



Published in final edited form as:

*J Am Chem Soc.* 2013 July 3; 135(26): 9785–9796. doi:10.1021/ja402525c.

## Relative stability of different DNA guanine quadruplex stem topologies derived using large-scale quantum-chemical computations

Jiří Šponer<sup>1,2,\*</sup>, Arnošt Mládek<sup>1,2</sup>, Nad'a Špačková<sup>1,3</sup>, Xiaohui Cang<sup>4</sup>, Thomas E. Cheatham III<sup>5</sup>, and Stefan Grimme<sup>6,\*</sup>

<sup>1</sup>Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic

<sup>2</sup>CEITEC - Central European Institute of Technology, Masaryk University, Campus Bohunice, Kamenice 5, 625 00 Brno, Czech Republic

<sup>3</sup>Department of Condensed Matter Physics, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic

<sup>4</sup>Institute of Genetics, School of Life Science, Zhejiang University, Hangzhou, China 310058

<sup>5</sup>Department of Medicinal Chemistry, College of Pharmacy, University of Utah, Salt Lake City, UT 84124, USA

<sup>6</sup>Mulliken Center for Theoretical Chemistry, Institute of Physical & Theoretical Chemistry, University of Bonn, Beringstrasse. 4, D-53115 Bonn, Germany

### Abstract

We provide theoretical predictions of the intrinsic stability of different arrangements of guanine quadruplex (G-DNA) stems. Most computational studies of nucleic acids have applied Molecular Mechanics (MM) approaches using simple pairwise-additive force fields. The principle limitation of such calculations is the highly approximate nature of the force fields. In this study we for the first time apply accurate QM computations (DFT-D3 with large atomic orbital basis sets) to essentially complete DNA building blocks, namely, seven different folds of the cation-stabilized 2-quartet G-DNA stem, each having more than 250 atoms. The solvent effects are approximated by COSMO continuum solvent. We reveal sizeable differences between MM and QM descriptions of relative energies of different G-DNA stems, which apparently reflect approximations of the DNA force field. Using the QM energy data, we propose correction to earlier free energy estimates of relative stabilities of different parallel, hybrid and antiparallel G-stem folds based on classical simulations. The new energy ranking visibly improves the agreement between theory and experiment. We predict the 5'-*anti-anti*-3' GpG dinucleotide step to be the most stable one, closely followed by the 5'-*syn-anti*-3' step. The results are in good agreement with known experimental structures of 2, 3 and 4-quartet G-DNA stems. Besides providing specific results for G-DNA, our study highlights basic limitations of force field modeling of nucleic acids. Although QM computations have their own limitations, mainly the lack of conformational sampling and the approximate description of the solvent, they can substantially improve quality of calculations currently relying exclusively on force fields.

\*Corresponding authors: sponer@ncbr.muni.cz, grimme@thch.uni-bonn.de.

#### Supporting Information Available:

Supporting information contains PDB files of all starting and optimized structures, details of calculations on MM<sub>opt</sub> geometries, series of calculations of stability order of different G-DNA stems, and several other analyses. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## Introduction

Nucleic acids adopt a wide variety of structures and possess diverse dynamical behavior, both of which are essential for their biological functions. Beyond experiments, structural dynamics of nucleic acids can be studied by computational techniques. The most common computational method for this purpose is classical (molecular mechanics based,<sup>1–7</sup> MM) atomistic molecular dynamics (MD) simulation.<sup>8–12</sup> Molecular simulations allow explicit inclusion of solvent and a rather large sampling of conformational space, with contemporary simulations reaching easily beyond the microsecond time scale. Both solvent and sampling of conformational space are essential for proper description of the structural dynamics. However, classical simulations of nucleic acids are limited by the approximate nature of the empirical force fields.<sup>6,13,14</sup>

Modern electronic structure (quantum chemical, QM) computations offer inherently more accurate description of a studied system compared to force fields.<sup>15–17</sup> However, the applicability of QM methods to nucleic acids is limited by the size of the systems that can be handled, difficulties in inclusion of solvent and essentially lack of any sampling. Thus, QM calculations in studies of nucleic acids have mostly been used for benchmark computations to derive the basic picture of molecular interactions such as base stacking<sup>18–20</sup> or to provide data for verification and parameterization of the simulation force fields.<sup>6,14</sup> Among studies related to this work, calculations on guanine-quartet (G-quartet) systems should be noticed.<sup>21–28</sup> QM computations have also characterized specific local molecular interactions seen in experimental nucleic acids structures,<sup>29,30</sup> while hybrid quantum-chemical/molecular mechanical (QM/MM) approaches have been used to address reaction mechanisms of RNA enzymes.<sup>31–33</sup> However, in general it is not easy to achieve biochemical relevance in QM computations of nucleic acids.<sup>17</sup> Reliable QM descriptions have been limited to only small fragments of the nucleic acids molecules (dozens of atoms when considering QM methods of quality comparable to the approach used in this study, e.g., up to ~60 atoms in studies of two stacked base pairs).

Considering the approximate nature of the simulation force fields, inclusion of accurate large-scale (hundreds of atoms) QM calculations into the portfolio of computational techniques to study nucleic acids is desirable. In the present work we for the first time apply such computations to a system that represents a sufficiently complete DNA fragment. Specifically, we study seven distinct folds of the cation-stabilized 2-quartet guanine quadruplex stem.<sup>34</sup>

The ability to apply large-scale QM computations to nucleic acids reflects method advances, such as development of fast and accurate computations capable of including the dispersion energy.<sup>35–38</sup> This allows a major increase in the size of the studied systems while retaining the accuracy of the structure-energy mapping.<sup>39</sup> Still, applications of the QM methodology to nucleic acids remain difficult. Even inclusion of hundreds of atoms does not always provide a complete enough system. Moreover, the conformational space accessible by nucleic acids and their fragments is complex, mainly due to the dihedral space of the backbone.<sup>40</sup> Optimization techniques do not allow its unambiguous characterization. The calculations may be easily trapped in local minima of unclear significance and/or be biased by spurious interactions not observed in complete solvated nucleic acids. Additionally, the negative charge associated with each phosphate group necessitates inclusion of a solvent. The solvent screening effects in QM calculations can only readily be incorporated using approximate continuum solvent models, which neglect specific hydration. Thus, even with the improved methods, QM computations of nucleic acids are not straightforward. However, despite these caveats, such computations can be very useful to address many specific

questions, especially in combination with experimental and simulation data. We anticipate that large-scale QM computations will become a common tool to study nucleic acids in the near future.

Guanine quadruplex molecules (G-DNA) are the most important non-canonical DNA structures.<sup>41–50</sup> G-DNA is formed by nucleic acids sequences containing short G-tracts, i.e., two or more consecutive guanines in the strand. The basic structural unit of G-DNA is a planar quartet of cyclically bound guanines stabilized by monovalent ions. Several consecutive quartets stack together to form the G-DNA stem consisting of four G-tracts with the monovalent ions lining up in its inner channel (Figure 1). The G-DNA molecule can consist of four or two separate sequences, or a single sequence, that come together with the G-tracts that make up the G-DNA stem in various combinations of parallel and anti-parallel strand orientations. The biochemically most relevant are single-stranded (monomer) topologies. Monomeric and dimeric quadruplexes need single-stranded loops to connect their G-tracts. The G-DNA molecules can adopt different topological variants that often are very sensitive to the base sequence and the surroundings.<sup>41,43,51–54</sup> One of the specific features of G-DNA topology is alternation of *syn* and *anti* (defined by glycosidic torsion  $\chi$ ) nucleotides in the individual quartets, which is imposed by orientation of the G-tracts.<sup>55</sup> In a given quartet (Figures 2 and 3, Supporting Information Figure S1), guanines belonging to G-tracts (G-strands) running *anti*-parallel with respect to each other must have opposite orientation of the bases along the glycosidic bond while guanines that belong to parallel strands must have identical orientation.<sup>55</sup> Supporting Information Figure S1 depicts all quadruplex topologies and types of G-tracts mentioned in this study.

G-DNA molecules have been often studied by simulation techniques.<sup>34,56–72</sup> Our recent study utilized post-processing MM-PBSA free energy computations to address the basic rules governing the topological preference of G-DNA stems.<sup>34</sup> The simulations were carried out for six two-quartet  $[d(GG)]_4$  stems, which included different strand orientations (Figure 3). The calculations predicted that the four possible 5'-GpG-3' steps (dinucleotides) show the following relative stability order: 5'-*syn-anti*-3' (SA) > *anti-anti* (AA) > *anti-syn* (AS) > *syn-syn* (SS). This subsequently determines the basic rules governing the preferred quadruplex stem topologies and their dependence on the number of consecutive G-quartets.<sup>34</sup> The data suggest that anti-parallel G-quadruplexes tend to maximize the number of *syn-anti* 5'-GpG-3' steps and avoid the unfavorable *anti-syn* and *syn-syn* steps. This rule is consistent with observed topologies of most of the monomeric and dimeric anti-parallel G-quadruplex structures.<sup>34</sup> Nevertheless, the computations were not capable to explain why tetrameric quadruplexes experimentally tend to form all-parallel stems with all-*anti* orientation of nucleobases.<sup>41,55</sup> Thus, we suggested that some force field bias may affect the calculated balance of all-parallel all-*anti* stems versus the antiparallel stems.<sup>34</sup>

In this study, we re-investigate the 2-quartet G-DNA stems geometries using modern QM dispersion-corrected density functional theory (DFT-D3) methods.<sup>73,74</sup> The DFT-D3 method obviously does not allow carrying out QM MD simulations for solvated nucleic acids. However, small but already complete DNA structures can be gradiently optimized while using continuum solvent model such as COSMO<sup>75,76</sup> to screen the electrostatic interactions. Such computations can then be compared with equivalent molecular mechanics optimizations with inclusion of Poisson-Boltzmann (PB) continuum solvent model that has been used in the earlier MM-PBSA free energy computations.<sup>34</sup> Provided the QM and MM optimized structures are similar, we can compare relative energies of the different stem topologies (Figure 3) as predicted by the MM and QM methods. One of the genuine requirements, a constraint that allows unambiguous comparison, is the similarity of the sugar-phosphate backbone conformations. Specifically, the backbone conformations should belong to the same rotameric family (e.g. the combined set of all the backbone

dihedrals)<sup>40,77</sup> in both the QM and MM structures. Our computations reveal surprisingly large differences between the MM and QM descriptions, leading to a major revision in the predicted energy order of different G-DNA stem arrangements.<sup>34</sup>

## Methods

### Initial structures

We have created eight starting structures. Six correspond to those investigated in the earlier MD free energy study (Figure 3).<sup>34</sup> In the AA (i.e., *anti-anti*) model, all the guanines have *anti* glycosidic bond orientations, and the structure has four parallel *anti-anti* steps. The coordinates were taken from the second and the third G-quartets of the first molecule (strands A-D) of the parallel tetrameric quadruplex [d(TG4T)]<sub>4</sub> 0.95 Å X-ray structure (PDB ID 352D).<sup>78</sup> This differs from the preceding MD study<sup>34</sup> where an NMR structure was used. The X-ray structure is more accurate than the NMR structure so it is more reliable for QM gradient geometry optimization. Due to genuine sampling of MD simulations, both starting structures are equally suitable for MD simulations. Three *syn-anti* (SA) models (SA-aabb, SA-abab and SA-aaab - Figure 3) were built as in ref<sup>34</sup>. They all have four SA steps but with different strand orientations ('a' and 'b' denote the strand directions around the G-quartet). The SA-aabb model was constructed using coordinates of the first and the second quartets obtained from the crystal structure of the diagonal antiparallel G-quadruplex [d(G4T4G4)]<sub>2</sub> (PDB 1JPQ; 1.9 Å).<sup>79</sup> The SA-abab structure was built from the first model in the NMR structure of the thrombin-binding DNA aptamer d(G2T2G2TGTG2T2G2), PDB: 148D.<sup>80</sup> The SA-aaab stem was built from the first two quartets in the first model of the NMR structure of the human telomere G-quadruplex d(T2G3T2AG3T2AG3T2AG3A), PDB: 2GKU.<sup>81</sup> The AS model was built from the second and the third quartets of 1JPQ. The '3AA+1SS' model, with three AA steps and one SS step, was obtained from the second and third quartets of 2GKU while using the first NMR model with / g-/g- backbone topology (refer to reference<sup>34</sup> for further discussion).

We have further prepared a second variant of the SA-abab fold (SA-abab-2) using the first and second quartet of the 1.5 Å 2AVH X-ray structure of [d(G4T3G4)]<sub>2</sub> edge-loop quadruplex.<sup>82</sup> Finally, entirely new 1.08 Å 3TVB crystal structure of parallel stranded tetrameric [d(G4)]<sub>4</sub> quadruplex appeared recently.<sup>83</sup> This structure, in contrast to all preceding parallel stranded structures, yet in line with our simulation predictions,<sup>34</sup> shows the first (5 -) quartet being all-*syn* while the remaining quartets are all-*anti*. The first quartet is converted to *syn* due to the absence of any additional nucleotide at the 5 -end of the strands in this structure.<sup>11</sup> This facilitates formation of *syn*-specific stabilizing terminal intra-nucleotide O5 ... N3(G) H-bonds (see below).<sup>34</sup>

All structures contained a single K<sup>+</sup> ion between the quartets, added manually in case of solution structures. The formal charge on the models is -3 (four phosphates and one cation). Our calculations do not require neutralization, since no periodic boundary conditions are used.

### Optimization of AA and AS structures with a subset of X-ray water molecules

The X-ray structures of AA and AS steps reveal bridging water molecules positioned ~3.0 Å from N2(G) of the 5 *anti* guanine and ~2.8 Å from O4 of the following 3 nucleotide (see below). We prepared additional structures of AA and AS systems with one such bridging water molecule per each GpG dinucleotide. Initial positions of oxygen atoms were taken from the experimental structures while the water molecules were oriented to serve as H-bond donor to O4 (G+1) and acceptor to N2(G).

## O5'H...N3(G) H-bonds in 5'-terminal *syn* G nucleosides

In G-DNA, we need to distinguish between guanines that are the 5'-terminal ones in the sequences and all the other guanines that are preceded by any stem, loop or flanking nucleotide.<sup>11,34</sup> The 5'-terminal guanine is the first nucleotide of the whole molecule and thus possesses free O5'-H termini. When such guanine adopts *syn* geometry in the G-tract it can form intramolecular O5 H...N3(G) H-bond (Figure 2). Such H-bonds substantially affect the *syn* vs. *anti* energy difference in computations as well as in experiments.<sup>11</sup> These H-bonds indeed form in experimental structures possessing 5'-terminal guanines. These H-bonds cannot form if there is any nucleotide preceding (in the 5'-direction, i.e., upstream) the *syn* guanosine since the O5' is not free. In computations, care is needed with models having 5'-terminal *syn* guanosines, including cases when the model structure is created by truncation of the experimental structure, since several scenarios can occur.

All X-ray structures with 5'-terminal *syn* guanosines indicate O5 ...N3(G) H-bonds with ~3.0 Å heteroatom distances. However, subsequent formation of these H-bonds in either MM or QM geometry optimizations may be influenced by the initial orientation of the H5T terminal hydrogen, depending also on the optimization protocol (see below). The geometry may be less well defined in NMR structures (see below).

If the *syn* guanosine is not the 5' terminal one in the experimental structure, then the backbone adopts different conformation with the O5 ...N3(G) distance 4.5–5.0 Å (Figure 2, Supporting Information Table S1). Even after truncation of a structure leading to a free 5'-terminus, the H-bonds do not form upon QM optimization using such starting structures.

In MD simulations these H-bonds form in all cases with 5'-terminal guanosines irrespective of the starting geometry due to the genuine sampling of the MD method.<sup>34</sup>

In our starting structures, there are four short O5 ...N3(G) distances in the SA-aaaa model and one in the SA-aabb and SA-abab-2 models. They all correspond to presence of 5'-terminal H-bond in the template part of the experimental structures. A specific case is the SA-abab structure where we observed an alternative O5 ...N2(G) H-bond with the terminal O5 H group serving as acceptor. This was a result of the initial NMR structure used for optimization. Note that the 148D NMR structure is somewhat ambiguous as its different models display a full spectrum of different positions for the terminal O5 H group of the G1 nucleotide.

The interaction is absent in the experimental structures used as templates for the remaining 5' terminal *syn* nucleotides in our 2-quartet stems due to presence of other upstream nucleotides. No terminal H-bonds form in these nucleotides using any minimization protocol, but they still form in simulations.

### QM methods

All *ab initio* QM calculations were performed with the TURBOMOLE 6.3 suite of programs.<sup>84</sup> The structures were fully optimized using the meta-GGA functional TPSS<sup>85</sup> and the D3 London dispersion correction with Becke-Johnson damping<sup>73,74</sup> (dubbed DFT-D3(BJ), or as shorthand in the manuscript as DFT-D3). This method has also been used to get the energy data. To estimate reliability of the calculations, a few single point calculations were reevaluated at higher PW6B95-D3(BJ) level.<sup>86</sup>

All geometries optimized in this study were obtained with inclusion of the continuum solvent, although in some analyses we evaluated their energies without the solvent contributions.

In general, the large all-electron Gaussian AO def-TZVP basis set<sup>87,88</sup> was used in the optimizations. In one case, enlargement of the polarization part of the AO basis to def2-TZVP was applied to understand the influence of the methods choices. All of the QM calculations employ the resolution of the identity (RI) approximation<sup>89–91</sup> for the two-electron Coulomb integrals and also apply the numerical quadrature grid  $m4$ <sup>92</sup> for the integration of the exchange-correlation contribution.

Bulk solvation effects were represented by the COSMO continuum solvation model<sup>75,76</sup> assuming the bulk dielectric constant of 78.5 for water. The atomic van der Waals radii for the molecular cavity construction in COSMO are taken as their defaults in TURBOMOLE (in Å for H: 1.3, C: 2.0, N: 1.83, O: 1.72, P: 2.11, K: 2.22).

Single-point gas phase energies of individual 5'-GpG-3' dinucleotides were evaluated at the same level of theory as the QM optimizations.

The choice of the density functional is key aspect in DFT studies and we have worked intensely in recent years to clarify this issue by extensive benchmarking ranging from thermochemistry to molecular interactions.<sup>93</sup> Structure optimizations are less sensitive and we are using the TPSS-D3 level as very reliable default for several years. The method is sufficient also for relative energies (see also below).

## MM methods

MM calculations were carried out with AMBER<sup>94</sup> and the AMBER12 codes using parmbsc0<sup>14</sup> and parmbsc0+parm\_OL4<sup>4</sup> force fields, which are recent refinements of the original Cornell *et al.* force field<sup>1</sup>. Parm\_OL4 is a parameterization of the torsion of DNA refining the description of the *syn* region.<sup>4</sup> Dang's parameters were used for the description of K<sup>+</sup> ions.<sup>95</sup> Two implicit solvation models were employed namely the Poisson-Boltzmann (PB) procedure used in the preceding MM-PBSA study<sup>34</sup> and the latest AMBER version of the generalized-Born (GB) method.

The PB optimization procedure included minimization (1000 steepest descent cycles followed by a variable number of conjugate gradient cycles with the convergence criterion for the energy gradient set to  $10^{-4}$  kcal/mol/Å) using the Poisson-Boltzmann (PB) procedure with the default AMBER9 settings (dielectric constants of 1.0 and 80.0 for the solute and the solvent, respectively, zero ionic strength, 1.6 Å probe radius, no cutoff for van der Waals and Coulombic interactions, 0.5 Å grid spacing). Atom-type/charge-based radii by Tan *et al.*<sup>96</sup> were applied. In our PB calculations, the nonpolar solvation interactions were separated into the attractive and repulsive terms<sup>97</sup> and the reaction field energy was computed using dielectric boundary surface charges.

GB optimization was carried out with AMBER12 using the Hawkins *et al.*<sup>98,99</sup> pairwise generalized Born model. The applied minimization methods were based on the LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton)<sup>100</sup> algorithm. Other parameter settings were left default (convergence criterion for the energy gradient  $10^{-4}$  kcal/mol/Å, zero ionic strength, no inclusion of the surface area in the solvation term).

There is a significant difference between the PB and GB optimization protocols. While the GB optimizations often resulted in visible geometrical changes due to jumping to nearby local minima, the PB minimization in the current AMBER version is inefficient, likely due to errors associated with inexact forces. The PB optimizations did not significantly depart from the starting structure and essentially only relaxed the bond lengths and angles.

## Estimation of errors in the QM treatment

Below we present a set of large-scale QM calculations on different topologies of 2-quartet G-DNA stems with inclusion of the sugar-phosphate backbone. Because the systems studied are large for a QM treatment with current hardware resources (250+ atoms, 3000–5000 basis functions) the calculations still necessarily involve some approximations. It is important to distinguish between numerical artifacts and inherent methodological problems of the basic quantum chemical approximation (DFT-D3/COSMO in our case). Note that in the QM treatment no system-specific empirical corrections are made. In our study, only relative energies of several closely related systems are important, reducing the requirement on the accuracy of the QM procedure. Further, the most important results of our study are visible from gas phase calculations of single GpG dinucleotides (see below) and are thus not affected by the solvent model.

We estimate the errors for relative energies to be on the order of a few kcal/mol. Many of the following conclusions are based on thousands of DFT-D3 benchmark calculations that we recently conducted for our GMTKN30 meta-data-base<sup>93</sup> and on other studies.<sup>73,101</sup> The capability of the chosen DFT-D3 method to accurately calculate intermolecular forces is documented in the literature beyond any doubt. We have shown recently that the TPSS-D3 method has excellent performance for relative energies of different DNA backbone substates with maximum and average deviations from reference CCSD(T) calculations of 0.5 and 0.2 kcal/mol for a full set of eighteen known DNA backbone families.<sup>102</sup> TPSS-D3 belongs to the most accurate and efficient methods for evaluation of DNA backbone energies, with accuracy almost an order of magnitude better than achieved by the MM in the same test.<sup>102</sup>

Numerical errors in the geometry optimizations are insignificant except for the molecular cavity construction/representation in the COSMO treatment. This prevents converging the QM optimizations very strictly (better than 0.1 kcal/mol).

Considering numerical issues, the RI-approximation, quadrature grid for the exchange-correlation energy and the finite AO basis set provide an estimated numerical precision of about 0.5–1 kcal/mol for relative energies compared to a non-RI, infinitely-sized grid based, and complete basis set result. The def-TZVP basis set is sufficient for structure optimization but yields basis set superposition errors (BSSE) in supermolecular interaction energy computations of 10–20%.<sup>103</sup> However, as we calculate primarily relative energies, the effects of intramolecular BSSE are diminished. Indeed, the test calculations using larger def2-TZVP basis set revealed only modest changes, 2.4 kcal/mol instead of 3.9 and 8.4 kcal/mol instead of 9.8 for the relative energies of the SA-aaab and SA-abab QM<sub>opt</sub> structures, respectively. We suggest that triple-zeta type AO basis sets provide sufficiently converged properties for the present conformational/isomerism problem.

The D3 correction for long and medium range London dispersion effects<sup>73,74</sup> together with standard density functionals has been tested thoroughly for non-covalent interactions in many structurally very diverse, small and large systems.<sup>73,74,93,104</sup> Compared to accurate wave function theory results, DFT-D3 provides a solid 5–10 % accuracy for non-covalent interactions. It can be expected that this high accuracy also transfers to the relative energies which should be accurate to within 0.5 kcal/mol with D3. Another issue is the choice of the basic exchange-correlation functional in DFT. The TPSS functional is of a semi-local type which is prone to so-called self-interaction error (SIE) which can lead to significant error in large systems with localized charges (for recent examples see ref.<sup>105</sup> and, for a general overview about related challenges in DFT see<sup>106</sup>). However, for relative (conformational) energies these errors mostly cancel as long as the basic electronic structure is reasonable. This is the case in our systems as exemplified by relatively large HOMO-LUMO gaps of 3–4 eV (at TPSS/def-TZVP level). As shown in Refs.<sup>101,107</sup> the mentioned SIE related

problems more or less disappear when the electrostatic interactions are screened by COSMO. We further checked this point by performing very costly PW6B95 functional based DFT-D3 calculations. In this hybrid method part of the semi-local exchange energy is replaced by non-local Fock-exchange, which is SIE free. In these test calculations again only minor differences compared to the TPSS results were found, 2.0 kcal/mol instead of 3.9 and 8.0 kcal/mol instead of 9.8 for the relative energies of the SA-aaab and SA-abab QM<sub>opt</sub> structures, respectively. Thus, our reported relative DFT-D3 energies should be in error by at most 2–3 kcal/mol (conservative estimate). This would have no effect on any conclusion of our study. The PW6B95-D3 method is probably the best method that can be used for such large systems. It is the overall best performing hybrid functional for our huge GMTKN30 database.<sup>93</sup> For the same reason (i.e., well documented robustness for energies) it was also recently applied for the interactions in supramolecular systems and very good results in comparison to experimental data (binding affinities) were obtained.<sup>108</sup> Thus, the agreement between the TPSS-D3 and PW6B95-D3 energies is a very important test.

The absolute accuracy of the COSMO model is difficult to estimate but of no major concern here as we merely use it to realistically describe the electrostatic interactions for a (in reality) solvated and charged system. As we are mainly interested in a comparison of MM and QM results and both methods employ conceptually similar continuum solvation models, we do not expect a big impact by the specific choice of COSMO. In test calculations employing different dielectric constants of 50.0 and 30.0, respectively, we obtained differences for the relative energy of 3AA+1SS and SA-aabb models with respect to the results for water of only 0.3 and 0.8 kcal/mol. Hence, we conclude that although the solvation model is absolutely essential for a system with several localized negative charges, its implementation details are unimportant. More importantly, the basic differences between the MM and QM results can be deduced from the gas phase computations of the 5'-GG-3' segments (see below) and thus the results are not due to solvation model.

## Results and Discussion

### QM optimizations

The 2-quartet stems were initially optimized using both the MM-PB and DFT-D3 – COSMO approaches (MM<sub>opt</sub> and QM<sub>opt</sub> structures), each starting from the above-described initial structures. As noted in the Method section, the MM-PB optimization procedure caused only insignificant changes in the dihedral angles of the backbone and other aspects of the starting structure. The minimization merely relaxed the bond lengths and angles. In contrast, the QM optimization resulted in larger structural changes (see below). Comparison of the QM and MM relative energies using such disparate structures would be inconsistent. To overcome this, we carried out MM-PB optimizations using the QM<sub>opt</sub> structure as the starting structure (QM – MM<sub>opt</sub>). Since the PB MM optimization again did not change the structures significantly, the QM – MM<sub>opt</sub> and QM<sub>opt</sub> structures allow consistent comparison of relative energies of different stem arrangements obtained by the MM and QM energy evaluations. All starting and optimized geometries are provided in the Supporting Information.

### Large energy difference between QM and MM descriptions

The first two lines of Table 1 summarize the relative energies obtained by the QM<sub>opt</sub> and QM – MM<sub>opt</sub> calculations, derived with respect to the AA stem (Figure 3). The values in parentheses show alternative energies after elimination of the 5'-terminal O5 ...N3(G) *syn* intramolecular H-bonds, where relevant (see Methods). This alternative calculation is explained in details in Supporting Information text and Figure S2.



Four points are striking: i) The differences between the QM and MM relative energies are large; ii) the variability of the G-DNA stem (i.e., two-quartet quadruplex) energies predicted by the QM approach is much smaller than obtained by the MM method, iii) there is a good correspondence between the MM and QM data when mutually comparing different SA structures but iv) the QM method systematically destabilizes the SA structures compared to the reference AA structure relative to the MM. The latter two points are essential to understand our modified prediction of relative stabilities of G-DNA stems presented below.

We suggest that the QM energies are in better agreement with known experimental data. Many G-DNA forming sequences adopt multiple and often co-existing folds, depending on details of experimental conditions.<sup>34,41,43,55,56</sup> This argues against large energy differences between different stem arrangements. In addition, tetrameric quadruplexes not constrained by loops form all-parallel all-*anti* stems in experiments, with the specific exception of one recent structure with 5'-terminal guanines<sup>83</sup> discussed in the Methods section. The low stability of the AA stem compared to the SA stems has been noticed in the preceding MD simulation MM-PBSA study as a potential force field imbalance.<sup>34</sup> Indeed, the present QM calculations shift the balance in favor of the parallel-stranded all-*anti* AA arrangement.

The differences between parmbsc0 and parmbsc0+parm\_OL4 force field computations are small (line 4 of Table 1) and we will further utilize only the parmbsc0 data. The last line of Table 1 shows that the earlier MM-PBSA free energy calculations derived from the parmbsc0 simulation trajectories<sup>34</sup> show similar, albeit muted, trends compared with the current single-point MM data. Notable also is the rather large energy difference between the two alternative SA-abab structures (14.2 kcal/mol, QM energies) favoring the SA-abab-2 model that was optimized starting from the high-resolution X-ray data. We assume that ~7 kcal/mol of this energy difference is due to the presence of a single O5 ...N3(G) interaction in the SA-abab-2 structure which is absent in the SA-abab structure (see Methods and Supporting Information). The remaining part of the energy difference is due to different backbone conformations seen in the respective X-ray and NMR starting structures.

In conclusion, the most interesting result suggested by Table 1 is the large difference between the QM and MM G-stem energies. In the following parts of the paper we present a series of investigations trying to identify the origin of the QM vs. MM difference. We will investigate two possibilities i) the difference is due to the solute force field not matching the reference QM data or ii) the difference is related to the uncertainty of the solvent model. Finally, we adjust the previous MM-PBSA free energy analysis of different G-stem.<sup>34</sup>

### The energy differences between the QM and MM descriptions originate in intrastrand contributions

We first calculated the energetics of the four isolated 5'-GpG-3' dinucleotides in all G-stems. The calculations were carried as single-point gas phase calculations using the 5'-GpG-3' dinucleotide geometries taken from the respective QM<sub>opt</sub> and QM-MM<sub>opt</sub> optimized stem structures (Table 2 and Supporting Information Table S2). These calculations include the intrinsic backbone energies, the intrastrand stacking energies and the base – backbone interactions. In contrast to Table 1, these single-strand gas phase calculations do not include the following terms: interstrand stacking, the stem-ion interactions and the solvation term. These calculations thus considerably help to understand the origin of the difference between QM and MM descriptions. Mainly, they can prove that the results are not caused by the solvent models.

The best comparison is to evaluate the QM<sub>opt</sub> vs. QM-MM<sub>opt</sub> relative energy differences of a given stem with respect to the reference AA G-stem, i.e., line 2 minus line 1 for both Tables 1 and 2. This result is summarized as the last line of Table 2. Clearly, the differences

in the gas phase energies of the 5 -GpG-3 steps in Table 2 explain most of the energy differences for the complete solvated G-stems in Table 1 (see also the correlation between the Table 1 and Table 2 data in Supporting Information Figure S3). Thus, the sizable difference between the QM and MM energies analyzed in our study is due to description of the intrinsic energetics of the sugar-phosphate backbone, intrastrand stacking and intrastrand base – backbone interactions. Only these terms are included in the gas phase 5 -GpG-3 dinucleotide energy calculations in Table 2. In other words, when considering the QM data as the benchmark, the differences between QM and MM data can be primarily attributed to the approximate nature of the DNA force field.

The values in parentheses in Table 2 are alternative estimates upon elimination of the O5 ... N3(G) H-bonds (see Supporting Information).

### Stacking calculations

Due to genuine G-DNA symmetry, the AA and 3AA+1SS stacking patterns differ from the SA and AS stacking patterns (Supporting Information).<sup>28,109</sup> Thus we have investigated if the difference between QM and MM descriptions could be caused by stacking many-body polarization effects.<sup>110</sup> The calculations show that base stacking is described similarly by QM and MM (Supporting Information Table S3).<sup>28</sup>

### The QM COSMO optimization leads to deviation of some structures from the native geometries and formation of non-native interactions

As noted above, the QM geometry minimization changes some of the studied structures. Let us first discuss the reference AA structure.<sup>78</sup> The QM optimization leads to visible increase of the helical twist while the quartets “buckle” towards each other (Figure 4). It is caused by formation of non-native intrastrand hydrogen bonds between N2 of the first guanine and O4' of the second sugar (Figure 5). This is not a consequence of the truncation of the model system. Such H-bonds could form in real systems, but are prevented by explicit hydration, which is visible in the X-ray structure (see below). The N2(G)...O4 (G+1) distance drops from the experimental value of 4.5 Å to 3.1 Å upon optimization. The explicit solvent MD simulations do not change the twist of the AA stem<sup>34</sup> and the N2(G)...O4 (G+1) H-bonds are not formed.

The geometrical change affects the backbone conformation (Figure 4, Supporting Information Table S4). The starting experimental geometry has common canonical DNA conformation, which is stretched in the course of the QM optimization, including high values around 240°. The optimized backbone conformation looks like mere extension of the canonical backbone conformation without changing the rotameric family. (By rotameric family we mean specific combination of the backbone dihedrals, since backbone of nucleic acids adopts distinct substates due to mutual correlation of the backbone dihedrals).<sup>18,77,102</sup>

The change of the AA structure upon QM optimization reflects lack of explicit solvent in the computations. Still, the structure can be used for comparison of QM and MM methods. We assume that the QM description of the potential energy surface (PES) is intrinsically more accurate than the MM description. We suggest that the AA QM<sub>opt</sub> structure is correct per se, i.e., it is the right structure for the given model system and the given approximation of the solvent description. However, caution should be exerted to not over-interpret the data when using them to assess the performance of MM in explicit solvent MD simulations. It is because the comparison is made on a structure, which is shifted away from the native region to another part of the PES. For example, if the MM backbone is stiffer than the QM backbone, the over-twisted structure with stretched backbone would be penalized by the MM description more than by the QM description (see Supporting Information text and

Figures S4–S6). Then the errors introduced by the force field into the explicit solvent simulations could be smaller than indicated by Tables 1 and 2, albeit the trends should be correct.

QM optimization of the 3AA+1SS stem also results in substantial increase of helical twist with formation of the non-native N2(G)...O4 (G+1) H-bonds in all three 5'-GpG-3' AA steps. No changes of helical twist occur in the remaining structures upon QM optimization (Figure S7). There are large changes of the backbone in the AS structure, forming the non-native N2(G)...O4 (G+1) H-bonds in three out of the four strands (Supporting Information Table S4, Figure S8). SA structures show only occasional local rearrangements of the backbone upon QM optimization (Supporting Information Table S4 and the accompanying text). Note that the N2(G)...O4 (G+1) H-bond can be established only when the first guanine is in *anti* conformation, e.g., in the AA and AS steps, but not in the SA and SS steps.

### Energy computations on the MM<sub>opt</sub> relaxed initial structures

We have also performed MM PB and DFT-D3 COSMO single point energy computations on the MM<sub>opt</sub> structures (Table 3), which differ only insignificantly from the starting structures due to inefficient optimization (see above). Obviously, performing QM single-point energy calculations on MM<sub>opt</sub> structures is somewhat inconsistent. However, these calculations are not biased by any non-native features due to optimization in continuum solvent. Still, the MM description overestimates the relative stability of the SA structures (Table 3). However, the difference is smaller than when using the QM<sub>opt</sub> structures (Table 1) with over-twisting of the AA and 3AA+1SS stems. The Table 3 reduces the energy difference between the MM and QM results for the SA stems (with exception of SA-aaaa) by ~13–15 kcal/mol, but the basic trend from the Table 1 is unchanged.

### Comparison of GB and PB computations

In order to get further insights into the role of different terms, we have reevaluated the MM energies using the GB model of solvent, by performing single point calculations on the QM<sub>opt</sub> structures (Table 4 and Supporting Information Table S5). The GB method modestly reduces the differences between the MM and QM data compared to the PB model. Still, the difference between the QM and MM energies remains sizable. Table 4 also compares gas phase energies of complete stems as evaluated by the QM and MM for the QM<sub>opt</sub> structures. Also this calculation confirms that the force field overstabilizes the SA structures compared to the QM level. Supporting Information text and Table S5 present some additional computations, including MM-GB optimizations. All calculations consistently point to a sizable overstabilization of the SA architectures with respect to the AA topology by the MM.

### Re-optimization of the AA and AS structures with a subset of X-ray water molecules

The main weakness of the above computations is over-twisting and deformation of the QM<sub>opt</sub> AA structure due to the non-native N2(G)...O4 (G+1) H-bonds. The X-ray structure reveals that the above atoms are bridged by water molecules. Similar hydration site is visible also in MD simulations (not shown) and is almost permanently occupied with dynamical water molecules with binding times of dozens to hundreds picoseconds. Therefore, we have carried out new QM optimizations of AA and AS structures, including these four water molecules based on X-ray data (see Methods and Figure 6). The water molecules prevent formation of the non-native H-bonds and eliminate over-twisting of the AA structure. The new QM<sub>opt</sub> structures remained in the native arrangements, although some tendency to adopt higher beta values remains visible (Supporting Information Figure S9 and Table S6). Still, even these native-like structures do not change the basic energy results of our study.

Compared to the Table 1, the difference between the AA structure and the remaining structures is shifted by +7.4 kcal/mol and +0.6 kcal/mol at the  $QM_{opt}$  and  $QM - MM_{opt}$  levels, respectively (Supporting Information Table S7). Thus, the key energy difference between the MM and QM descriptions of reference AA stem and various SA structures is reduced by 6.8 kcal/mol (~25%). It is approximately at midpoint between the data from Tables 1 and 3.

### Significance of the energy difference between the QM and MM descriptions

The computations reveal a non-negligible difference between the internal structure-energy relation (potential energy surface) of the studied G-DNA systems as described by the MM and QM approaches. We suggest that the difference primarily reflects inaccuracy of the MM method. We are not yet able to exactly quantify the origin of the difference. However, the main part of the QM vs. MM difference is primarily attributed to the DNA force field, namely the description of the backbone and coupling between backbone and bases (cf. Table 2 and the pertinent discussion). For reasons explained above, it is not caused by the continuum solvent model. Although there is no straightforward way to use the present QM data directly for force field parametrization, they can serve as important benchmarks for testing force fields (including future polarization force fields) and fast QM methods.

We found yet another result that supports the above suggestions. During explicit solvent simulations (at ~40 ns) the simulated AA structure converts from right-handed to incorrect left-handed quadruplex.<sup>34</sup> We have taken two left-handed AA structures (taken at 50 and 55 ns) and derived their  $QM_{opt}$  structures. The DFT-D3/COSMO approach penalizes these structures by 18 and 23 kcal/mol compared to the reference right-handed AA structure. In contrast, the MM/PB method stabilizes these structures ( $QM - MM_{opt}$  data) by -14 and -21 kcal/mol. Single-strand gas phase calculations equivalent to those in Table 2 show that the QM method penalizes these two left handed structures by 15 and 23 kcal/mol relatively to the MM description.

### Reevaluation of the stability order of different G-stem arrangements

The earlier MD simulation study predicted that the four possible 5'-GpG-3' G-DNA dinucleotides have the following relative stability order: 5'-SA-3' ~ -4 kcal/mol > AA 0 kcal/mol > AS ~ +4 kcal/mol > SS ~5 kcal/mol.<sup>34</sup> The SA and SS steps with 5'-terminal ends have estimated stabilities of ~ -8 and ~ 0 kcal/mol.

This would suggest that G-DNA stems maximize the number of SA steps, avoid AS and SS steps, and the 5-terminal Gs are *syn*. It gives reasonable prediction for the 2-quartet Thrombin Binding Aptamer (15-TBA), which has the SA-abab fold with one terminal *syn* G.<sup>80</sup> For the 3-quartet stems typical for the human telomeric quadruplexes, the prediction would be: hybrid 3+1 topology -9 kcal/mol, basket -5 kcal/mol, parallel all-*anti* 0 kcal/mol and chair 0.0 kcal/mol (see Supporting Information Figure S1, Tables S8–S11 and the accompanying text). The calculations thus predict that the different G<sub>3</sub> stems have rather similar stabilities. Indeed, hybrid, parallel all-*anti* and basket topologies are known from atomistic experiments.<sup>41,43,47,81</sup> The chair topology with three G-quartets has not yet been observed but is not ruled out.<sup>111</sup> It is also fairly consistent with unexpected formation of 2-quartet SA G-stem of the K<sup>+</sup> form of the [G<sub>3</sub>(T<sub>2</sub>AG<sub>3</sub>)T] sequence. The 5'-terminal G imposes *syn* orientation of its first guanine while the loss of four less favorable AS steps is compensated for by the loop and flanking base interactions.<sup>112</sup> However, for the intramolecular parallel G<sub>3</sub> stem the energy order would incorrectly predict fold with four 5'-SAA-3' strands (the first quartet flipped to *syn*) to be more stable than the observed all parallel all-*anti*-structure with four 5'-AAA-3' strands.<sup>113</sup> The model fails for tetrameric G<sub>4</sub> stems lacking 5'-terminal Gs (e.g., (TGGGT)<sub>4</sub>), as the parallel structures with four SAAA or

SASA strands as well as anti-parallel four SASA strands structure are predicted to be considerably more stable than the observed parallel four AAAA strands structure.<sup>78,114,115</sup> For (GGGG)<sub>4</sub>, the observed structure with four SAAA strands<sup>83</sup> is correctly predicted<sup>11</sup> and also the antiparallel four-tetrad stems (1JPQ and 2AVH) are correctly predicted to have four SASA strands.<sup>79,116,117</sup> The prediction likely overestimates stability of the SA step compared to AA step. In order to prefer all-*anti* arrangement for parallel stems, the SA step should be slightly less stable than the AA step.

Therefore, we have added a correction to the original data<sup>34</sup> based on present QM results. Such correction is justified, since the preceding and present study use the same force field and PB model of the solvent. The refined prediction has been obtained in the following manner. For the SA 5'-GG-3' step, we took the sum of MM-QM energy difference for the SA-aaab, SA-abab, SA-abab-2 and SA-abbb stems divided by 16 (the structures have 4×4 SA steps). We did not use the SA-aaaa structure as it was not included in the original study. For reasons explained above (the over-twisting and non-native interactions), we have taken average of the values from Tables 1 and 3, as the Table 1 likely overestimates and Table 3 underestimates the difference. We have taken data with eliminated 5'-terminal H-bonds (in parentheses in Table 1). Alternative calculations with inclusion of the 5'-terminal H-bonds or with inclusion of the SA-aaaa structure would not change the results dramatically. Our data give +4.7 kcal/mol correction, predicting the SA step to be by ~ +1.2 kcal/mol less stable than the AA one. Alternatively, we could use the energy data derived for the improved AA QM<sub>opt</sub> geometry optimized with the X-ray water molecules (Supporting Information Table S7); in this case we would not need to consider Table 3 data for balancing the results. This would lead to essentially identical correction of +4.9 kcal/mol. The calculation for the remaining steps is given in Supporting Information.

We obtain the following stability order of the GpG G-DNA steps (kcal/mol): AA (0), SA (+1.2), AS (+3.5), SS (+7.8), S<sub>5 term</sub>A (-3.2) and S<sub>5 term</sub>S (+3.4). ("S<sub>5 term</sub>" means 5'-terminal *syn*, see Supporting Information Figure S1). Although the data were derived by studies of formally intermolecular 2-quartet stems, the prediction is applicable to any N-quartet tetrameric, dimeric and intramolecular G-DNA stem. It gives estimate of relative intrinsic G-DNA stem stability with the following factors taken into account: number of quartets, strand orientation pattern, *syn-anti* pattern and presence or absence of additional nucleotide at the 5'-end of the strands. The model does not take into account additional effects of loops, ion types and some other effects. The model takes into account solvation effects via the MD simulation runs<sup>34</sup> and using continuum solvent approximation for the energies. Although our stability order is only approximate, it helps separate intrinsic properties of G-stems from the remaining effects.

The adjusted energy order improves the predictions (Supporting Information Tables S8–S11). Mainly, the observed all-*anti* parallel topology is now correctly predicted for the tetrameric stems while four SAAA strands topology (4SAAA) remains correctly predicted for the tetrameric G<sub>4</sub> parallel stem with 5'-terminal guanines. All-*anti* pattern is also correctly predicted for the intramolecular parallel-stranded stems. The 4SASA structure is correctly predicted for the antiparallel G<sub>4</sub> stem, since other antiparallel variants such as 2SAAA + 2ASSS are less stable. The stability order of different G<sub>3</sub> topologies is now as follows: parallel all-*anti* 0 kcal/mol, hybrid ~ +13 kcal/mol, and basket and chair ~ +20 kcal/mol. These differences may look rather large, but as noted above, the results should be primarily considered as relative trends. In addition, the estimates of the AS and SS steps relative energies are less certain than the estimate for the SA step. Importantly, the calculation now predicts the observed *syn-anti* patterns for different G<sub>3</sub> stem topologies: all-*anti* for the all parallel stem, 3SAA+1SSA for the hybrid arrangement, and for chair and basket there are two variants with similar energies, namely 2SSA+2SAA and 2SAS+2ASA.

The later is observed for the basket topology (See Supporting Information). The prediction is also consistent with recent NMR study showing that all-parallel all-*anti* form of d(TGGGT)<sub>4</sub> and d(TGGGGT)<sub>4</sub> is accompanied by minor species having 5'-end *syn* tetrad.<sup>118</sup> This indicates that the AA step is only slightly more stable than the SA step.

## Concluding remarks

Structural dynamics and energetics of nucleic acids have been traditionally studied by Molecular Mechanics (MM) approaches using simple force fields. The main limitation of the MM approaches is the highly approximate nature of the force fields. Therefore, it is tempting to include the recently developed accurate large-scale (hundreds of atoms) QM calculations into the portfolio of computational techniques to study nucleic acids. Here we apply such computations for the first time to nucleic acids. Namely, we study different folds of the cation-stabilized 2-quartet guanine quadruplex stem, complementing recent classical simulation and free energy analysis.<sup>34</sup> Our results are relevant not only for G-DNA molecules. Comparably large differences between MM and QM descriptions may be common also for other nucleic acids systems and may affect their MM-based computational studies. Before advance of large-scale QM computations, we had no appropriate tool to capture such differences.

QM calculations provide assessment of the bias introduced by force fields and include physical-chemistry effects neglected in MM description. Most notable are electronic polarization (induction) effects, which contribute the bulk of important non-additive (many-body) interactions in polar systems. The pair-additive force fields with constant atom-centered point charges neglect sensitivity of the electronic structure of the system to external electric fields while in reality the electronic structures of the individual bases and backbone units respond to their environment.<sup>119</sup> The force field also does not allow sufficiently accurate description of different conformations of the flexible and highly polarizable negatively charged DNA sugar phosphate backbone. The missing contributions are unphysically compensated for by the dihedral terms of the force field.<sup>11,40</sup>

Our study reveals large systematic differences between the MM and QM relative energy order of stability of the different G-DNA stem topologies. The MM approach underestimates the relative stability of G-DNA stems with dominant 5'-*anti-anti-3* steps (AA and 3AA+1SS folds) compared to topologies with 5'-*syn-anti-3* steps (SA-aaab, SA-aabb and SA-abab). Although the magnitude of overstabilization of the SA steps by force field varies in different sets of computations, the trend is unambiguously seen from all computations. It means that MD simulation and free energy studies of G-DNA stems are affected by systematic bias in favor of structures with antiparallel stem arrangements with SA steps. However, we would like to underline that the MM description is certainly sufficiently balanced for many analyses.<sup>11</sup> Nevertheless, combining MM and QM analyses can improve future theoretical predictions. Our study indicates that the limitations of MM methods in studies of nucleic acids may be larger than usually assumed.

When using the QM energy data as a correction to earlier free energy estimates of relative stability of different G-stem folds,<sup>34</sup> improved agreement with the experimental data is achieved. The new model uses both explicit solvent simulation MM data and QM data, with the following factors taken into account: number of quartets, strand orientation pattern, *syn-anti* pattern and presence or absence of additional nucleotide at the 5'-end of the strands. The remaining limitations of the model are explained above. According to the new prediction, the 5'-*syn-anti-3* G-DNA dinucleotide step is slightly less stable than the 5'-*anti-anti-3* step. We confirm the stabilizing role of 5'-terminal H-bonds of 5'-terminal *syn* G-DNA nucleotides.

We also illustrate difficulties facing large-scale QM computations of nucleic acids. The first issue is limited sampling, which depends on the availability of relevant starting structures. Future QM studies could profit from analysis of series of starting structures sampled by the classical explicit solvent MD simulations. The other issue is the remaining incompleteness of the studied system, which in our particular case is representation of the environment via continuum solvent model. We have observed excessive geometrical changes of some molecules including formation of non-native intramolecular interactions. We show that at least in some cases such undesired rearrangements can be prevented by including a subset of explicit water molecules. However, the difference between the MM and QM descriptions is consistently visible already from the “in vacuo” energy calculations (Tables 2 and 4). Thus our basic results are not caused by approximations of the continuum solvent models and may be considered as estimation of genuine errors inherent to the MM description. Further investigations will be needed to see if such large discrepancies between QM and MM descriptions are common also for other types of nucleic acids structures.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was supported by the project “CEITEC - Central European Institute of Technology” (CZ. 1.05/1.1.00/02.0068) from European Regional Development Fund and the Grant Agency of the Czech Republic, grant number P208/11/1822. The present study was also financially supported by the South Moravian Centre for International Mobility within the framework of the “Brno Ph.D. Talent” scholarship program. TEC acknowledges support by NIH R01-GM081411, NSF XRAC MCA01S027 and the University of Utah Center for High Performance Computing.

## References

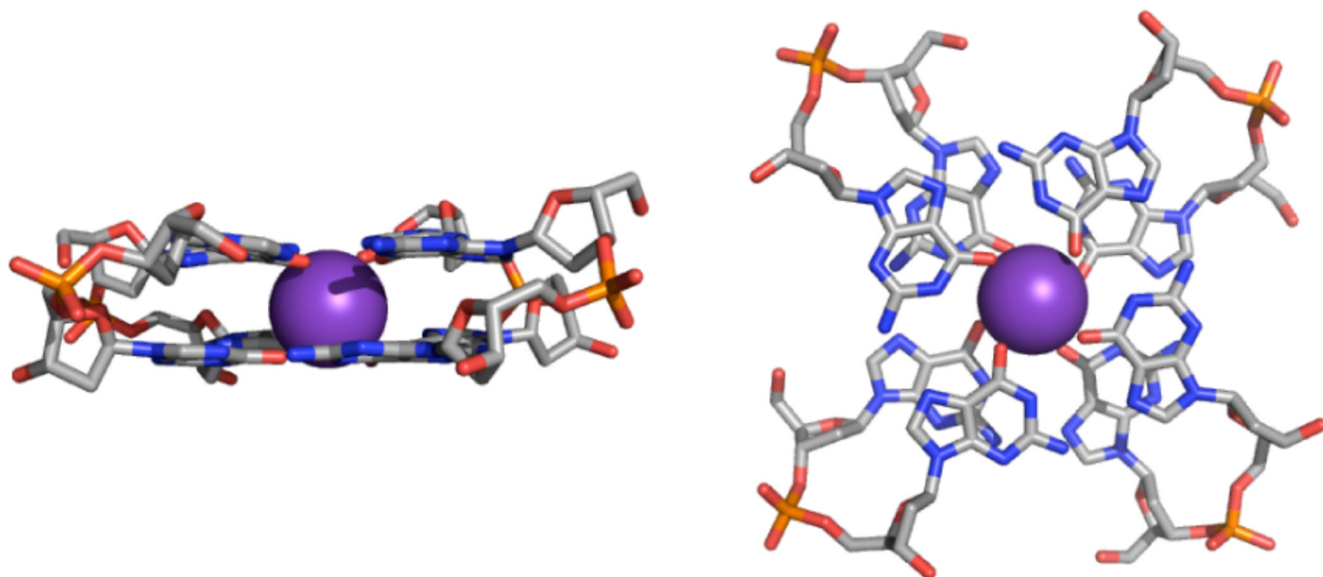
1. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. *J. Am. Chem. Soc.* 1995; 117:5179–5197.
2. Foloppe N, MacKerell AD. *J. Comput. Chem.* 2000; 21:86–104.
3. Hart K, Foloppe N, Baker CM, Denning EJ, Nilsson L, MacKerell AD. *J. Chem. Theory Comput.* 2012; 8:348–362. [PubMed: 22368531]
4. Krepl M, Zgarbova M, Stadlbauer P, Otyepka M, Banas P, Koca J, Cheatham III TE, Jurecka P, Sponer J. *J. Chem. Theory Comput.* 2012; 8:2506–2520. [PubMed: 23197943]
5. Mackerell AD. *J. Comput. Chem.* 2004; 25:1584–1604. [PubMed: 15264253]
6. Zgarbova M, Otyepka M, Sponer J, Mladek A, Banas P, Cheatham TE, Jurecka P. *J. Chem. Theory Comput.* 2011; 7:2886–2902. [PubMed: 21921995]
7. Denning EJ, Priyakumar UD, Nilsson L, Mackerell AD. *J. Comput. Chem.* 2011; 32:1929–1943. [PubMed: 21469161]
8. Cheatham TE. *Curr. Opin. Struct. Biol.* 2004; 14:360–367. [PubMed: 15193317]
9. Ditzler MA, Otyepka M, Sponer J, Walter NG. *Accounts Chem. Res.* 2010; 43:40–47.
10. Orozco M, Noy A, Perez A. *Curr. Opin. Struct. Biol.* 2008; 18:185–193. [PubMed: 18304803]
11. Sponer J, Cang X, Cheatham TE III. *Methods.* 2012; 57:25–39. [PubMed: 22525788]
12. Perez A, Luque FJ, Orozco M. *Accounts Chem. Res.* 2012; 45:196–205.
13. Banas P, Hollas D, Zgarbova M, Jurecka P, Orozco M, Cheatham TE III, Sponer J, Otyepka M. *J. Chem. Theory Comput.* 2010; 6:3836–3849.
14. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE III, Laughton CA, Orozco M. *Biophys. J.* 2007; 92:3817–3829. [PubMed: 17351000]
15. Banas P, Jurecka P, Walter NG, Sponer J, Otyepka M. *Methods.* 2009; 49:202–216. [PubMed: 19398008]
16. Sponer J, Leszczynski J, Hobza P. *Biopolymers.* 2001; 61:3–31. [PubMed: 11891626]

17. Sponer J, Sponer JE, Petrov AI, Leontis NB. *J. Phys. Chem. B.* 2010; 114:15723–15741. [PubMed: 21049899]
18. Mladek A, Sponer JE, Jurecka P, Banas P, Otyepka M, Svozil D, Sponer J. *J. Chem. Theory Comput.* 2010; 6:3817–3835.
19. Sponer J, Leszczynski J, Hobza P. *J. Phys. Chem.* 1996; 100:5590–5596.
20. Parker TM, Hohenstein EG, Parrish RM, Hud NV, Sherrill CD. *J. Am. Chem. Soc.* 2012; 135:1306–1316. [PubMed: 23265256]
21. Gresh N, Pullman B. *Int. J. Quantum Chem.* 1986; 12:49–56.
22. Gu JD, Leszczynski J. *J. Phys. Chem. A.* 2000; 104:6308–6313.
23. Meyer M, Steinke T, Brandl M, Suhnel J. *J. Comput. Chem.* 2001; 22:109–124.
24. Gu JD, Leszczynski J. *J. Phys. Chem. A.* 2002; 106:529–532.
25. Louit G, Hocquet A, Ghomi M, Meyer M, Suhnel J. *Phys. Chem. Commun.* 2003; 6:1–5.
26. van Mourik T, Dingley AJ. *Chem.-Eur. J.* 2005; 11:6064–6079. [PubMed: 16052652]
27. Guerra CF, Zijlstra H, Paragi G, Bickelhaupt FM. *Chem.-Eur. J.* 2011; 17:12612–12622. [PubMed: 21997949]
28. Lech CJ, Heddi B, Phan AT. *Nucleic Acids Res.* 2013; 41:2034–2046. [PubMed: 23268444]
29. Sponer J, Mokdad A, Sponer JE, Spackova N, Leszczynski J, Leontis NB. *J. Mol. Biol.* 2003; 330:967–978. [PubMed: 12860120]
30. Vlieghe D, Sponer J, Van Meervelt L. *Biochemistry.* 1999; 38:16443–16451. [PubMed: 10600105]
31. Mlynsky V, Banas P, Walter NG, Sponer J, Otyepka M. *J. Phys. Chem. B.* 2011; 115:13911–13924. [PubMed: 22014231]
32. Nam KH, Gaot JL, York DM. *J. Am. Chem. Soc.* 2008; 130:4680–4691. [PubMed: 18345664]
33. Trobro S, Aqvist J. *Proc. Natl. Acad. Sci. U. S. A.* 2005; 102:12395–12400. [PubMed: 16116099]
34. Cang XH, Sponer J, Cheatham TE III. *Nucleic Acids Res.* 2011; 39:4499–4512. [PubMed: 21296760]
35. Grimme S. *WIREs Comput. Mol. Sci.* 2011; 1:211–228.
36. Jurecka P, Cerny J, Hobza P, Salahub DR. *J. Comput. Chem.* 2007; 28:555–569. [PubMed: 17186489]
37. Klimes J, Michaelides A. *J. Chem. Phys.* 2012; 137:120901–120912. [PubMed: 23020317]
38. Zhao Y, Truhlar DG. *Accounts Chem. Res.* 2008; 41:157–167.
39. Svozil D, Hobza P, Sponer J. *J. Phys. Chem. B.* 2010; 114:1191–1203. [PubMed: 20000584]
40. Sponer J, Mladek A, Sponer JE, Svozil D, Zgarbova M, Banas P, Jurecka P, Otyepka M. *Phys. Chem. Chem. Phys.* 2012; 14:15257–15277. [PubMed: 23072945]
41. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. *Nucleic Acids Res.* 2006; 34:5402–5415. [PubMed: 17012276]
42. Lane AN, Chaires JB, Gray RD, Trent JO. *Nucleic Acids Res.* 2008; 36:5482–5515. [PubMed: 18718931]
43. Neidle S. *Curr. Opin. Struct. Biol.* 2009; 19:239–250. [PubMed: 19487118]
44. Davis JT. *Angew. Chem. Int. Edit.* 2004; 43:668–698.
45. De Cian A, Lacroix L, Douarre C, Temime-Smaali N, Trentesaux C, Riou JF, Mergny JL. *Biochimie.* 2008; 90:131–155. [PubMed: 17822826]
46. Mergny JL, Mailliet P, Lavelle F, Riou JF, Laoui A, Helene C. *Anti-Cancer Drug Des.* 1999; 14:327–339.
47. Neidle S, Parkinson GN. *Curr. Opin. Struct. Biol.* 2003; 13:275–283. [PubMed: 12831878]
48. Li J, Correia JJ, Wang L, Trent JO, Chaires JB. *Nucleic Acids Res.* 2005; 33:4649–4659. [PubMed: 16106044]
49. Balasubramanian S, Hurley LH, Neidle S. *Nat. Rev. Drug Discov.* 2011; 10:261–275. [PubMed: 21455236]
50. Qin Y, Hurley LH. *Biochimie.* 2008; 90:1149–1171. [PubMed: 18355457]
51. Ambrus A, Chen D, Dai JX, Bialis T, Jones RA, Yang DZ. *Nucleic Acids Res.* 2006; 34:2723–2735. [PubMed: 16714449]

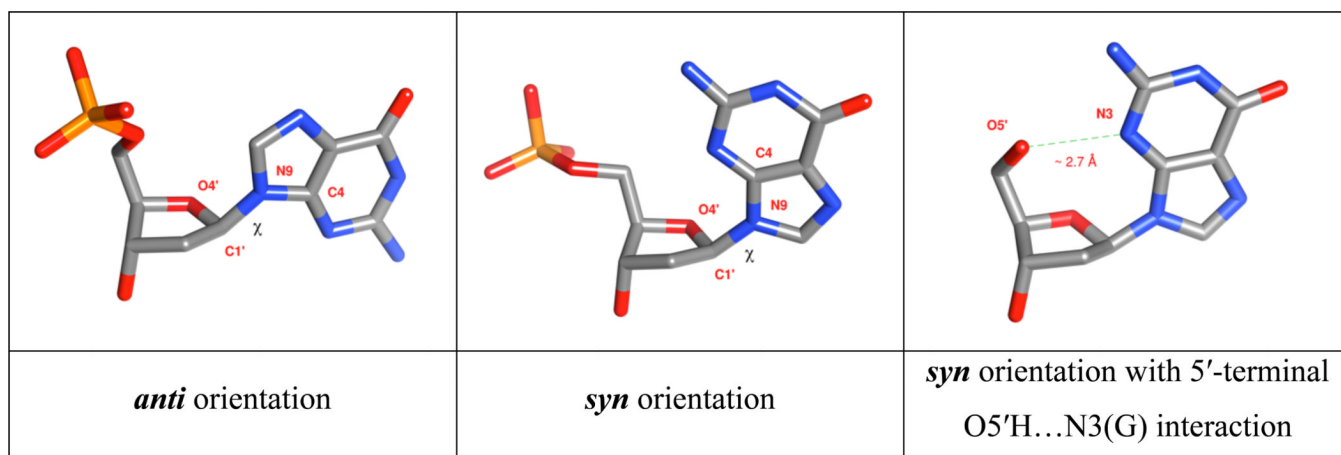


52. Crnugelj M, Sket P, Plavec J. *J. Am. Chem. Soc.* 2003; 125:7866–7871. [PubMed: 12823005]
53. Dai JX, Carver M, Yang DZ. *Biochimie.* 2008; 90:1172–1183. [PubMed: 18373984]
54. Phan AT, Kuryavyi V, Patel DJ. *Curr. Opin. Struct. Biol.* 2006; 16:288–298. [PubMed: 16714104]
55. Silva MW. *Chem.-Eur. J.* 2007; 13:9738–9745. [PubMed: 17972263]
56. Cang XH, Sponer J, Cheatham TE III. *J. Am. Chem. Soc.* 2011; 133:14270–14279. [PubMed: 21761922]
57. Collie GW, Haider SM, Neidle S, Parkinson GN. *Nucleic Acids Res.* 2010; 38:5569–5580. [PubMed: 20413582]
58. Reshetnikov RV, Sponer J, Rassokhina OI, Kopylov AM, Tsvetkov PO, Makarov AA, Golovin AV. *Nucleic Acids Res.* 2011; 39:9789–9802. [PubMed: 21893589]
59. Cavallari M, Calzolari A, Garbesi A, Di Felice R. *J. Phys. Chem. B.* 2006; 110:26337–26348. [PubMed: 17181293]
60. Fadrna E, Spackova N, Sarzynska J, Koca J, Orozco M, Cheatham TE, Kulinski T, Sponer J. *J. Chem. Theory Comput.* 2009; 5:2514–2530.
61. Haider SM, Patel JS, Poojari CS, Neidle S. *J. Mol. Biol.* 2010; 400:1078–1098. [PubMed: 20595043]
62. Hazel P, Parkinson GN, Neidle S. *Nucleic Acids Res.* 2006; 34:2117–2127. [PubMed: 16641317]
63. Pagano B, Mattia CA, Cavallo L, Uesugi S, Giancola C, Fraternali F. *J. Phys. Chem. B.* 2008; 112:12115–12123. [PubMed: 18767769]
64. Rueda M, Luque FJ, Orozco M. *J. Am. Chem. Soc.* 2006; 128:3608–3619. [PubMed: 16536534]
65. Akhshi P, Acton G, Wu G. *J. Phys. Chem. B.* 2012; 116:9363–9370. [PubMed: 22780684]
66. Cavallari M, Garbesi A, Di Felice R. *J. Phys. Chem. B.* 2009; 113:13152–13160. [PubMed: 19754127]
67. Haider S, Parkinson GN, Neidle S. *Biophys. J.* 2008; 95:296–311. [PubMed: 18375510]
68. Mashimo T, Yagi H, Sannohe Y, Rajendran A, Sugiyama H. *J. Am. Chem. Soc.* 2010; 132:14910–14918. [PubMed: 20882978]
69. Petraccone L, Spink C, Trent JO, Garbett NC, Mekmaysy CS, Giancola C, Chaires JB. *J. Am. Chem. Soc.* 2011; 133:20951–20961. [PubMed: 22082001]
70. Haider S, Neidle S. 2008; Vol. 608:17–37.
71. Han HY, Langley DR, Rangan A, Hurley LH. *J. Am. Chem. Soc.* 2001; 123:8902–8913. [PubMed: 11552797]
72. Petraccone L, Garbett NC, Chaires JB, Trent JO. *Biopolymers.* 2010; 93:533–548. [PubMed: 20095044]
73. Grimme S, Antony J, Ehrlich S, Krieg H. *J. Chem. Phys.* 2010; 132:154104–154122. [PubMed: 20423165]
74. Grimme S, Ehrlich S, Goerigk L. *J. Comput. Chem.* 2011; 32:1456–1465. [PubMed: 21370243]
75. Klamt A. *WIREs Comput. Mol. Sci.* 2011; 1:699–709.
76. Klamt A, Schuurmann G. *J. Chem. Soc.-Perkin Trans.* 1993; 2:799–805.
77. Svozil D, Kalina J, Omelka M, Schneider B. *Nucleic Acids Res.* 2008; 36:3690–3706. [PubMed: 18477633]
78. Phillips K, Dauter Z, Murchie AIH, Lilley DMJ, Luisi B. *J. Mol. Biol.* 1997; 273:171–182. [PubMed: 9367755]
79. Haider S, Parkinson GN, Neidle S. *J. Mol. Biol.* 2002; 320:189–200. [PubMed: 12079378]
80. Schultze P, Macaya RF, Feigon J. *J. Mol. Biol.* 1994; 235:1532–1547. [PubMed: 8107090]
81. Luu KN, Phan AT, Kuryavyi V, Lacroix L, Patel DJ. *J. Am. Chem. Soc.* 2006; 128:9963–9970. [PubMed: 16866556]
82. Hazel P, Parkinson GN, Neidle S. *J. Am. Chem. Soc.* 2006; 128:5480–5487. [PubMed: 16620121]
83. Clark GR, Pytel PD, Squire CJ. *Nucleic Acids Res.* 2012; 40:5731–5738. [PubMed: 22373921]
84. Ahlrichs R, Bar M, Haser M, Horn H, Kolmel C. *Chem. Phys. Lett.* 1989; 162:165–169.
85. Staroverov VN, Scuseria GE, Tao JM, Perdew JP. *J. Chem. Phys.* 2003; 119:12129–12137.
86. Zhao Y, Truhlar DG. *J. Phys. Chem. A.* 2005; 109:5656–5667. [PubMed: 16833898]

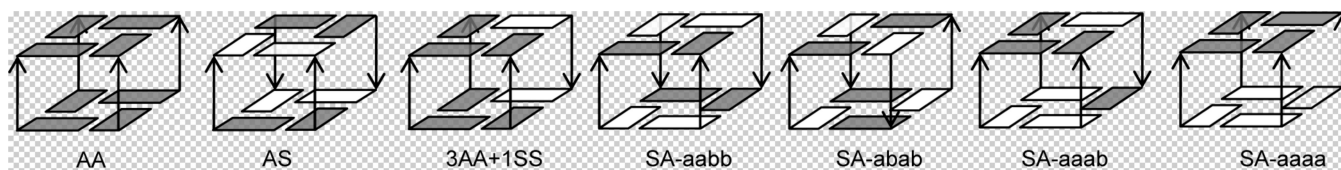
87. Schafer A, Huber C, Ahlrichs R. *J. Chem. Phys.* 1994; 100:5829–5835.
88. Weigend F, Ahlrichs R. *Phys. Chem. Chem. Phys.* 2005; 7:3297–3305. [PubMed: 16240044]
89. Eichkorn K, Treutler O, Ohm H, Haser M, Ahlrichs R. *Chem. Phys. Lett.* 1995; 240:283–289.
90. Eichkorn K, Weigend F, Treutler O, Ahlrichs R. *Theor. Chem. Acc.* 1997; 97:119–124.
91. Weigend F, Kohn A, Hattig C. *J. Chem. Phys.* 2002; 116:3175–3183.
92. Treutler O, Ahlrichs R. *J. Chem. Phys.* 1995; 102:346–354.
93. Goerigk L, Grimme S. *Phys. Chem. Chem. Phys.* 2011; 13:6670–6688. [PubMed: 21384027]
94. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. *J. Comput. Chem.* 2005; 26:1668–1688. [PubMed: 16200636]
95. Dang LX, Kollman PA. *J. Phys. Chem.* 1995; 99:55–58.
96. Tan CH, Yang LJ, Luo R. *J. Phys. Chem. B.* 2006; 110:18680–18687. [PubMed: 16970499]
97. Tan C, Tan YH, Luo R. *J. Phys. Chem. B.* 2007; 111:12263–12274. [PubMed: 17918880]
98. Hawkins GD, Cramer CJ, Truhlar DG. *Chem. Phys. Lett.* 1995; 246:122–129.
99. Hawkins GD, Cramer CJ, Truhlar DG. *J. Phys. Chem.* 1996; 100:19824–19839.
100. Liu DC, Nocedal J. *Math. Program.* 1989; 45:503–528.
101. Grimme S, Hujo W, Kirchner B. *Phys. Chem. Chem. Phys.* 2012; 14:4875–4883. [PubMed: 22378355]
102. Mladek A, Krepl M, Svozil D, Cech P, Otyepka M, Banas P, Zgarbova M, Jurecka P, Šponer J. *Phys. Chem. Chem. Phys.* 2013; 15:7295–7310. [PubMed: 23575975]
103. Kruse H, Grimme S. *J. Chem. Phys.* 2012; 136:154101–154116. [PubMed: 22519309]
104. Goerigk L, Kruse H, Grimme S. *Chem. Phys. Chem.* 2011; 12:3421–3433. [PubMed: 22113958]
105. Rudberg E. *J. Phys.-Condes. Matt.* 2012; 24:072202.
106. Cohen AJ, Mori-Sanchez P, Yang WT. *Chem. Rev.* 2012; 112:289–320. [PubMed: 22191548]
107. Antony J, Grimme S. *J. Comput. Chem.* 2012; 33:1730–1739. [PubMed: 22570225]
108. Grimme S. *Chem.-Eur. J.* 2012; 18:9955–9964. [PubMed: 22782805]
109. Wang Y, Patel DJ. *Structure.* 1993; 1:263–282. [PubMed: 8081740]
110. Šponer J, Gabb HA, Leszczynski J, Hobza P. *Biophys. J.* 1997; 73:76–87. [PubMed: 9199773]
111. Xu Y, Noguchi Y, Sugiyama H. *Bioorgan. Med. Chem.* 2006; 14:5584–5591.
112. Lim KW, Amrane S, Bouaziz S, Xu WX, Mu YG, Patel DJ, Luu KN, Phan AT. *J. Am. Chem. Soc.* 2009; 131:4301–4309. [PubMed: 19271707]
113. Parkinson GN, Lee MPH, Neidle S. *Nature.* 2002; 417:876–880. [PubMed: 12050675]
114. Gros J, Rosu F, Amrane S, De Cian A, Gabelica V, Lacroix L, Mergny JL. *Nucleic Acids Res.* 2007; 35:3064–3075. [PubMed: 17452368]
115. Mergny JL, De Cian A, Ghelab A, Sacca B, Lacroix L. *Nucleic Acids Res.* 2005; 33:81–94. [PubMed: 15642696]
116. Hud NV, Smith FW, Anet FAL, Feigon J. *Biochemistry.* 1996; 35:15383–15390. [PubMed: 8952490]
117. Smith FW, Feigon J. *Nature.* 1992; 356:164–168. [PubMed: 1545871]
118. Sket P, Virgilio A, Esposito V, Galeone A, Plavec J. *Nucleic Acids Res.* 2012; 40:11047–11057. [PubMed: 22977177]
119. Song JN, Ji CG, Zhang JZH. *Phys. Chem. Chem. Phys.* 2013; 15:3846–3854. [PubMed: 23399949]



**Figure 1.**  
2-quartet G-DNA stem in the 5'-anti-anti-3' (AA) arrangement.

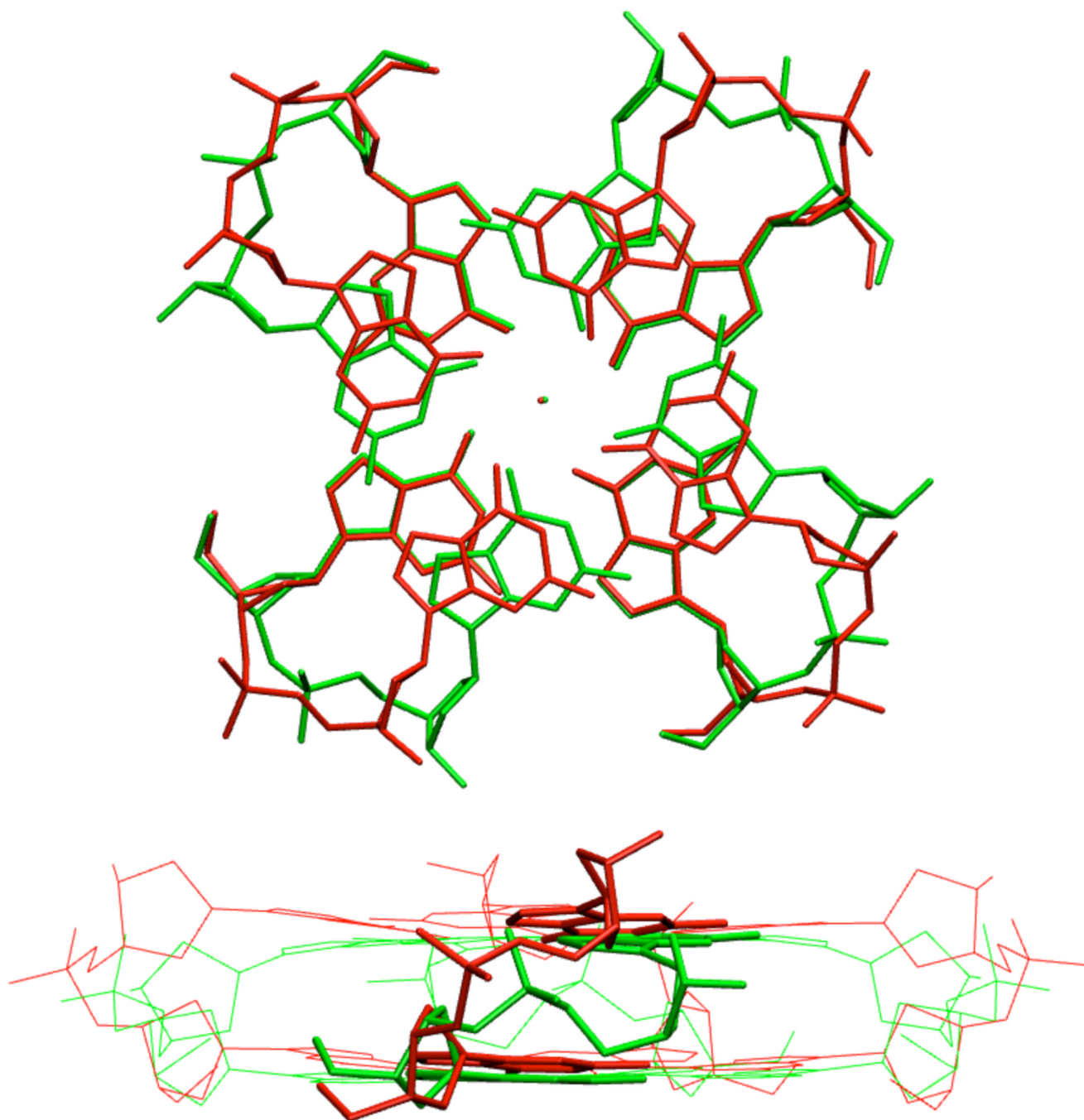
**Figure 2.**

Two orientations of the glycosidic torsion angle  $\chi$ : *anti* (left) and *syn* (center and right). The  $\chi$  torsion angle is defined as O4'-C1'-N9-C4 and O4'-C1'-N1-C2 for purines and pyrimidines, respectively. The O5'H...N3(G) hydrogen bond typical for 5'-terminal *syn* guanosines in the G-DNA is depicted right.<sup>34</sup>

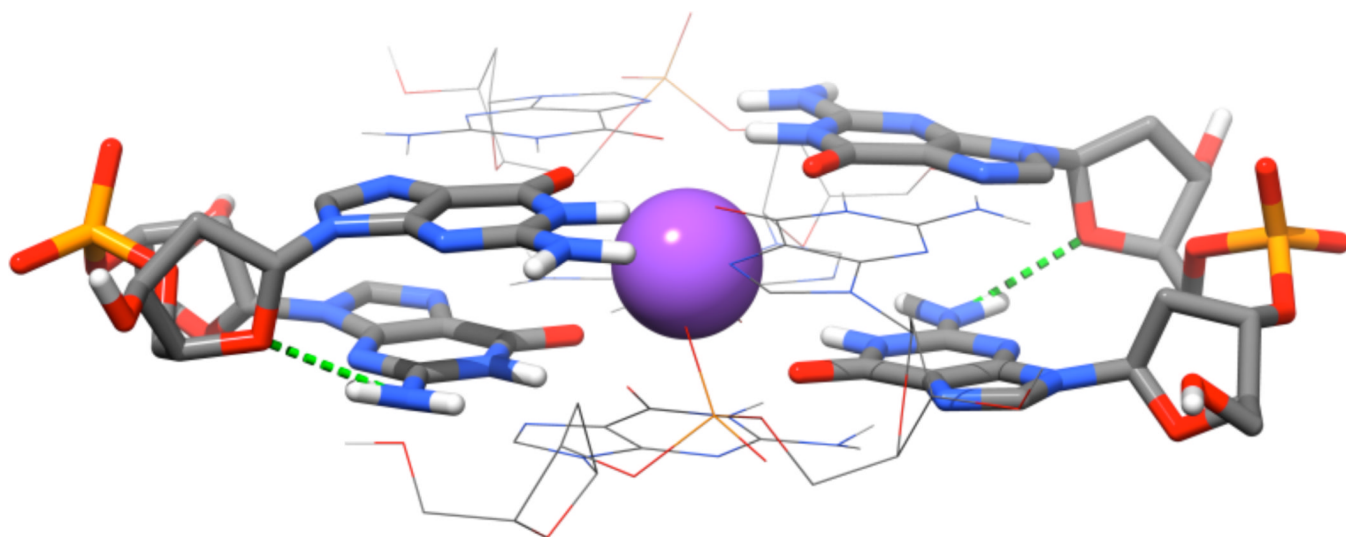


**Figure 3.**

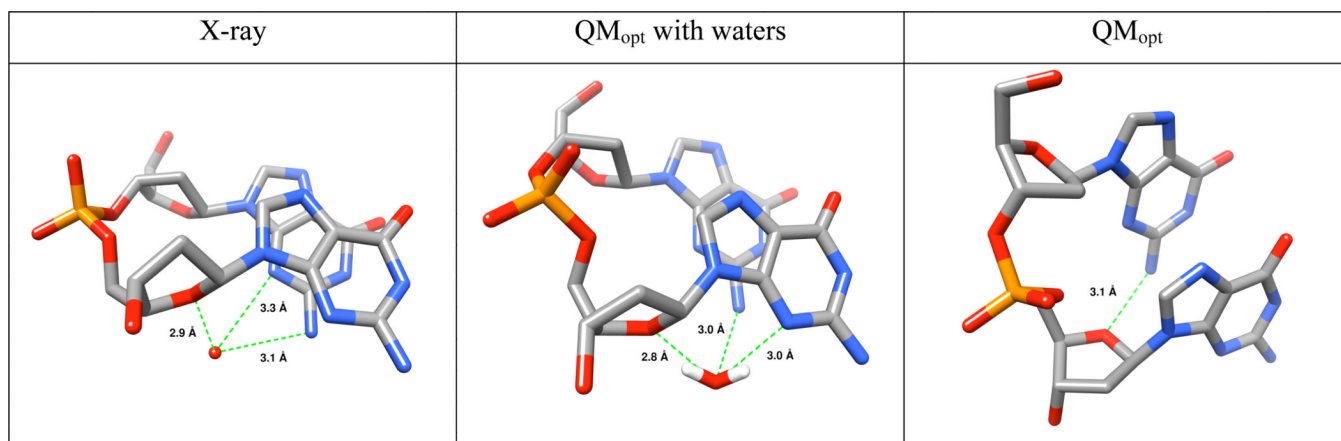
Two-quartet G-DNA stems used in our computations. From left to right: AA, AS, 3AA + 1SS, SA-aabb, SA-abab, SA-aaab and SA-aaaa structures. Each stem consists of four 5'-GG-3' dinucleotide steps (strands); S and A stand for *syn* and *anti* guanines. The notation 'a' and 'b' is used to distinguish SA stems with different relative 5'-3' orientation of the adjacent strands around the structure; a is used for strands oriented upward in our Figure while b for strands oriented downward; c.f., the names of the structures with the relative orientation of the arrows showing the 5'-3' directionality of the strands. White rectangle marks *syn* and grey *anti* nucleosides. The channel cation ( $K^+$ ) is not shown. The first six arrangements were considered in the preceding MD simulation study,<sup>34</sup> the seventh arrangement has been added in the course of this study.



**Figure 4.**  
QM<sub>opt</sub> (green) and X-ray (red) AA structures.



**Figure 5.** Detailed view on the non-native N2(G) ... O4 (G+1) H-bonds in the QM<sub>opt</sub> structure of the AA model.

**Figure 6.**

Explicit water molecules prevent non-native N2(G)...O4 (G+1) H-bonds in the QM<sub>opt</sub> AA structure. Left - the X-ray water position (red ball, only oxygen is visible), which is essentially identical in all observed instances. Center - example of the water position after the QM optimization; note that the water molecule has somewhat relocated. Right - the non-native H-bond in absence of the explicit water molecule.



Table 1

Relative energies (kcal/mol) of 2-quartet G-DNA stems. Solvent screening is included by the method used in the optimization.<sup>a</sup>

Method/ Stem	AA	SA-aabb	SA-abab	SA-abab-2	SA- aaab	3AA+ISS	AS	SA-aaaa
QM <sub>opt</sub> <sup>b</sup>	0	+0.4 (+7.4)	+9.8	-4.4 (+2.6)	+3.9	+6.8	+16.7	-21.5 (+6.5)
QM MM <sub>opt</sub> <sup>c</sup>	0	-27.0 (-20.3)	-14.8	-31.3 (-25.6)	-22.9	+2.1	-2.7	-53.9(-20.9)
<b>Difference<sup>d</sup></b>	<b>0</b>	<b>-27.4 (-27.7)</b>	<b>-24.6</b>	<b>-26.9 (-28.2)</b>	<b>-26.8</b>	<b>-4.7</b>	<b>-19.4</b>	<b>-32.4 (-27.4)</b>
QM MM <sub>opt</sub> OL4 <sup>e</sup>	0	-34.0	-17.7	-39.6	-31.4	+3.7	-4.9	-58.4
Cang <i>et al.</i> <sup>34f</sup>	0	-14.6(-20.2)	-14.4(-20.0)	-14.4(-20.0)	-13.4(-19.0)	+4.6(+6.0)	+ ~14	N/A

<sup>a</sup>Intramolecular O5 H...N3(G) H-bonds in 5'-terminal *syn* guanosines are present in the following structures: SA-aabb, one interaction; SA-abab-2, one interaction; SA-aaaa, four interactions. All these interactions exist in the original X-ray structures. There is one O5 H...N2(G) H-bond in the SA-abab system. The H-bonds are consistently present/absent in both QM<sub>opt</sub> and QM MM<sub>opt</sub> structures and thus should not systematically affect the comparison of QM and MM relative energies.

<sup>b</sup>Optimization using the DFT-D3 COSMO approach starting from the initial structure. The estimates in parentheses were obtained by deducting 7 kcal/mol per each *syn* nucleotide terminal O5 H...N3(G) H-bond (see the text).

<sup>c</sup>Optimization using the parmbsc0 force field and PB method starting from the QM<sub>opt</sub> structure. Values in parentheses show the energies with elimination of the terminal O5 H...N3(G) H-bonds (see the text).

<sup>d</sup>Difference between MM and QM relative energies. The estimates in parentheses were obtained when considering the deductions of the energies of the terminal O5 H...N3(G) H-bonds.

<sup>e</sup>Optimization using the parmbsc0+ OL4 force field and PB method starting from the QM<sub>opt</sub> structure.

<sup>f</sup>MD MM-PBSA free energy computations compiled from the Table 1 by Cang *et al.*<sup>34</sup> with -4.4 (-3.0 in parenthesis) kcal/mol correction per each 5'-terminal *syn* base to deduct contributions of all O5 H...N3(G) intramolecular H-bonds (for more details see the original study and the Method section).

Table 2

Sum of the relative energies of the four 5'-GpG-3' dinucleotides of the 2-quartet G-DNA stems evaluated as gas phase single points (kcal/mol).<sup>a, b, c, d, e</sup>

Method/ Stem	AA	SA-aabb	SA-abab	SA-abab-2	SA- aaab	3AA+1S S	AS	SA-aaaa
QM <sub>opt</sub> <sup>b</sup>	0	+16.6 (+26.6)	+28.7	+4.6 (+14.6)	+19.1	+17.0	-6.3	-18.7 (+21.3)
QM MM <sub>opt</sub> <sup>c</sup>	0	-9.9 (+5.3)	+5.7	-25.1 (-15.6)	-5.0	+17.9	-12.4	-60.3 (+1.0)
<b>Difference<sup>d</sup></b>	0	-26.5 (-21.3)	-23.0	-29.7 (-30.2)	-24.1	+0.9	-6.1	-41.6 (-20.3)
<b>T2-T1<sup>e</sup></b>		+0.9 (+6.4)	+1.6	-2.8 (-2.0)	+2.7	+5.6	+13.3	-9.2 (+7.1)

<sup>a</sup>For energies of all the individual GpG dinucleotides, see Supporting Information Table S2. The values in parentheses are without the 5' terminal O5 H...N3(G) H-bonds (cf. Table 1).

<sup>b</sup>Optimization using the DFT-D3 COSMO approach starting from the initial structure.

<sup>c</sup>Optimization using the parmbsc0 force field and PB method starting from the QM<sub>opt</sub> structure.

<sup>d</sup>Difference between MM and QM relative energies. The estimates in parentheses were obtained when considering the deductions of the energies of the terminal O5 H...N3(G) H-bonds.

<sup>e</sup>Difference between predictions from Tables 2 and 1 (lines "Difference").

Table 3

Relative energies (kcal/mol) calculated for the MM<sub>opt</sub> structures.

Method/ Stem	AA	SA-aabb	SA-abab	SA-abab-2	SA- aabb	3AA+IS S	AS	SA-aaaa
QM on MM <sub>opt</sub> <sup>a</sup>	0	+8.3	+40.4	+7.0	+11.3	+17.0	+14.5	+27.8
MM <sub>opt</sub> <sup>b</sup>	0	-2.0 (-7.7) <sup>c</sup>	+27.7(+4.8) <sup>d</sup>	-6.2 (-16.2) <sup>c</sup>	+1.7	+15.6	+17.6	-1.6 (-34.1) <sup>c</sup>
<b>Difference 1. 2-1. 1<sup>e</sup></b>	<b>0</b>	<b>-10.3</b>	<b>-12.7</b>	<b>-13.2</b>	<b>-9.6</b>	<b>-1.4</b>	<b>+3.1</b>	<b>-29.4</b>
QM, Table 3-Table 1 <sup>f</sup>	0	+7.9	+30.6	+11.4	+7.4	+10.2	-2.2	+49.3
MM, Table 3-Table 1 <sup>g</sup>	0	+25.0	+42.5	+25.1	+24.6	+13.5	+20.3	+52.3
<b>Difference 1. 5 - 1. 4<sup>h</sup></b>	<b>0</b>	<b>+17.1</b>	<b>+11.9</b>	<b>+13.7</b>	<b>+17.2</b>	<b>+3.3</b>	<b>+22.5</b>	<b>+3.0</b>

<sup>a</sup>DFT-D3 COSMO single point energy on the MM<sub>opt</sub> structure.

<sup>b</sup>MM<sub>opt</sub> using the parmbsc0 force field and PB method starting from the initial (experimental) structure; the optimization procedure essentially keeps the initial structure (see Methods).

<sup>c</sup>Data in parenthesis show alternative energies when the O5 H...N3 H-bonds are formed by adjusting position of the 5'-terminal hydrogen compared to the automated Leap AMBER assignment (see the text) for those nucleotides where there is such an H-bond in the X-ray experimental structure. This allows formation of the O5 H...N3(G) H-bonds.

<sup>d</sup>The initial structure has one perturbed quartet which is not repaired by the MM optimization. This explains the high relative energy. The number in parentheses is for alternative structure where we established the O5 H...N3(G) H-bond by initial reorientation of the whole O5 H group.

<sup>e</sup>Difference between lines 2 and 1, i.e., difference between the MM and QM relative energies. To be compared with line 3 of Table 1.

<sup>f</sup>Change of the QM relative energies of the stems predicted by data from the Table 3 compared to the Table 1 (line 1, Table 3 minus line 1, Table 1).

<sup>g</sup>Change of the MM relative energies of the stems predicted by data from the Table 3 compared to the Table 1 (line 2, Table 3 minus line 2, Table 1).

<sup>h</sup>Change of the predicted MM - QM energy difference when Table 3 is used instead of Table 1. It can be calculated either as difference between the preceding lines 4 and 5 in this Table or as difference between lines 3 from Tables 3 and 1).

**Table 4**

Single point energies (kcal/mol) for QM<sub>opt</sub> structures (with no further relaxation) obtained using different methods.

Method/stem	AA	SA-aabb	SA-abab	SA-abab2	SA-aaab	3AA+ISS	AS	SA-aaaa
PB <sup>a</sup>	0	-24.6	-10.0	-37.2	-22.5	-0.4	+1.7	-57.8
GB	0	-16.0	-3.2	-28.3	-15.1	+5.1	+6.2	-37.7
QM <sup>b</sup>	0	+0.4	+9.8	-4.4	+3.9	+6.8	+16.7	-21.5
gas phase QM	0	+27.3	+26.1	+23.1	+30.8	+12.2	+44.6	-7.6
gas phase force field	0	+12.8	+16.6	+2.9	+17.7	+10.6	+53.6	-38.2

<sup>a</sup>The PB data in this Table marginally differ from the Table 1 QM<sub>MMopt</sub> data since additional PB relaxation has been done in the Table 1. Note that the purpose of computations in Tables 1 and 4 was slightly different. The Table 1 QM<sub>MMopt</sub> calculations were done to obtain MM structures and energies using the QM<sub>opt</sub> geometry as the start, while the purpose of Table 4 was to derive energies by all methods on entirely unprocessed QM<sub>opt</sub> structures.

<sup>b</sup>The reference QM<sub>opt</sub> data from the Table 1.