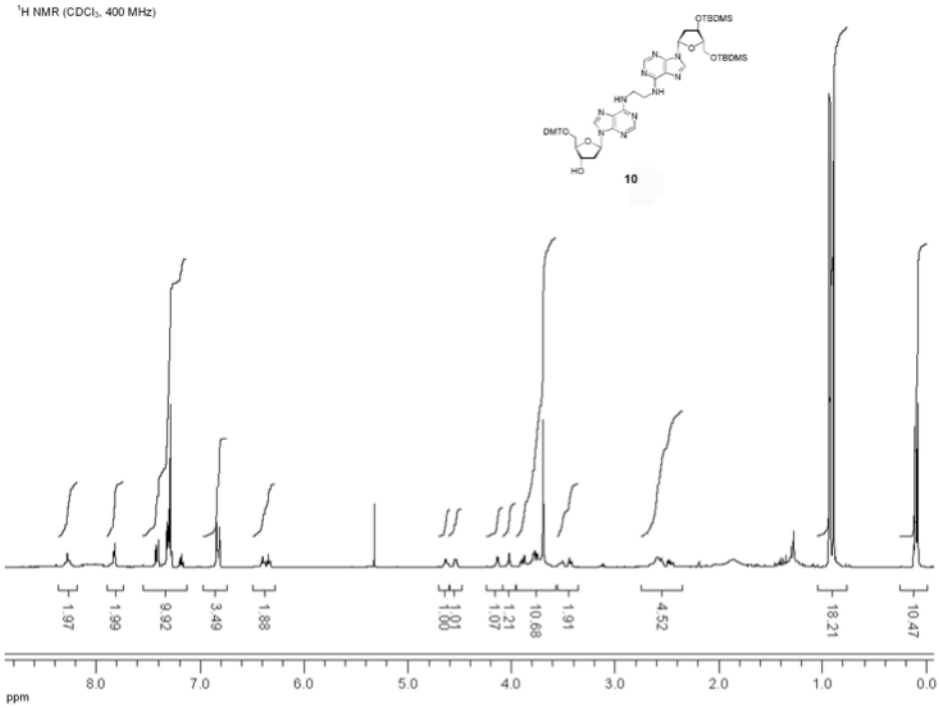


Figure S25. ^{1}H NMR spectrum of compound **10**

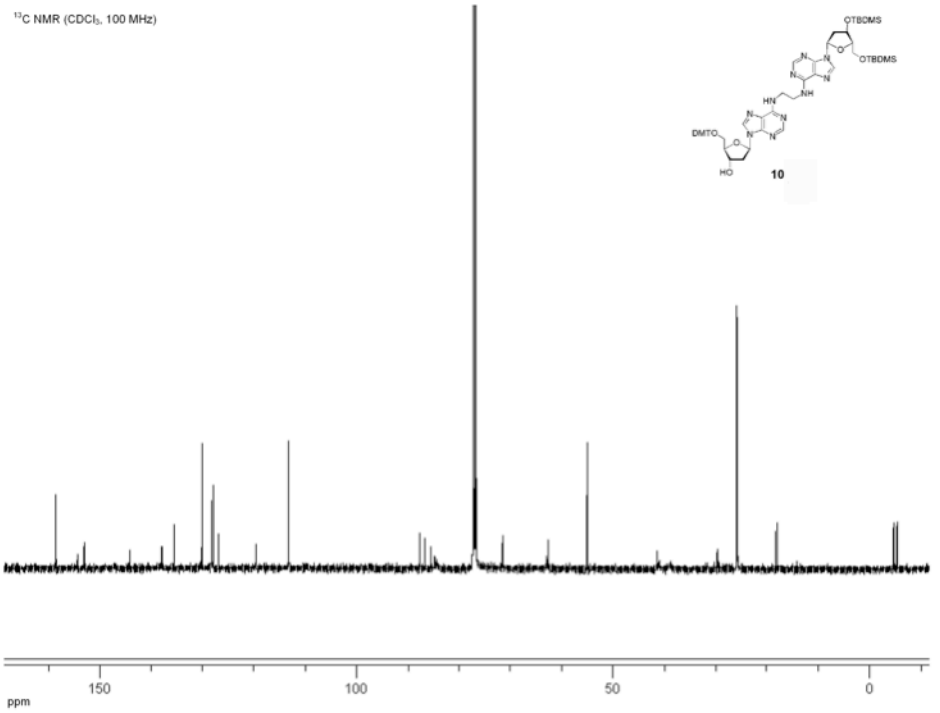


Figure S26. ^{13}C NMR spectrum of compound **10**

Table S2. T_m data of unmodified, 3'-B^C-modified and 3'-B^A-modified double-stranded DNAs. Conditions: 1 μM siRNA, 10 mM phosphate buffer (pH 7.0), 0.1 mM EDTA and 100 mM NaCl.

ON	Sequence	T_m (°C)
12 15	3'-AGGTCTTGTTTCCTTTGC-5' 3'-GCAAAGGAAACAAGACCT-5'	56.7
13 15	3'-B^CAGGTCTTGTTTCCTTTGC-5' 3'-GCAAAGGAAACAAGACCT-5'	56.4
14 15	3'-B^AAGGTCTTGTTTCCTTTGC-5' 3'-GCAAAGGAAACAAGACCT-5'	56.2

Figure S27. Overlay of CD spectra of unmodified and 3'-B^C-modified DNA (A) and siRNA (B) duplexes. Conditions: 1 μM siRNA, 10 mM phosphate buffer (pH 7.0), 0.1 mM EDTA and 100 mM NaCl.

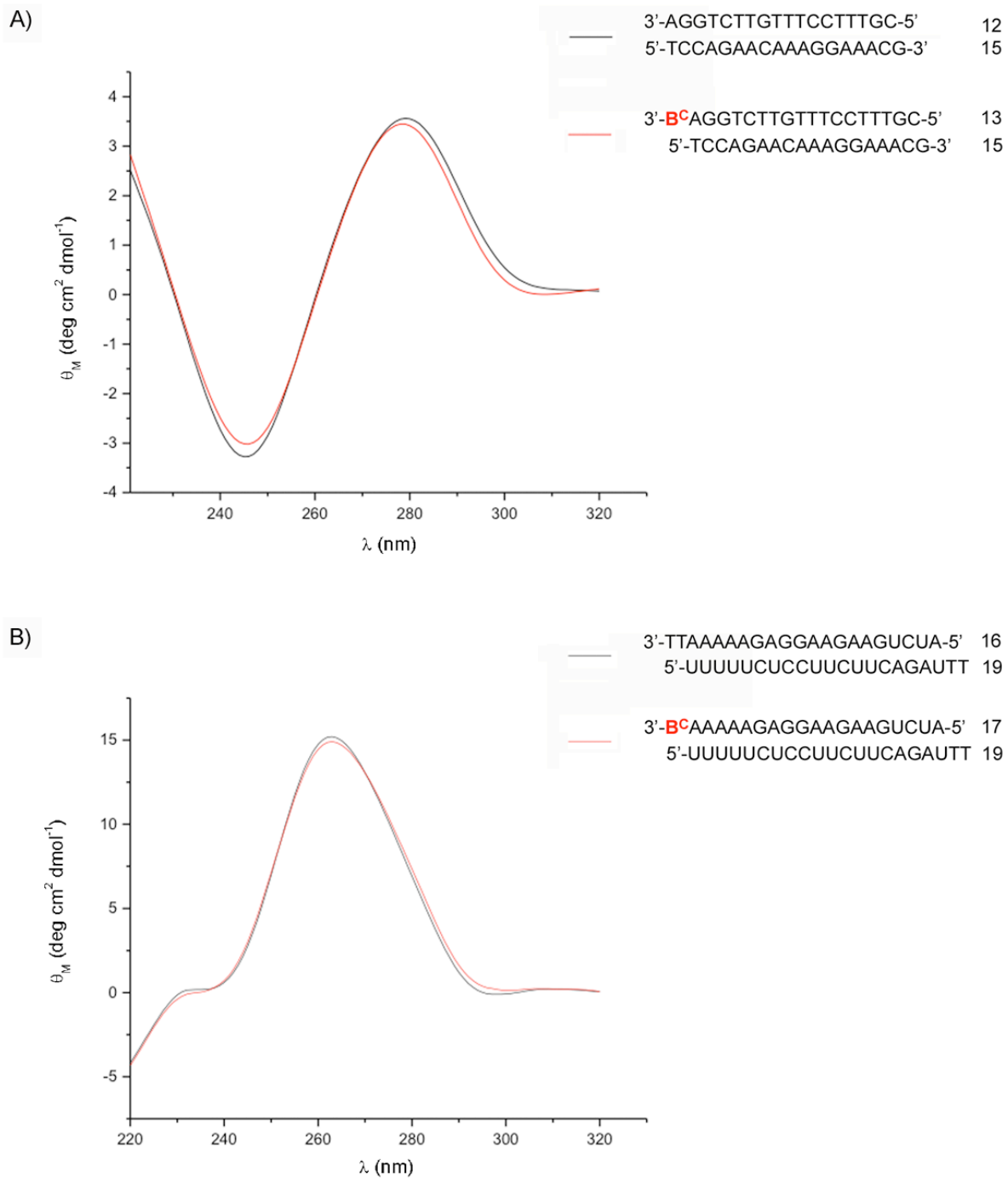
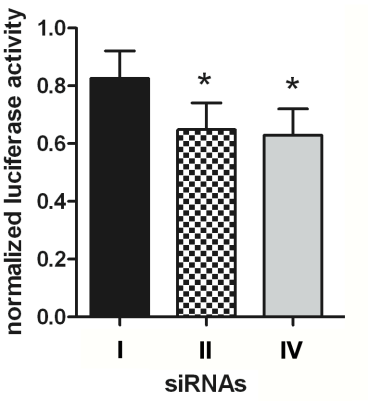


Figure S28. Separate gene silencing activities for unmodified and modified siRNAs **I**, **II** and **IV** targeting the *Renilla* luciferase mRNA in SH-SY5Y cells. SH-SY5Y cells were transfected with dual reporter plasmids that express *Renilla* luciferase (the target) and non-targeted firefly luciferase as an internal non-targeted control and with siRNAs (8 pM and 2 pM per well; panels A and B, respectively) containing 3'-sense-B^C and 3'-sense-B^A substitutions (**II** and **IV**) and compared with the unmodified counterpart (**I**). Normalized *Renilla* luciferase activity is shown. Experiments were carried out in triplicate. Bars indicate standard deviation. A Bonferroni test was conducted to evaluate B^C and B^A modifications to the unmodified control (**I**). In all figures, * indicate significant changes in *Renilla* luciferase expression from unmodified siRNA **I** ($p<0.05$).

A) 8 pM siRNA per well



B) 2 pM siRNA per well

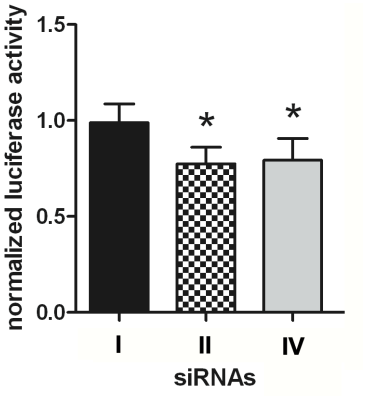


Figure S29. Nuclease stability studies. 20% Denaturing polyacrylamide gels depicting the time course of KF-catalyzed degradation of fully-PS-modified single-stranded DNA **30** (A), SNVPD-catalyzed degradation of 3'-4PS-modified and fully PS-modified single-stranded RNAs **32** and **33** (B), and SNVPD-catalyzed degradation of unmodified, 3'-B^C-modified, 3'-B^A-modifed, 3'-PS-modified and fully PS-modifed single-stranded DNAs **12-14**, **29** and **30**.

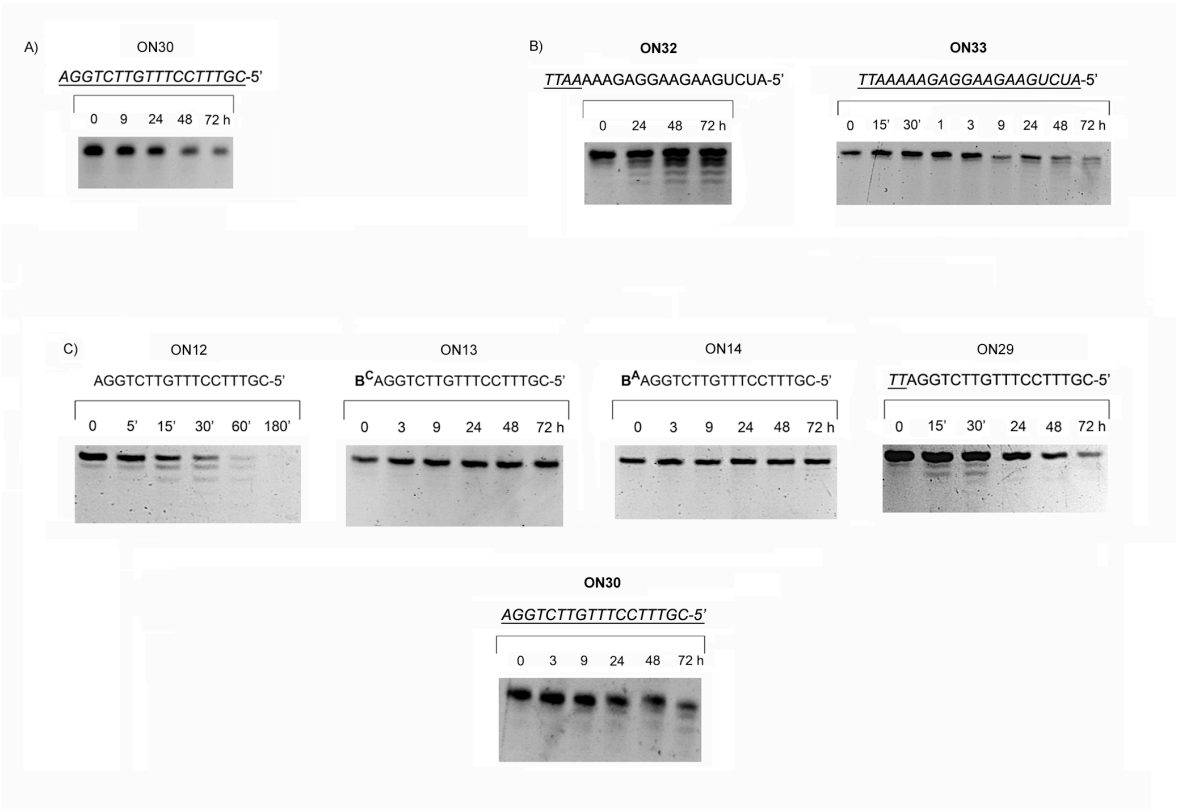
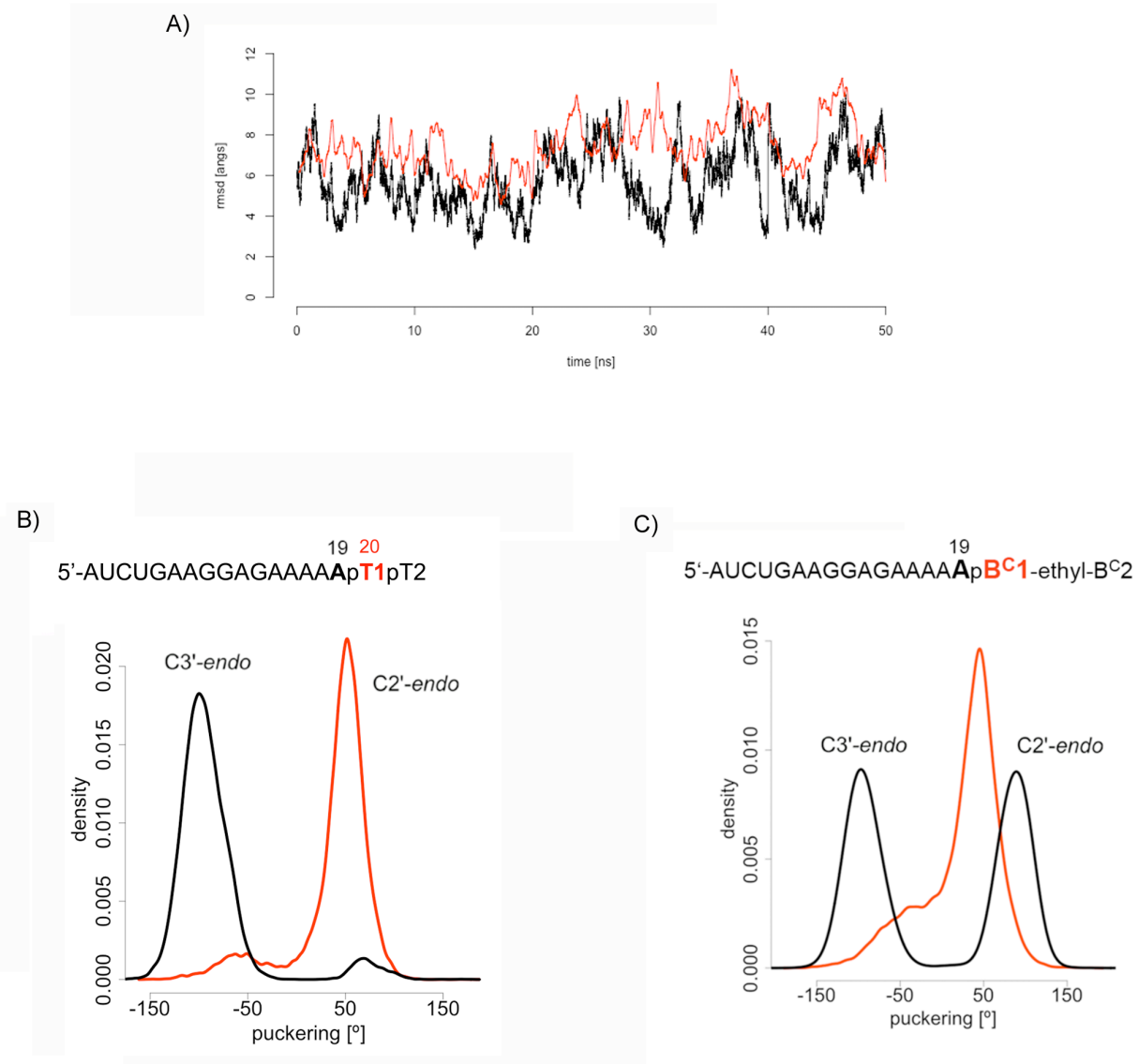


Figure S30. MD simulations of unmodified and 3'-B^C-modified single-stranded RNA with sequences 5'-AUCUGAAGAAGGAAGAAAAA(19)TT and 5'-AUCUGAAGAAGGAAGAAAAA(19)B^C (B^C = B^C1-ethyl-B^C2). B^C1 and B^C2 refer to each of the units composing the dimer). (A) Backbone RMSD calculations with native average structure took as reference. In black RMSD for the native structure simulation and in red RMSD for the 3'-B^C-modified structure simulation. The variations of RMSD along the MD trajectories are quite similar. (B) Populations of delta torsion angles related to the sugar pucker of the residues 19 (A; black) and 20 (T1; red) of the unmodified 21nt-oligonucleotide single-stranded RNA. (C) Populations of delta torsion angles related to the sugar pucker of the residues 19 (A; black) and 20 (B^C1; red) of the 3'-B^C-modified 21nt-oligonucleotide single-stranded RNA.



A.4 Toward a consensus view of duplex RNA flexibility.

Ignacio Faustino, Alberto Pérez, and Modesto Orozco.
Biophysical Journal. 2010, 99(6), pp.1876–1885.

Biophysical Journal, Volume 99

Supporting Material

Towards a Consensus View of Duplex RNA Flexibility?

Ignacio Faustino, Alberto Perez, and Modesto Orozco

SUPPLEMENTARY METHODS

Simulation details. All CHARMM27 simulations were carried out with NAMD (43) computer program applying the Langevin piston method and RESPA (Reversible System Propagator Algorithm) for enhancing the computation speed. The AMBER-suite code pmemd (44) was used for all parmbsc0 calculations using the Berendsen thermostat.

Entropy Calculations. Schlitter's (49) and Andricioaei-Karplus (50) methods were used for post-processing of MD trajectories as described elsewhere (37). When comparing equivalent DNA_2 and RNA_2 sequences, equivalent atomic composition was obtained by transforming (in the trajectory files) methyls at position 5' of thymines and 2'OH of riboses into hydrogens (note that entropies computed in this way have not physical sense in absolute terms, but are useful to discuss DNA vs RNA flexibility). Similarly, when comparing entropies in a duplex type for different sequences the bases were neglected upon the N9 or N1 atom.

Comparison of flexibility patterns. The absolute (γ) and relative (Γ) similarity measures between two trajectories are computed as:

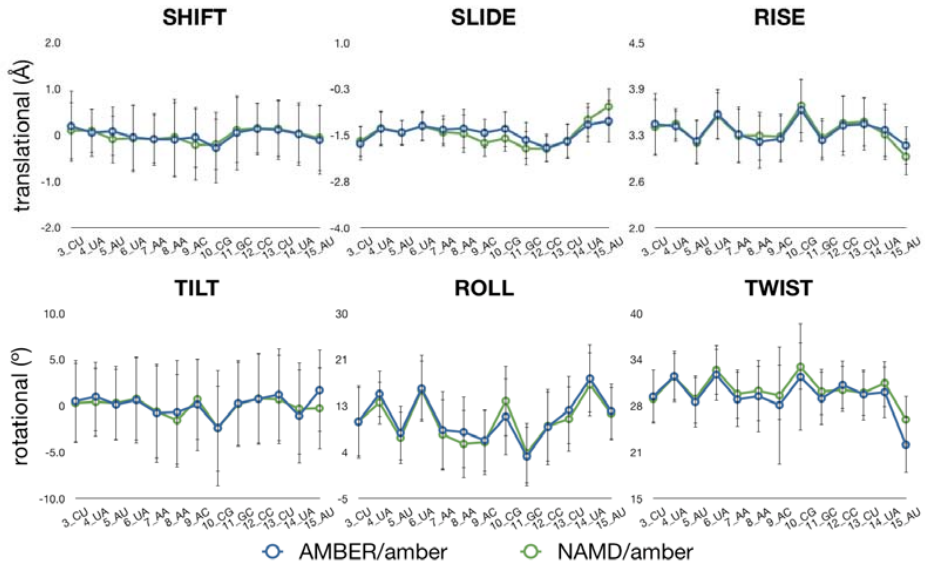
$$\gamma_{AB} = \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\mathbf{v}_i^A \bullet \mathbf{v}_j^B\right)^2 \qquad (1)$$

where M is the minimum number of eigenvectors defining a threshold of variance (between 80-90 % in our case).

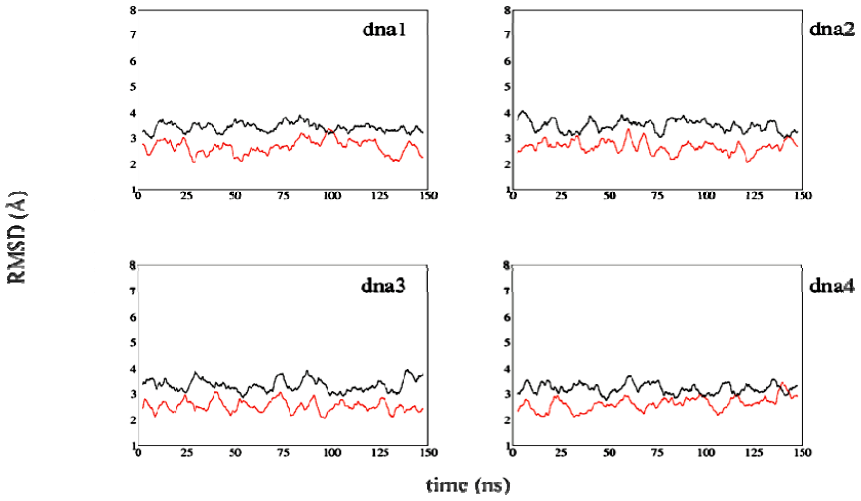
$$\Gamma = \frac{2\gamma_{AB}}{\left(\gamma_{AA} + \gamma_{BB}\right)} \qquad (2)$$

where the self-similarity indexes γ_{AA} and γ_{BB} were obtained by comparing first and second halves of trajectories and 1 means identity and 0 orthogonal trajectories.

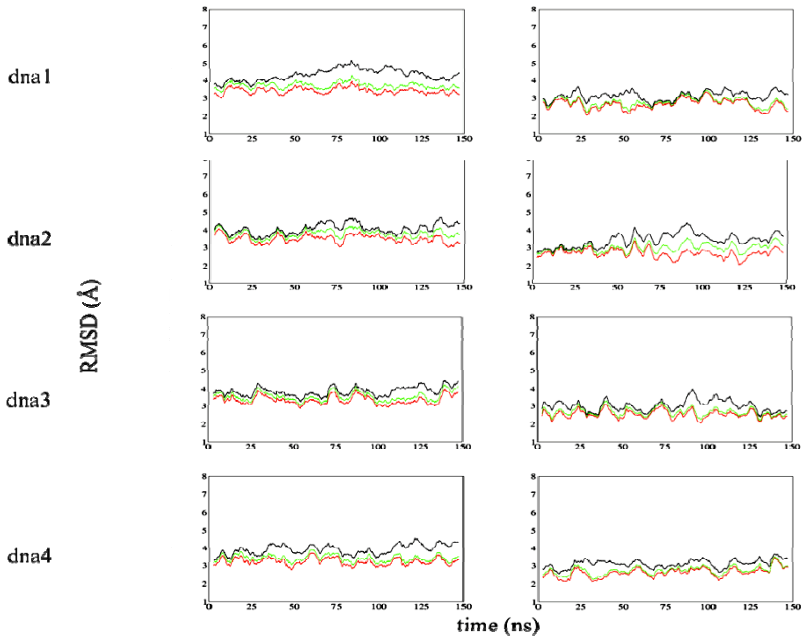
SUPPLEMENTARY FIGURES



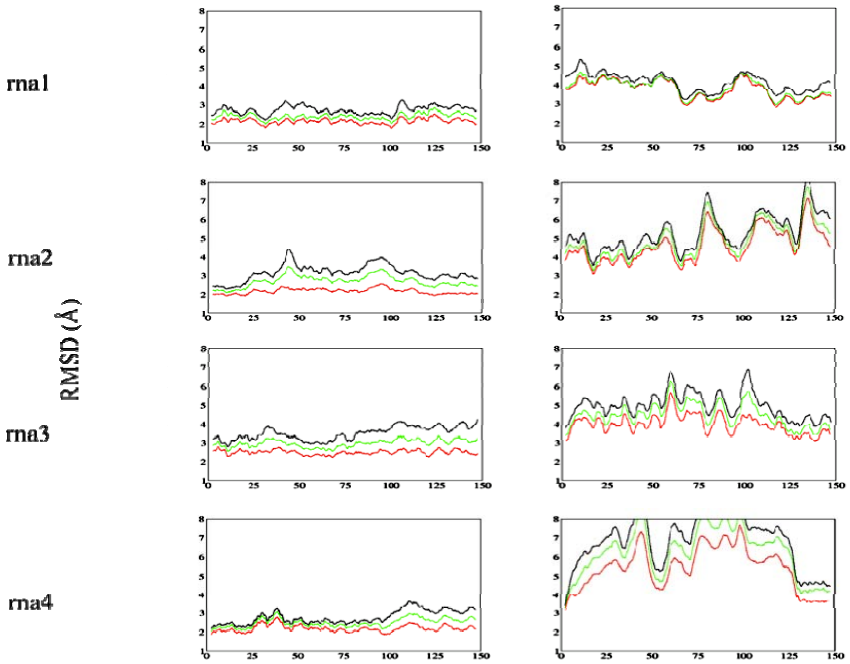
Supplementary Figure 1. Average helical parameters along RNA sequence 1 for simulations run with the amber force-field and NAMD engine (in green) and the amber force-field with AMBER MD engine (in blue).



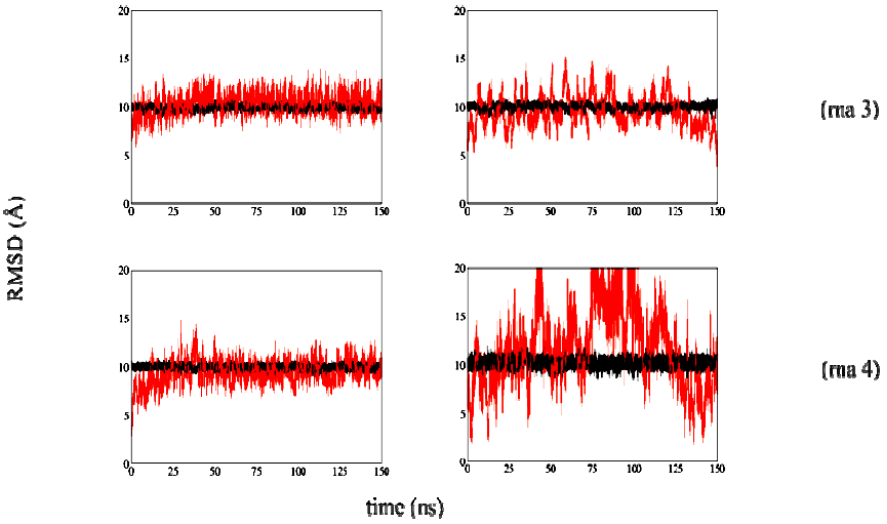
Suplementary Figure 2. Smoothed RMSd (in Å) from fibre conformation for DNA structures for parmbsc0 (in black) and charmm27 simulations (in red).



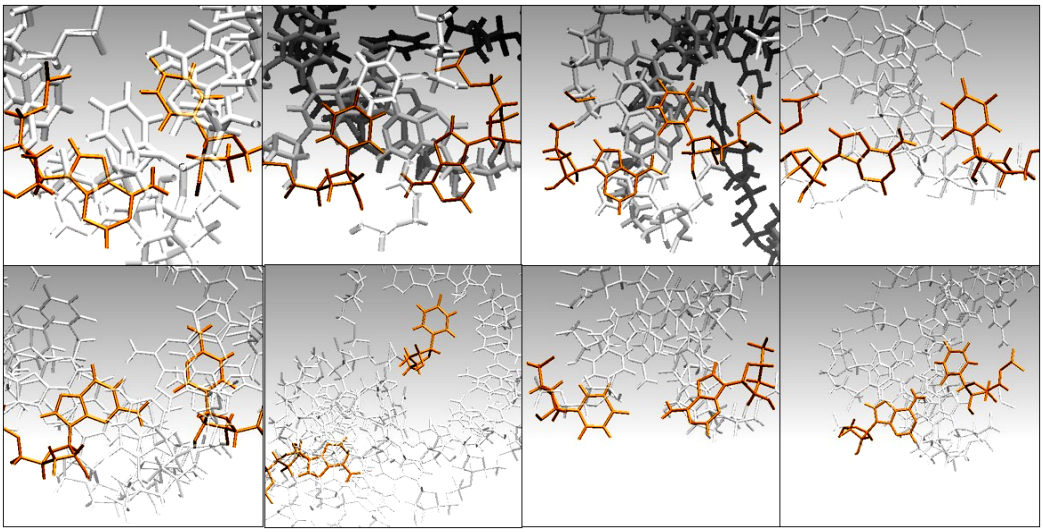
Supplementary Figure 3. Smoothed RMSd (in Å) from average structure for DNA/parmbsc0 (on the left) and DNA/charmm27 simulations (right) for 18-mer (in black), for central 16-mer (in green) and the central 14-mer (in red).



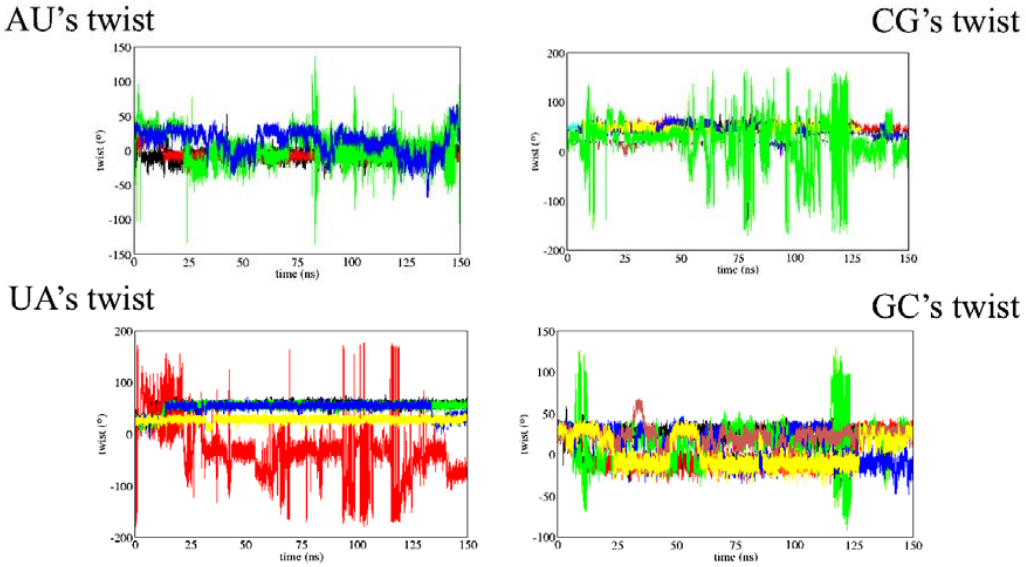
Supplementary Figure 4. Smoothed RMSd (in Å) from average structure for RNA/parmbsc0 (on the left) and charmm27 simulations (right) for 18-mer (in black), for central 16-mer (in green) and the central 14-mer (in red).



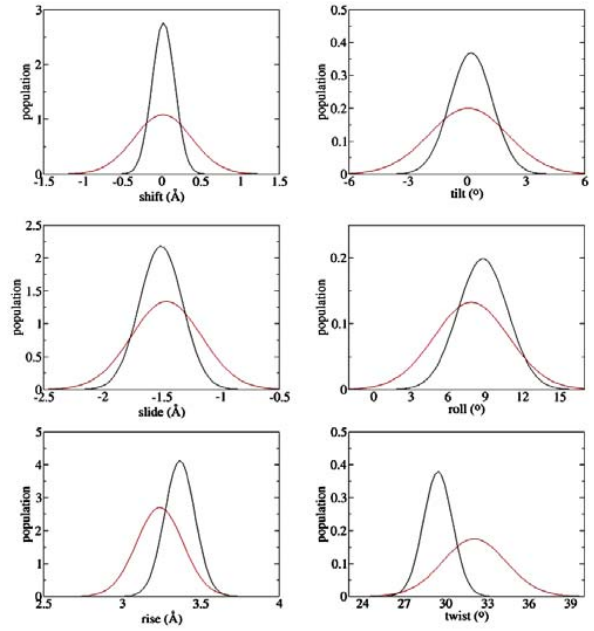
Supplementary Figure 5. Comparison of the average time evolution of the minor (in black) and major grooves (red) geometries for the central 14-mer of sequences 3 and 4 for the corresponding parmbsc0 (on the left) and charmm27 (right) trajectories.



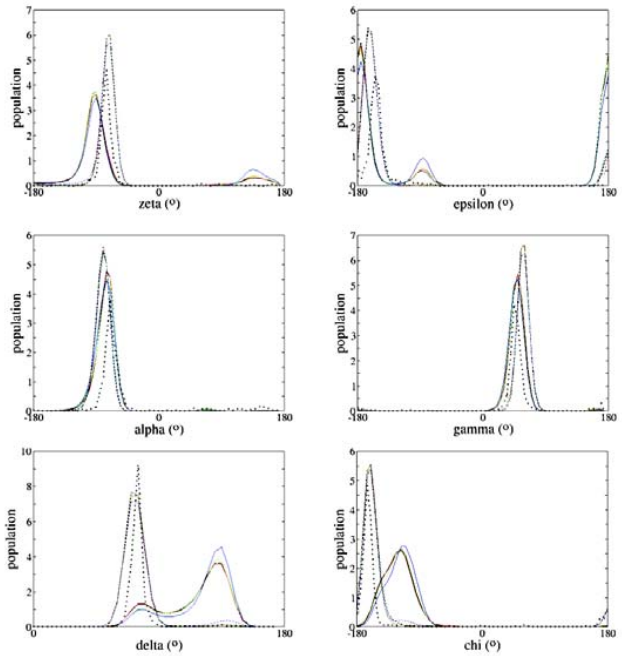
Supplementary Figure 6. Examples of opened base pairs from last structures taken from RNA/CHARMM27 trajectories.



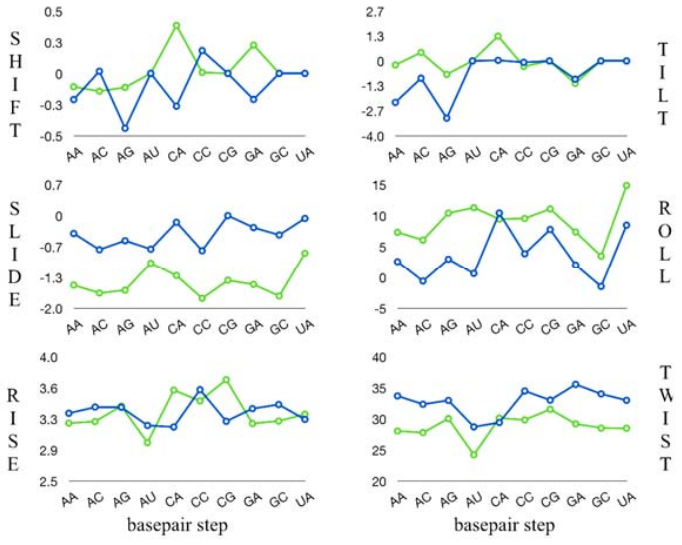
Supplementary Figure 7. Time evolution of twist helical parameter for all AU (top on the left), CG (top on the right), UA (bottom on the left) and GC (bottom on the right) dinucleotide steps present in the four RNA molecules simulated with charmm27 force-field.



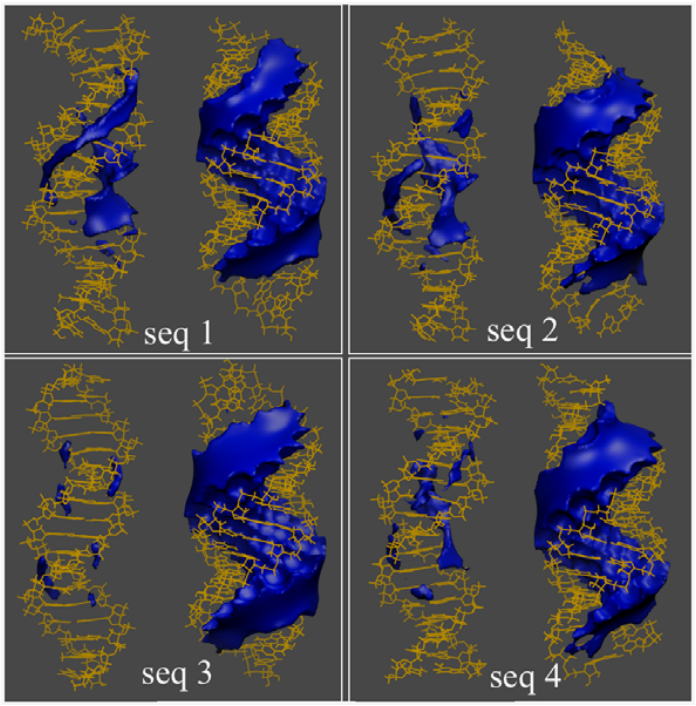
Supplementary Figure 8. Averaged distributions of the base pair step helical properties over the central 14-mer of each sequence for RNA/parmbsc0 trajectories (in black) and averaged distributions derived from RNA X-ray structural data (red).



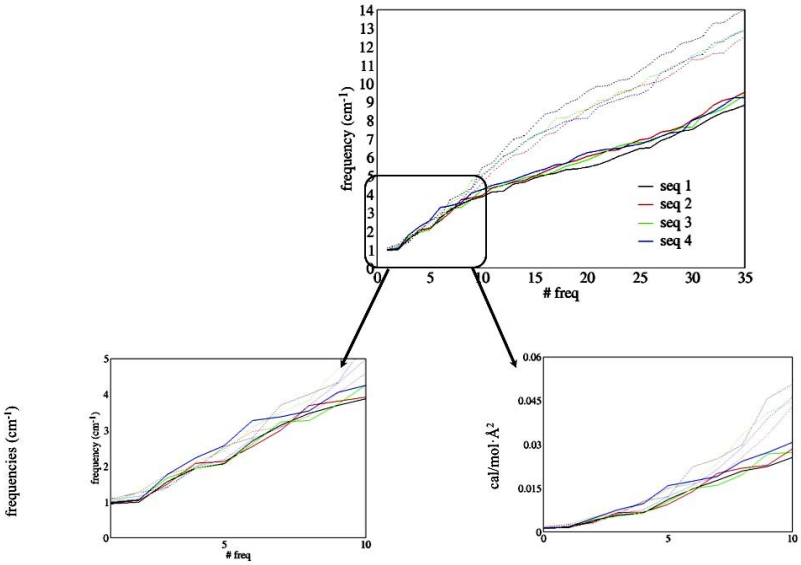
Supplementary Figure 9. Histograms of the six backbone dihedral angles for DNA (straigth lines), RNA (dotted) and X-ray RNA database (dotted black line).



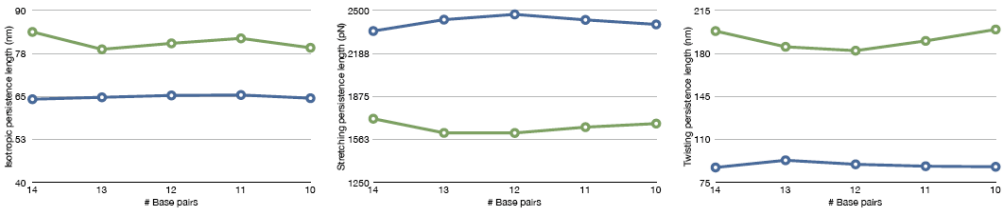
Supplementary Figure 10. Average helical parameters for the 10 unique base pair steps for DNA (blue) and RNA (green) from parmbsc0 simulations.



Supplemenatary Figure 11. Classical molecular interaction potential maps for the four sequences being DNA on the left and RNA on the right computed for the corresponding time-averaged structure. Countour plots showed here correspond to -5 kcal/mol in all cases and correspond to a Na^+ probe particle.



Supplementary Figure 12. On the top, frequency (in cm^{-1}) of the first 35 principal components for the four sequences considering only values derived from parmbsc0 simulations. At the bottom, representations of the corresponding frequencies (on the left) and the force constants (on the right, in cal/mol·$Å^2$) associated with the first 10 deformation modes of the four DNA and RNA duplexes. Straight lines correspond to DNA simulations and dotted lines to RNA trajectories.



Supplementary Figure 13. Variation of DNA (in blue) and RNA (in green) bending (anisotropic and isotropic) and twisting persistence lengths (in nanometers) and stretch modulus (in picoNewtons) with fragment length.

Biophysical Journal, Volume 99

Supporting Material

Towards a Consensus View of Duplex RNA Flexibility?

Ignacio Faustino, Alberto Perez, and Modesto Orozco

Step		Shift		Slide		Rise		Tilt		Roll		Twist	
AA·TT	(4/4)	-0.11	0.02	-1.50	0.10	3.20	0.07	-0.22	0.71	7.31	0.86	28.02	0.77
	(4/2)	*0.30*	*0.35*	*-1.68*	*0.32*	*3.37*	*0.01*	*-0.81*	*0.33*	*8.02*	*0.18*	*25.83*	*6.63*
	(73/21)	**0.23**	**0.84**	**-1.40**	**0.47**	**3.04**	**0.40**	**-1.20**	**5.44**	**7.67**	**4.72**	**30.79**	**5.88**
AC·GT	(7/7)	-0.14	0.09	-1.67	0.03	3.22	0.02	0.45	0.26	6.06	0.49	27.76	0.25
	(7/4)	*-0.58*	*1.07*	*-1.57*	*0.23*	*3.26*	*0.02*	*-1.24*	*0.35*	*8.08*	*1.25*	*30.37*	*3.24*
	(27/39)	***0.12***	***0.87***	***-1.21***	***0.51***	***2.92***	***0.75***	***0.50***	***3.23***	***5.21***	***6.43***	***34.26***	***7.06***
AG·CT	(4/4)	-0.11	0.05	-1.61	0.04	3.40	0.04	-0.72	1.26	10.40	0.14	30.02	1.36
	(4/2)	*-0.32*	*0.15*	*-1.52*	*0.05*	*3.25*	*0.05*	*-0.67*	*0.55*	*7.51*	*1.72*	*27.32*	*0.76*
	(18/37)	**-0.05**	**0.78**	**-1.38**	**0.44**	**3.26**	**0.17**	**-0.38**	**2.45**	**8.34**	**3.61**	**33.21**	**5.04**
AT·AT	(3/3)	0.00	0.00	-1.02	0.02	2.96	0.10	0.00	0.00	11.27	2.58	24.18	0.32
	(3/0)	-	-	-	-	-	-	-	-	-	-	-	-
	(55/15)	**0.00**	**0.64**	**-1.39**	**0.40**	**3.24**	**0.20**	**0.00**	**2.50**	**7.23**	**4.30**	**34.39**	**5.28**
CA·TG	(5/5)	0.39	0.45	-1.29	0.00	3.60	0.24	1.33	2.13	9.43	3.11	30.11	1.41
	(5/1)	*0.29*	-	*-1.53*	-	*3.64*	-	*0.55*	-	*9.40*	-	*26.00*	-
	(40/33)	**0.29**	**0.72**	**-1.48**	**0.41**	**3.19**	**0.41**	**-0.11**	**3.02**	**8.51**	**5.80**	**30.65**	**6.16**
CC·GG	(8/8)	0.01	0.08	-1.79	0.06	3.47	0.03	-0.30	0.74	9.53	0.63	29.82	0.44
	(8/8)	*0.08*	*0.12*	*-1.82*	*0.04*	*3.36*	*0.06*	*1.71*	*0.61*	*8.86*	*0.66*	*27.92*	*0.37*
	(63/81)	**0.11**	**0.63**	**-1.67**	**0.36**	**3.19**	**0.38**	**0.75**	**4.27**	**7.30**	**3.87**	**32.21**	**5.71**
CG·CG	(8/8)	0.00	0.00	-1.40	0.17	3.72	0.28	0.00	0.00	11.06	0.52	31.52	0.29
	(8/2)	*0.00*	-	*-1.73*	-	*3.73*	-	*0.00*	-	*12.54*	-	*25.02*	-
	(168/27)	**0.00**	**1.43**	**-1.62**	**0.55**	**3.37**	**0.47**	**0.00**	**4.06**	**11.03**	**5.11**	**29.84**	**8.61**
GA·TC	(5/5)	0.23	0.01	-1.49	0.06	3.19	0.01	-1.20	0.44	7.36	0.67	29.17	0.39
	(5/3)	*0.08*	*0.00*	*-1.69*	*0.05*	*3.41*	*0.03*	*-0.83*	*0.25*	*9.57*	*0.82*	*30.07*	*0.66*
	(68/27)	**-0.09**	**0.67**	**-1.63**	**0.40**	**3.31**	**0.49**	**0.28**	**4.36**	**8.40**	**5.19**	**30.81**	**5.67**
GC·GC	(5/5)	0.00	0.00	-1.74	0.18	3.22	0.03	0.00	0.00	3.46	0.67	28.49	0.01
	(5/0)	-	-	-	-	-	-	-	-	-	-	-	-
	(85/41)	**0.00**	**0.92**	**-1.27**	**0.44**	**2.90**	**0.82**	**0.00**	**4.32**	**4.44**	**5.42**	**35.49**	**6.96**
TA·TA	(3/3)	0.00	-	-0.81	-	3.31	-	0.00	-	14.83	-	28.47	-
	(3/2)	*0.00*	-	*-1.48*	-	*3.63*	-	*0.00*	-	*14.54*	-	*29.98*	-
	(11/26)	**0.00**	**0.90**	**-1.39**	**0.24**	**3.20**	**0.32**	**0.00**	**3.60**	**10.97**	**3.18**	**31.24**	**5.66**
Generic		0.00	0.20	-1.54	0.24	3.35	0.21	-0.00	0.94	8.50	2.50	28.96	1.66
		0.00	*0.52*	*-1.66*	*0.17*	*3.37*	*0.13*	*0.00*	*1.23*	*9.21*	*1.54*	*28.18*	*2.78*
		0.00	**0.12**	**-1.44**	**0.15**	**3.16**	**0.16**	**-0.00**	**0.53**	**7.91**	**2.11**	**32.29**	**1.93**

Supplementary Table 1. Average and standard deviations for helical parameters for the 10 unique RNA dinucleotide steps removing those base pairs which lose the interaction with its complementary for at least 50% of the time of simulation in charmm27 trajectories. In roman parmbsc0-derived values, in italics charmm27 values and in bold, data from X-ray structures database. In parenthesis it is showed the number of dinucleotide steps taken for averaging; note that often charmm27 values are obtained for a number of steps smaller than the maximum available one.

	Nucleobase		Backbone	
AT (AU)	11	*12*	29	*31*
CG	12	*12*	29	*32*

	central 14-mer		minor groove		major groove	
seq 1	371	*383*	36	*41*	63	*60*
seq 2	373	*386*	37	*42*	64	*63*
seq 3	373	*389*	39	*43*	62	*64*
seq 4	374	*386*	41	*43*	59	*60*
Average	373	*386*	38	*42*	62	*62*

Supplementary Table 2. Averaged number of water molecules (standard deviations in parenthesis) interacting with the base and the backbone of A·T/A·U and C·G base pairs for DNA (roman) and RNA (italics) (top). Time-averaged number of water molecules interacting with the central 14-mer, in the minor and major grooves for DNA (roman) and RNA (italics) (bottom). Water molecules around 3.5 A around oligo atoms. "Nucleobase" atoms were restricted to base atoms and C1' while "backbone" atoms to the rest of the corresponding nucleotide.

sequence	Γ
seq 1	0.637
seq 2	0.347
seq 3	0.380
seq 4	0.543
Average	0.477

Supplementary Table 3. Relative similarity indexes (see Methods) between DNA and RNA parmbsc0 trajectories considering all atoms with the exception of the atoms attached to ribose position 2 and position 5 of the ring of thymines and uracils.

	Seq 1	Seq 2	Seq 3	Seq 4
Seq 1	1.000	0.393	0.514	0.729
	1.000	*0.526*	*0.477*	*0.813*
	0.824	**0.634**	**0.564**	**0.767**
Seq 2		1.000	0.393	0.393
		1.000	*0.485*	*0.659*
		0.627	**0.479**	**0.404**
Seq 3			1.000	0.457
			1.000	*0.501*
			0.459	**0.463**
Seq 4				1.000
				1.000
				0.795

Supplementary Table 4. Relative similarity indexes (see Methods) for the four RNA simulations by only taking backbone atoms. Values restricted to central 14-mer for comparative purposes. Values in roman style: AMBER/AMBER; italics: CHARMM/CHARMM and bold: AMBER/CHARMM.

	AA	AC	AG	AT/AU	CA	CC	CG	GA	GC	TA/UA
SHIFT	1.76	1.17	1.46	1.23	1.05	1.48	1.11	1.28	1.28	0.62
	1.32	*1.10*	*1.24*	*0.49*	*1.40*	*2.19*	*1.10*	*1.55*	*1.82*	*1.40*
	-	**-**	**-**	**-**	**-**	**1.61**	**-**	**-**	**1.30**	**-**
	0.14	***0.17***	***0.62***	***-***	***0.29***	***0.79***	***0.16***	***0.45***	***-***	***0.40***
SLIDE	2.49	3.00	1.95	4.30	1.96	1.88	2.02	1.63	2.78	1.34
	3.82	*2.85*	*3.84*	*1.83*	*3.52*	*4.70*	*2.91*	*3.90*	*3.07*	*3.36*
	-	**-**	**-**	**-**	**-**	**5.44**	**-**	**-**	**4.89**	**-**
	1.02	***2.92***	***3.24***	***-***	***3.47***	***4.36***	***2.38***	***2.59***	***-***	***3.51***
RISE	7.97	9.16	6.64	9.43	6.54	7.73	6.66	7.79	9.95	6.32
	9.78	*10.96*	*8.81*	*6.22*	*4.33*	*8.99*	*4.06*	*9.84*	*11.40*	*4.34*
	-	**-**	**-**	**-**	**-**	**16.53**	**-**	**-**	**16.95**	**-**
	3.03	***7.68***	***7.74***	***-***	***2.85***	***7.81***	***2.26***	***5.87***	***-***	***3.22***
TILT	0.038	0.038	0.037	0.038	0.026	0.043	0.029	0.037	0.041	0.019
	0.034	*0.032*	*0.035*	*0.013*	*0.032*	*0.054*	*0.029*	*0.045*	*0.036*	*0.017*
	-	**-**	**-**	**-**	**-**	**0.040**	**-**	**-**	**0.030**	**-**
	0.017	***0.024***	***0.030***	***-***	***0.024***	***0.045***	***0.019***	***0.030***	***-***	***0.023***
ROLL	0.022	0.022	0.019	0.024	0.017	0.019	0.016	0.020	0.023	0.015
	0.020	*0.023*	*0.021*	*0.012*	*0.016*	*0.025*	*0.014*	*0.020*	*0.025*	*0.014*
	-	**-**	**-**	**-**	**-**	**0.050**	**-**	**-**	**0.030**	**-**
	0.014	***0.018***	***0.017***	***-***	***0.016***	***0.022***	***0.012***	***0.018***	***-***	***0.014***
TWIST	0.033	0.036	0.030	0.045	0.020	0.035	0.019	0.026	0.030	0.024
	0.058	*0.056*	*0.050*	*0.032*	*0.052*	*0.060*	*0.045*	*0.060*	*0.052*	*0.036*
	-	**-**	**-**	**-**	**-**	**0.060**	**-**	**-**	**0.040**	**-**
	0.004	***0.018***	***0.034***	***-***	***0.016***	***0.030***	***0.017***	***0.020***	***-***	***0.033***

Supplementary Table 5. Force constants for the unique dinucleotide steps associated to local helical parameters for DNA/parmbsc0 (roman), RNA/parmbsc0 (italics), X-ray RNA database analysis (bold) and RNA/charmm27 (bold italic). Translational force constants in kcal/mol Å^2 and rotational ones in kcal/mol deg^2. See text for details.

Base pair step	AA	AC	AG	AU	CA	CC	CG	GA	GC	UA	Average
TRANSLATIONAL	4.07	4.44	3.35	4.99	3.18	3.70	3.27	3.57	4.65	2.76	3.80
	4.97	*4.97*	*4.63*	*2.85*	*3.08*	*5.29*	*2.69*	*5.09*	*5.43*	*3.03*	*4.20*
ROTATIONAL	0.092	0.095	0.086	0.107	0.063	0.096	0.063	0.082	0.094	0.058	0.084
	0.112	*0.111*	*0.105*	*0.057*	*0.100*	*0.140*	*0.088*	*0.124*	*0.112*	*0.066*	*0.102*

Supplementary Table 6. Average translational (in kcal/mol·Å^2) and rotational (in kcal/mol·deg^2) stiffness indexes for the ten unique dinucleotide steps for DNA (roman) and RNA (italics).