THE UNIQUE TAUTOMERIC AND RECOGNITION PROPERTIES OF THIO-THYMINES?

Ignacio Faustino[1], Anna Aviño[2], Ivan Marchán[1], F.Javier Luque[3], Ramon Eritja[2] and Modesto Orozco[1,4,5]*

The tautomeric and recognition properties of thymine, 2- and 4-thiothymines 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 sulphur 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 thiothymines induces only mild changes in DNA structure, stability and fidelity. Despite the fact that mismatching can largely stabilize minor tautomeric forms, thiothymines are found in the canonical thio-form irrespective of the paired base. Our theoretical results, confirmed by new experimental studies, describe the complete tautomeric and recognition characteristics of thiothymines and demonstrate that both 2-thio and 4-thiothymine are excellent molecules to introduce special chemical properties in modified DNA.

---

[1] 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
[2] Institute of Research in Biomedicine, IQAC-CSIC, CIBER-BBN Networking Centre on Bioengineering, Biomaterials and Nanomedicine, Baldiri Reixac 15, Barcelona 08028. Spain
[3] Departament de Fisicoquímica and Institut de Biomedicina (IBUB), Facultat de Farmàcia, Universitat de Barcelona, Avgda Diagonal 643, Barcelona 08028, Spain.
[4] National Institute of Bioinformatics, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain
[5] Departament de Bioquímica, Facultat de Biología, Avgda Diagonal 647, Barcelona 08028, Spain.

- Corresponding author modesto@mmb.pcb.ub.es

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 non-standard bases are in fact spontaneously found in DNA due to 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, non-standard nucleobases can change the functionality of DNA, opening a wide range of biotechnological and biomedical applications[10].

Non-standard 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 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-thiothymine (^{2}S and ^{4}S; Figure 1) are relevant as the reactivity of the thiocarbonyl group can be exploited in nucleophilic and photo-cross-linking reactions in DNA[15]. Moreover, the possibility to photoactivate thiothymines in conditions where coding nucleobases are unaltered confer potential interest as pro-drugs in phototherapy of psoriasis and skin cancer[16]. Early hybridization experiments suggested that insertion of thiothymidine 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 behaviour has been challenged by more recent analysis, which suggest

that one of the thio-derivatives: ^{4}S shows mutagenic properties due to stabilization of G-mismatching[18]. These findings raise doubts about the prevalence of keto/thio tautomers in thiopyrimidines, suggesting a partner-dependent tautomeric preference (see below), as noted previously by other modified nucleboases[13,19]. This behaviour, if confirmed, will open interesting possibilities for the design of promiscuous nucleobases, though will seriously challenge the real impact of thiothymines in the chemistry and phototherapy of nucleic acids.

A detailed comparison of the tautomeric preferences of thymine and thiothymines in different environments, including DNA, is presented. Attention is paid on the impact of thiothymines 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 thiothymines. Thus, the presence of thymine surrogates does not significantly affect the DNA structure and only introduce marginal destabilization in the duplex, without altering the DNA pairing fidelity and therefore lacking mismatch-related mutagenic properties. Overall, our results demonstrate that thiothymines 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 o ^{4}S. For this purpose all possible tautomers of the N1-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 these corrections we obtain our *best estimate* that is expected to be within a few tenths of kcal/mol from the real value.

To check the goodness of classical force fields to represent canonical and non-canonical pairings of thymine and thiothymines 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 thio-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 free energies of those tautomers (ΔG_{hyd}^{A} and ΔG_{hyd}^{B}) 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 ($\Delta\Delta G_{hyd}^{B-A}$) 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 $\Delta\Delta G_{hyd}^{B-A}$ 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 ($\Delta G_{taut}^{A->B}(sol)$; Eq. 1) was derived by adding the best estimate of the gas phase tautomerization free energy ($\Delta G_{taut}^{A->B}(gas)$) to the differential hydration free energy between tautomers ($\Delta\Delta G_{hyd}^{B-A}$).

$$\Delta G_{taut}^{A->B}(sol) = \Delta G_{taut}^{A->B}(gas) + \Delta\Delta G_{hyd}^{A->B} \quad (1)$$

DNA simulations. In order to examines the impact of thiothymines 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 thiothymine-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(thio), G·^{4}S(thio), 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.

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(thio)/amino tautomers of T and thiothymines 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). For this purpose mutations keto/thio-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 ($\Delta\Delta G^{A->B}_{solvDNA}$), which can be combined with the intrinsic gas phase tautomerization ($\Delta G^{A->B}_{taut}(gas)$) to obtain the tautomerization free energy in the duplex ($\Delta G^{A->B}_{taut}(DNA)$; Eq. 2).

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

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

Note that in case that Eq. 2 reveals that for one of the pairings the canonical keto/thio 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. 3).

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

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

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

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

where A and A' stand for the canonical tautomer (keto/thio) of thymine (T) and thiothymine (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 thiothymines were obtained from different sources: i) equilibrium parameters for bonded terms were taken from B3LYP optimized geometries, ii) van der Waals parameters for sulphur were taken from a previous work[19], and iii) atomic charges for every tautomer were fitted using the standard RESP procedure[33]. Note that 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. 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 supplementary Figure S1). Accordingly, the largest source of errors is expected to be statistical noise and uncertainties related to incomplete samplings in the simulations.

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. Due to 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 thiothymines. Several sequences (including those studied in previous works) were considered, but detailed results are reported only for 5'-GCAATGGAXCCTCTA-3'/3'-CGTTACCTYGGAGAT-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-thiothymidine and 4-thiothymidine 2-cyanoethyl phosphoramidites were from commercial sources. 2-Thiothymidine (*Glen Research, USA*) was protected with the toluoyl group. 4-Thiothymidine (*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 (50ºC, 1 hr). Three different protocols were considered for deprotection of oligonucleotides carrying ^{4}S: i) concentrated ammonia (50 ºC, 1 hr), ii) 1M 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) in acetonitrile (room temperature, 3 hr) followed by 50 mM NaSH in concentrated ammonia (room temperature, 24 hr; supplier's recommendation), and iii) 50 mM NaSH in concentrated ammonia (room temperature, 24 hr). 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 acetate (pH 6.5), and solvent B: 70% acetonitrile in 100 mM triethylammonium acetate pH 6.5. Columns: Nucleosil 120C18 (10 μm), 200 x 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 90ºC, allowed to cool slowly to room temperature, and stored at 4ºC. 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 1ºC/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 Tm values was estimated at ±0.7ºC.

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/thio 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_H4t) 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 thiothymine analogues are almost identical and strongly favour Watson-Crick pairing, which is in contrast with the behaviour 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/thio forms are predicted to be the dominant species for T, ^{2}S and ^{4}S in water. In particular, the keto/thio→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 Supplementary Figure S2), and to the tautomerization free energy in aqueous solution, for which errors are not expected to be larger than 1 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/thio tautomeric form), were performed to study the structural impact of introducing: i) G·T mismatches and ii) thiothymines into DNA duplexes. All trajectories were stable and the helical B-like structures were always well preserved during the entire simulation (Supplementary Figure S3). Thus, the DNA appeared as a very robust structure, and able to accommodate local distortions 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 (Supplementary Table S1) are within the thermal noise of the reference simulation. G·T and G·S mismatches (for standard keto/thio tautomers) lead to wobble pairings associated with local distortions clearly visible in roll and twist parameters (see examples in supplementary Figure S4). 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 steps (see Figure 4). No major structural differences are found in d(G·S) pairs compared to the d(G·T) ones, though the presence of sulphur atom at position 2 decreases the percentage of time in which bases are hydrogen-bonded (see Figure 4 and below).

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 (Supplementary Figure S3) and displaying the expected triple Watson-Crick-like hydrogen-bond scheme at the substitution site (see Figures 4 and 5) 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/thio form) are drastically reduced when X is in the enol/thiol form (Figures S4 and 4). As previously noted by others[18], these findings suggest that the DNA environment favours 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).

In order 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/thio forms, ii) the change in stability due to the T→^{2}S and T→^{4}S mutations (in their standard keto/thio 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/thio 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 thiothymines.

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/thio 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 Supplementary Figure S5 for examples). The independent estimates of the reversible work associated to each mutation also agreed well (see Supplementary Figure S6 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 supplementary Figure 5) 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 values reported by Karran[17b], but deviates more than 2 kcal/mol from more recent estimates[18]. In order 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 thio 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 Supplementary Figure S6). 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 reference 18. Inspection of the different interaction terms (see Table 4) suggests that the T→S mutation (for keto/thio tautomers) is small due to the balance between the lost of hydrogen-bond energy (sulphur is a poorer H-bond acceptor than oxygen) and the gain in stacking energy related to the stronger dispersive interactions of the sulphur.

The preceding results demonstrate that statistical errors in the simulations protocols cannot be responsible of the sizeable 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) ant thio (^{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 revert 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 disfavoured by 3 kcal/mol, and such a difference amounts to 1.1-2.4 kcal/mol for the thio→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 sizeable 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 thiothymines, but even in the case of A→G transductions keto/thio 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 that ii) thiothymines, 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 thiothymines 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 thiothymines paired with both G and A were synthesized using solid-phase 2-cyanoethylphosphoramidite chemistry. The required synthons to incorporate 2-thiothymidine and 4-thiothymidine into oligodeoxynucleotides were obtained from commercial sources[17a,36]. Since 4-thiothymidine 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 hrs or at 55ºC for 1 hr to minimize decomposition of the 4-

thiothymidine residue. Also we studied the use of 50 mM NaSH or the removal of the 2-cyanoethyl group with 1M 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) solution in acetonitrile at room temperature for 3 hrs. In all cases the desired oliognucleotide was obtained as the major compound. The resulting oligonucleotides were purified by HPLC and gave the expected 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 (ΔG = -13.7 Kcal/mol), and the T·G pair (ΔG = -11.2 Kcal/mol) is the most stable of the T mismatches (ΔG = -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 references 37,38). The most stable pairing for ^{2}S also involves A (ΔG = -13.4 kcal/mol), followed by the mismatch with G (ΔG = -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 (ΔG = -12.5 kcal/mol), followed by the mismatch with G (ΔG = -10.4 kcal/mol). As suggested by theoretical calculations, ^{4}S does not change the pairing scheme of DNA, and 4-thiothymine 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 thiothymines, 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 thiothymines. It is found that keto/thio tautomers are the prevalent ones for both thymine and thiothymine in gas phase, the situation being mostly unaltered in water. When inserted into DNA duplex, thiothymines induces small changes in structure and stability. A guanine paired to thymine or thiothymine help to stabilize minor enol/thiol forms, but this effect is not large enough as to change the tautomeric preferences of the pyrimidines considered here. Overall, our theoretical results, confirmed by experimental measures, support a quite classical behaviour for thiothymines, which are incorporated in 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.

ACKNOWLEDGMENTS

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.

REFERENCES

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, 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.
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 (Amst)* **2002**, *1*, 275-286.
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. *Chemical Physics Letters* **1998**, *294*, 45-48.
21. Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W. *Chem. Phys. Lett.* **1999**, *302*, 437-446.
22. Gaussian 03, Revision C.02, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; 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.
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. Chem. Soc.* **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. *Annual review of physical chemistry* **2000**, *51*, 435-471, (d) Cheatham, T. E.; Kollman, P. A. *Ann. 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.
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.; III, T. E. C.; 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.
35. Kobayashi, R. *J. Phys. Chem. A* **1998**, *102*, 10813-10817.
36. Kuimelis, R. G.; Nambiar, K. P. *Nucleic Acids Res.* **1994**, *22*, 1429-1436.
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. (Oxf)* **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. *Helvetica Chimica 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.
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.

LEGENDS TO FIGURES

Figure 1. Schematic representation of thymine, 2-thiothymine and 4-thiothymine and their tautomeric species.

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

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

Figure 4. Variation along the last 10 ns of the number of hydrogen bonds at the mutation site for different pairings.

Figure 5. 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.

LEGENDS TO SUPPLEMENTARY FIGURES

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. 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. 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. 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 S6. 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.

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*

Table 1. Tautomerization free energy (kcal/mol) for tautomers of N1-methyl-thymine (T), 2-thiothymine (^{2}S) and 4-thiothymine (^{4}S) in the gas phase determined at different levels of theory. The best estimate (quality CBS/CCSD(T)) is displayed in italics. All values are referred to the canonical keto/amino or thio/amino tautomers (see Figure 1 for nomenclature).

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

Table 2. Relative (to the respective canonical tautomers) hydration free energy of the different tautomers of N1-methyl-thymine (T), 2-thiothymine (^{2}S) and 4-thiothymine (^{4}S) determined from MST-SCRF and MD/TI simulations. 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.

Mutation	Central triad (**X**)	Comp. base	ΔΔG(stab)	Exp. Data (literature)[#]
A → G	A**X**G	T	1.5 ± 0.2	1.7 [*] / 1.7
	A**X**T	T	1.6 ± 0.2	1.7 [&] / 2.4
A → G	A**X**G	^{2}S	2.2 ± 0.2	1.7 [*] / 3.4
	A**X**T	^{2}S	2.1 ± 0.2	--
A → G	*A**X**G*	*^{4}S*	*1.7 ± 0.1*	*-0.5 [*] / -0.7*
	A**X**T	^{4}S	1.5 ± 0.2	1.9 [&] / 2.6

Table 3. Change in the free energy (kcal/mol) associated to the A→G mutation in the two sequences considered here (identified by the central triad). Values were computed/measured with the complementary pyrimidine equal to thymine (T), 2-thiothymine (^{2}S) and 4-thiothymine (^{4}S). Standard errors are also shown. *The case of large discrepancy with the literature is in italics.*

[*] Data from reference 18.

[&] Data from reference 17b.

[#] Values after the slash refers to estimates obtained from a linear regression (see ref. 40) with melting temperatures reported in the corresponding paper.

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	*AXG*	-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

Table 4. Hydrogen-bond interaction (in the central pair) and stacking (both intra- and inter-strand for the central triad) energies for the substitution site in the two different sequences used in our simulations. All values are in kcal/mol.

Mutation	Central triad (**X**)	Comp. base	ΔΔG(stab)	Exp. Data (literature)[#]
T→^{4}S	A**X**G	A	-0.3 ± 0.2	0.4 [*] / 0.8
	A**X**T	A	0.4 ± 0.1	0.4 [&] / 0.8
T→^{4}S	*AXG*	*G*	*0.3 ± 0.2*	*-1.8 [*] / -2.2*
	A**X**T	G	-0.3 ± 0.1	0.4 [&] / 1.1
T→^{2}S	A**X**G	A	-0.5 ± 0.2	0.1 [*] / 0.0
	A**X**T	A	-0.7 ± 0.2	-0.9 [$] / -0.9
T→^{2}S	A**X**G	G	0.5 ± 0.2	0.1 [*] / 1.1
	A**X**T	G	0.4 ± 0.2	0.5 [$] / 0.5

Table 5. Free energy change associated to the substitution of thymine by thiothymine (keto and thio tautomers) in the two sequences considered here (identified by the central triad). 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.*

[*] Data from reference 18.

[&] Data from reference 17b.

$ Data from Haynes's group on LNA, reference 39.

[#] Values after the slash refers to estimates obtained from a linear regression (see ref. 40) with melting temperatures reported in the corresponding paper.

Mutation	Central triad (**X**)	ΔG(solv)	ΔG(int)	ΔΔG(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

Table 6. Change in stability of the duplex due to the tautomeric change from keto/thio forms to the enol/thiol species for T and ^{4}S paired to G (ΔΔG(stab)). 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 QM level (see Table 1). All values are in kcal/mol.

Basepair	ΔH (kcal/mol)	ΔS (cal/K.mol)	ΔG(kcal/mol)	Tm (ºC)
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 7. Thermodynamic parameters of double-stranded DNA to single-stranded DNA transition. 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.

(A)

Parameter	T·A	T·G	^{2}S·A	^{2}S·G	^{4}S·A	^{4}S·G	4S$_{thiol}$·G	T(e)·G
Rise	3.44 3.41	3.42 3.38	3.37 3.38	3.38 3.42	3.35 3.36	3.34 3.43	3.45 3.47	3.42 3.37
Roll	5.22 5.05	5.77 4.52	4.68 4.62	5.10 4.97	5.29 4.48	5.13 5.17	3.64 4.15	4.71 4.25
Twist	32.04 32.13	32.33 31.95	32.25 32.42	31.59 32.47	31.68 32.50	32.57 31.96	32.67 32.22	32.50 32.45
Inclination	9.66 9.54	11.40 9.39	8.91 8.64	11.09 9.41	10.31 8.69	10.31 10.06	7.04 8.06	9.03 8.39
Opening	0.04 0.26	1.21 0.47	-0.49 -0.40	-0.19 -0.90	1.17 1.37	0.95 0.66	0.58 0.73	-0.70 -0.49
Prop twist	-11.15 -11.53	-11.82 -11.28	-11.82 -11.25	-11.06 -12.00	-10.67 -11.64	-11.69 -11.33	-10.94 -11.26	-11.27 -10.76
Phase	129.20 129.45	130.36 129.67	127.07 129.49	131.80 128.82	123.14 128.47	128.25 129.91	133.62 130.36	132.05 129.87
mG width	12.94 12.81	13.11 12.74	13.19 12.82	13.34 13.28	13.69 12.73	13.20 12.74	12.36 12.59	13.25 12.96
MG width	19.48 19.25	19.06 19.54	19.13 19.10	19.33 19.16	20.41 20.12	19.02 19.81	19.01 20.01	19.25 19.77

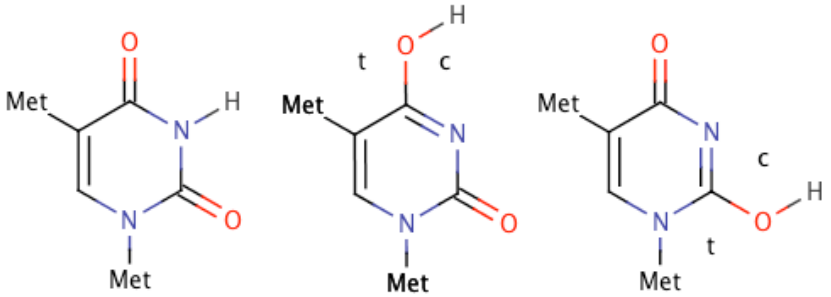
(B)

Parameter	T·A	T·G	^{2}S·A	^{2}S·G	^{4}S·A	^{4}S·G	^{4}S(t)·G	T(e)·G
Rise	3.41 3.32	3.15 3.26	3.31 3.28	3.08 3.07	3.40 3.42	2.98 3.24	3.63 3.50	3.33 3.44
Roll	11.65 3.12	14.34 6.77	11.40 2.94	13.63 6.00	13.37 -0.12	13.15 6.65	9.80 2.75	12.64 4.63
Twist	30.57 33.67	20.19 23.07	30.23 33.73	16.77 21.31	28.99 37.28	19.19 22.17	33.15 28.94	29.98 33.20
Inclination	21.03 5.55	36.30 16.57	21.00 5.18	37.14 15.93	24.75 0.34	34.68 16.64	16.83 5.91	23.21 8.11
Opening	0.26 0.67	12.82 3.54	-2.14 -1.12	2.83 -5.89	10.18 9.71	10.02 4.39	7.98 7.24	-4.12 -4.09
Prop twist	-12.08 -14.40	-15.66 -15.72	-11.66 -15.77	-12.59 -15.32	-15.13 -20.04	-12.77 -14.78	-18.19 -16.84	-11.69 -14.14
Phase	108.36 119.38	84.83 89.06	97.71 106.25	84.65 84.50	89.52 112.88	78.69 86.68	109.51 96.06	102.19 114.77
mG width	13.00 12.60	13.20 12.33	13.47 12.56	13.89 13.37	13.88 12.50	13.65 12.37	11.61 12.05	13.46 12.34
MG width	19.29 19.49	18.83 19.26	18.78 19.10	19.27 19.45	20.36 20.86	18.98 19.80	18.92 20.34	18.66 19.18

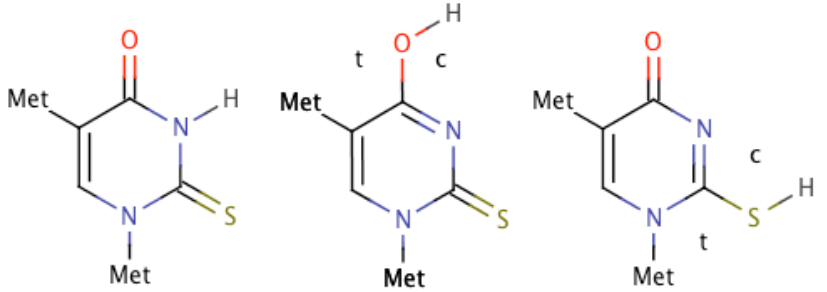
Table S1. DNA descriptors for the two sequences considered here with the different pairings. In each cell top values correspond to the duplex with the central A**X**G triad and down values to those for the duplex with the central A**X**T triad. (A) Values for the entire DNA (excluding ends). (B) Values for the central triads. Angular parameters are in degree, translational parameters are in Å.

Figure 1

Thymine (T)



2-Thiothymine (^{2}S)



4-Thiothymine (^{4}S)

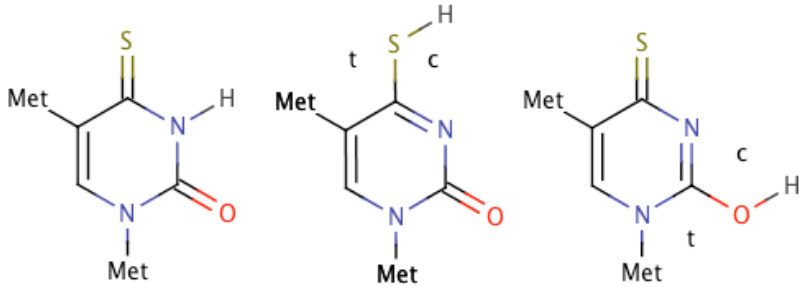
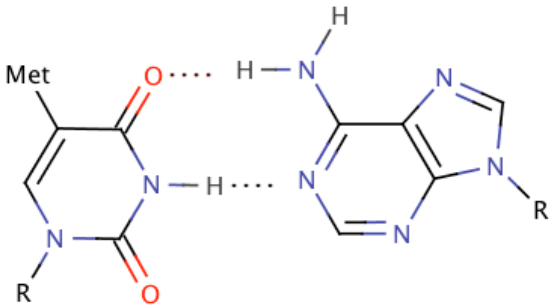
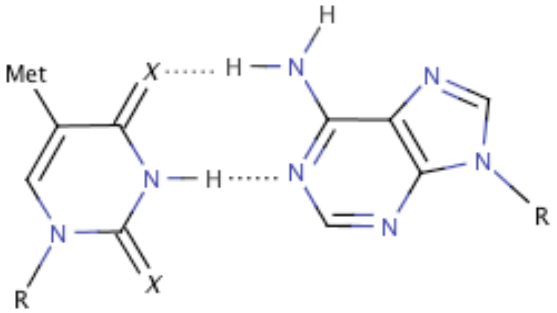


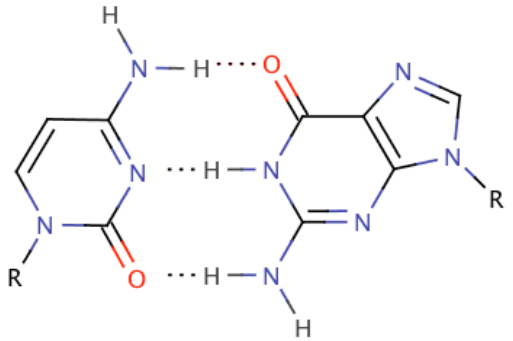
Figure 2



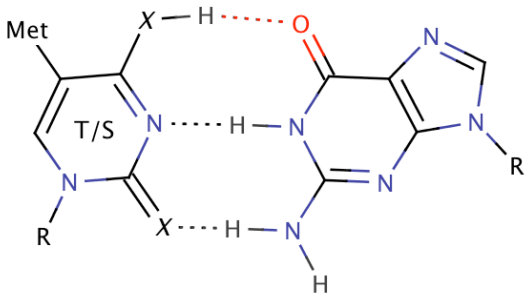
Canonical A·T



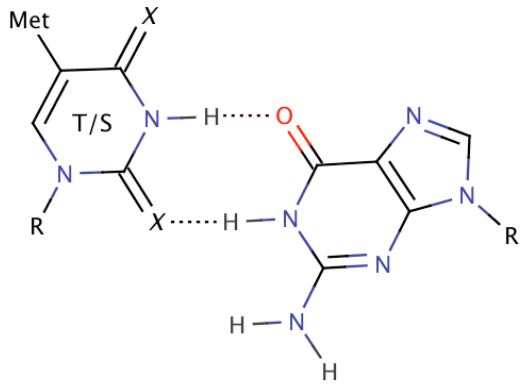
WC-like A·S



Canonical G·C

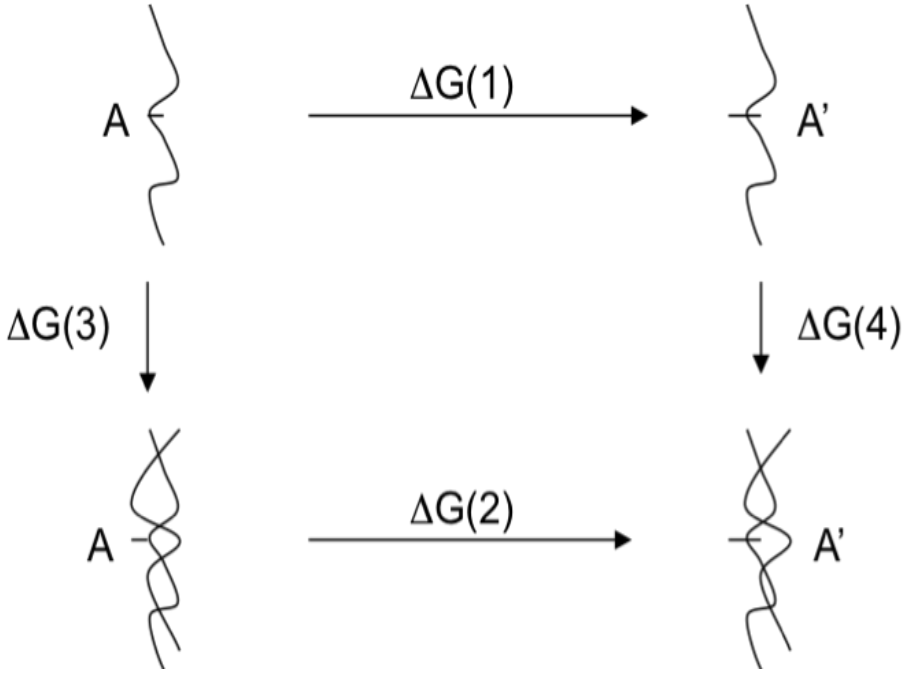


WC-like G·T/S(enol/thiol)



Wobble G·T/S

Figure 3



$$\Delta\Delta G_{stab}(X \rightarrow Y) = \Delta G(4) - \Delta G(3) = \Delta G(2) - \Delta G(1)$$

Figure 4

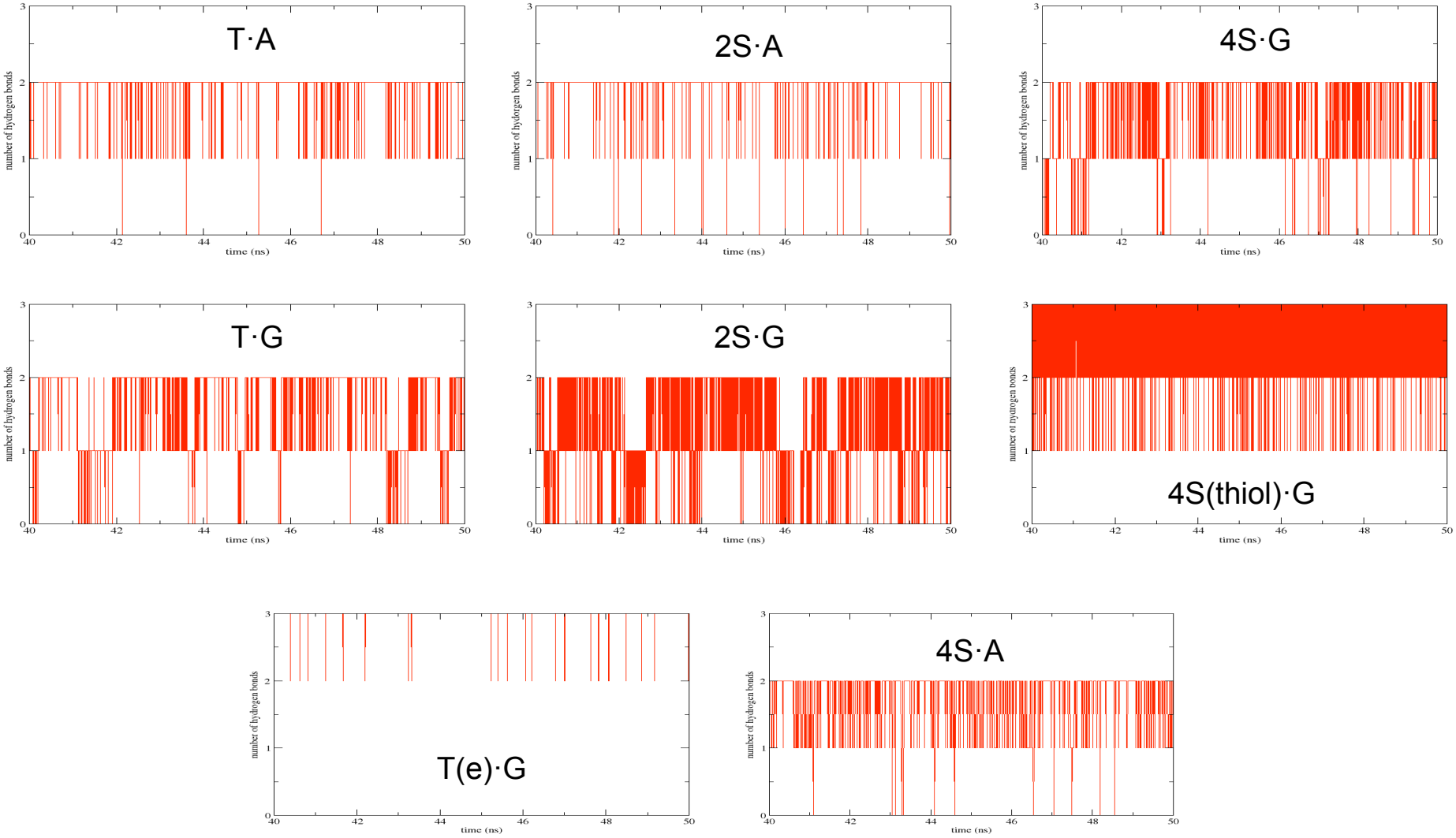
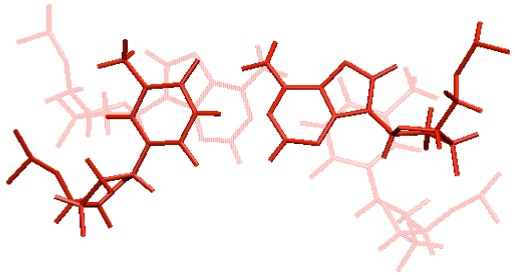
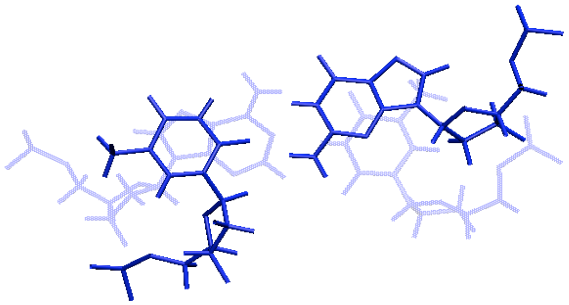

T·A
2S·A
4S·G
T·G
2S·G
4S(thiol)·G
T(e)·G
4S·A
number of hydrogen bonds
time (ns)

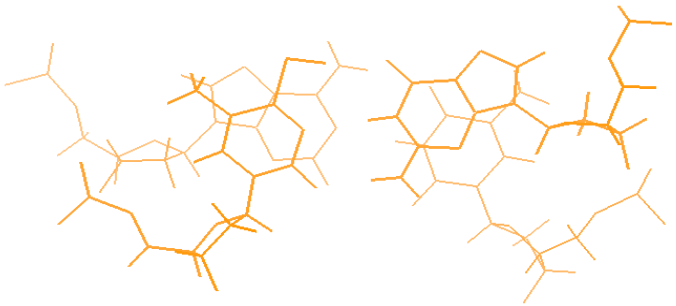
Figure 5



Canonical A·T or A·S pair



Wobble G·T or G·S pair



WC-like G·T(enol) or G·S(thiol) pair

Figure S1

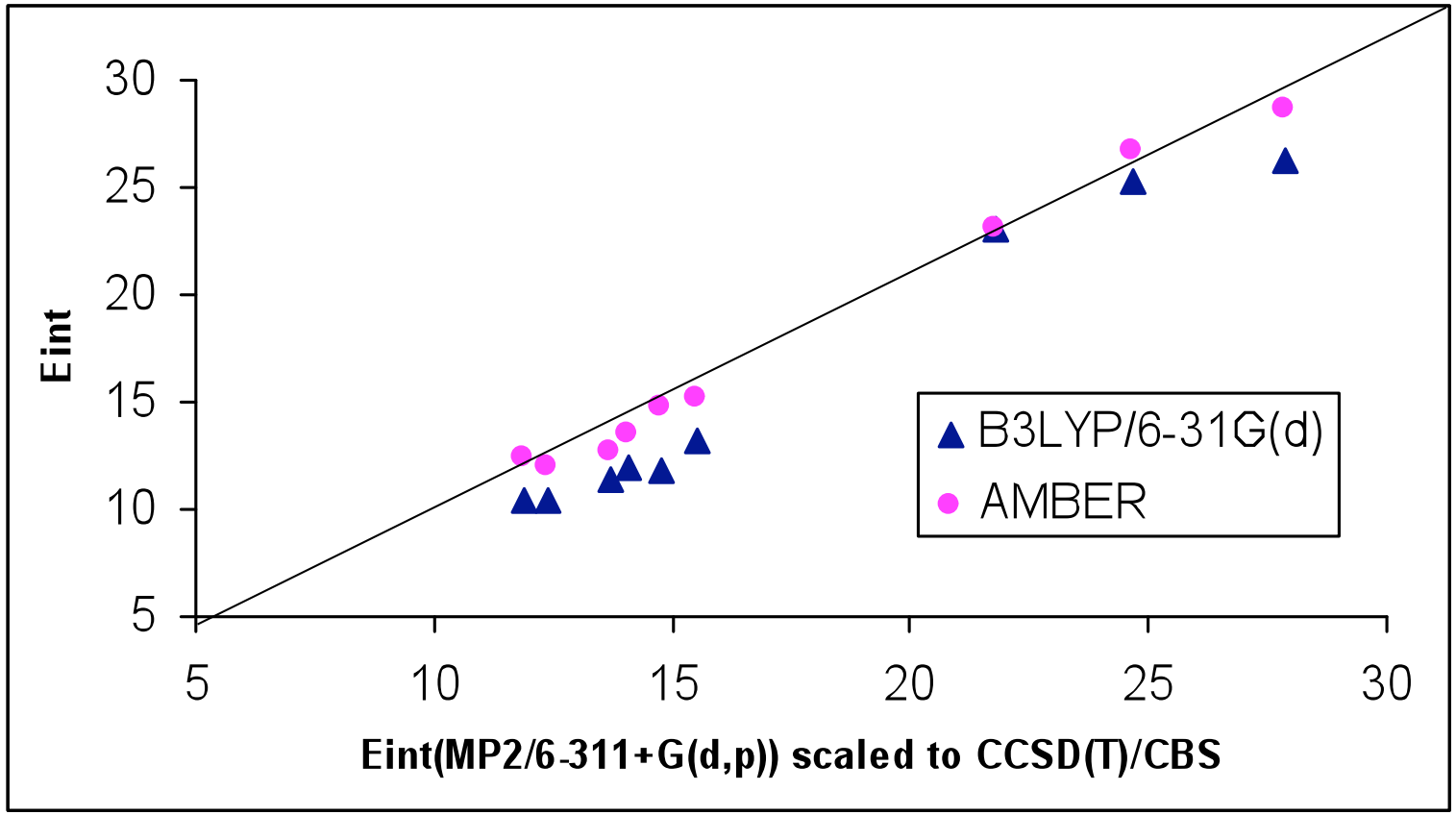

Eint
30
25
20
15
10
5
5
10
15
20
25
30
Eint(MP2/6-311+G(d,p)) scaled to CCSD(T)/CBS
B3LYP/6-31G(d)
AMBER

Figure S2

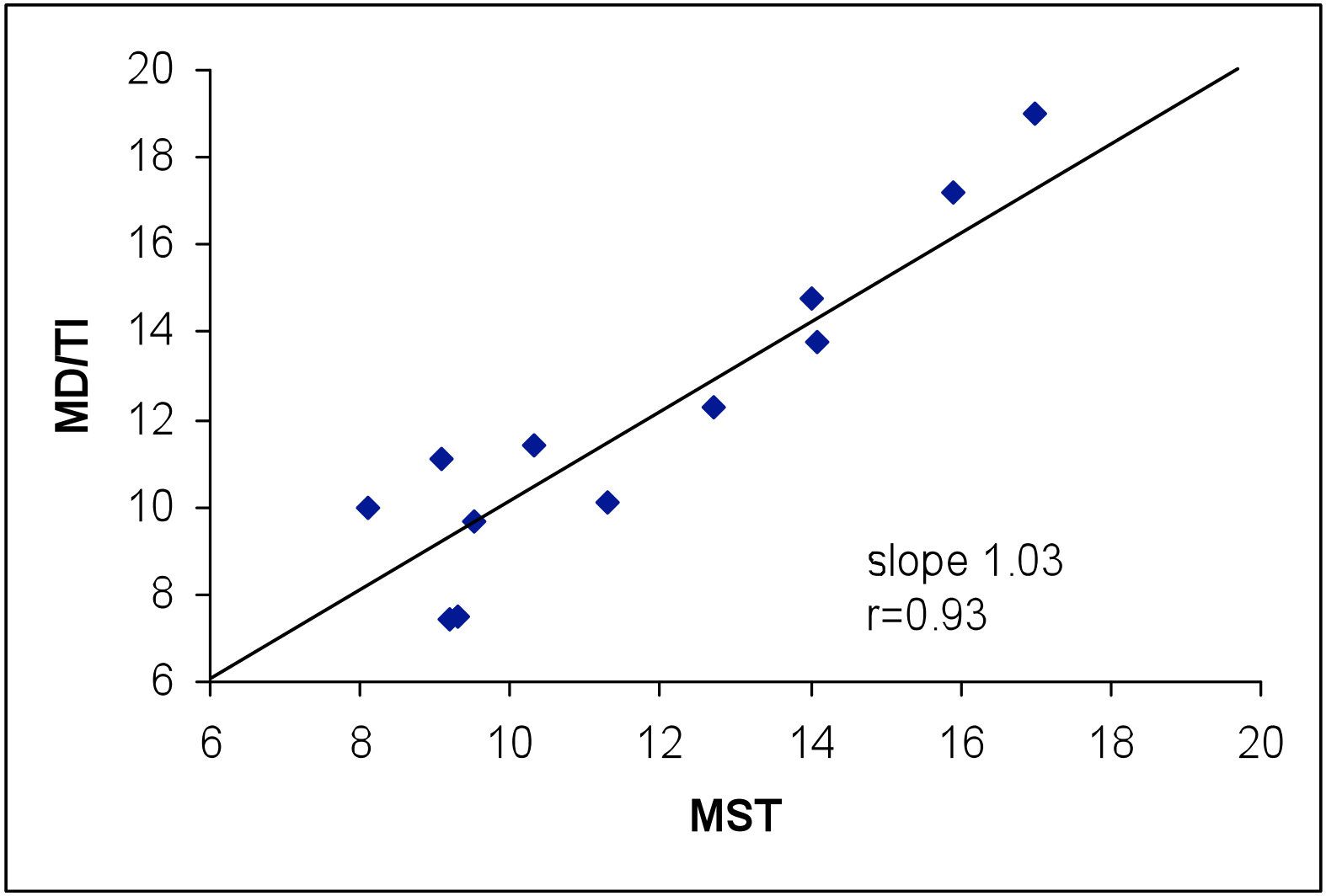


Figure S3

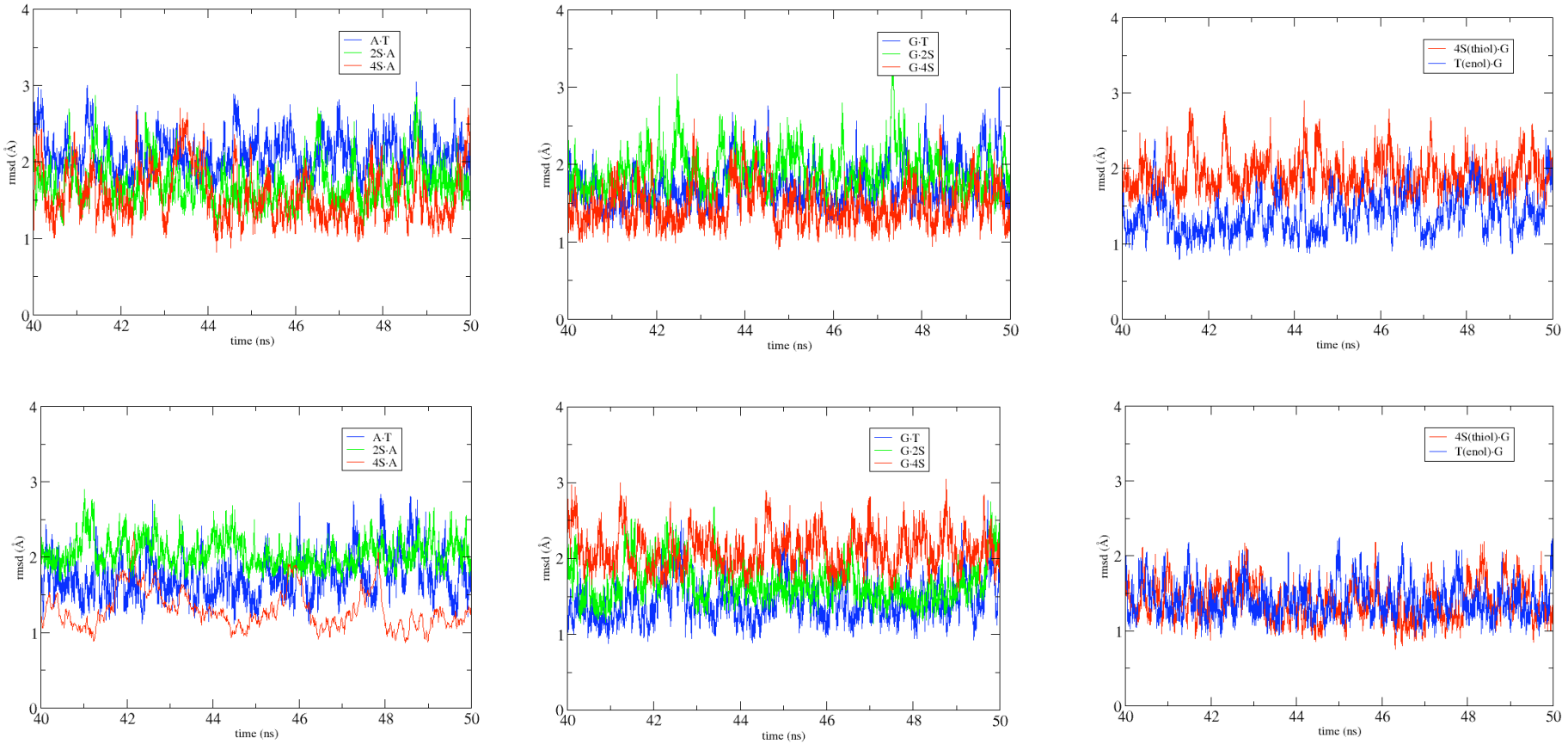


Figure S4

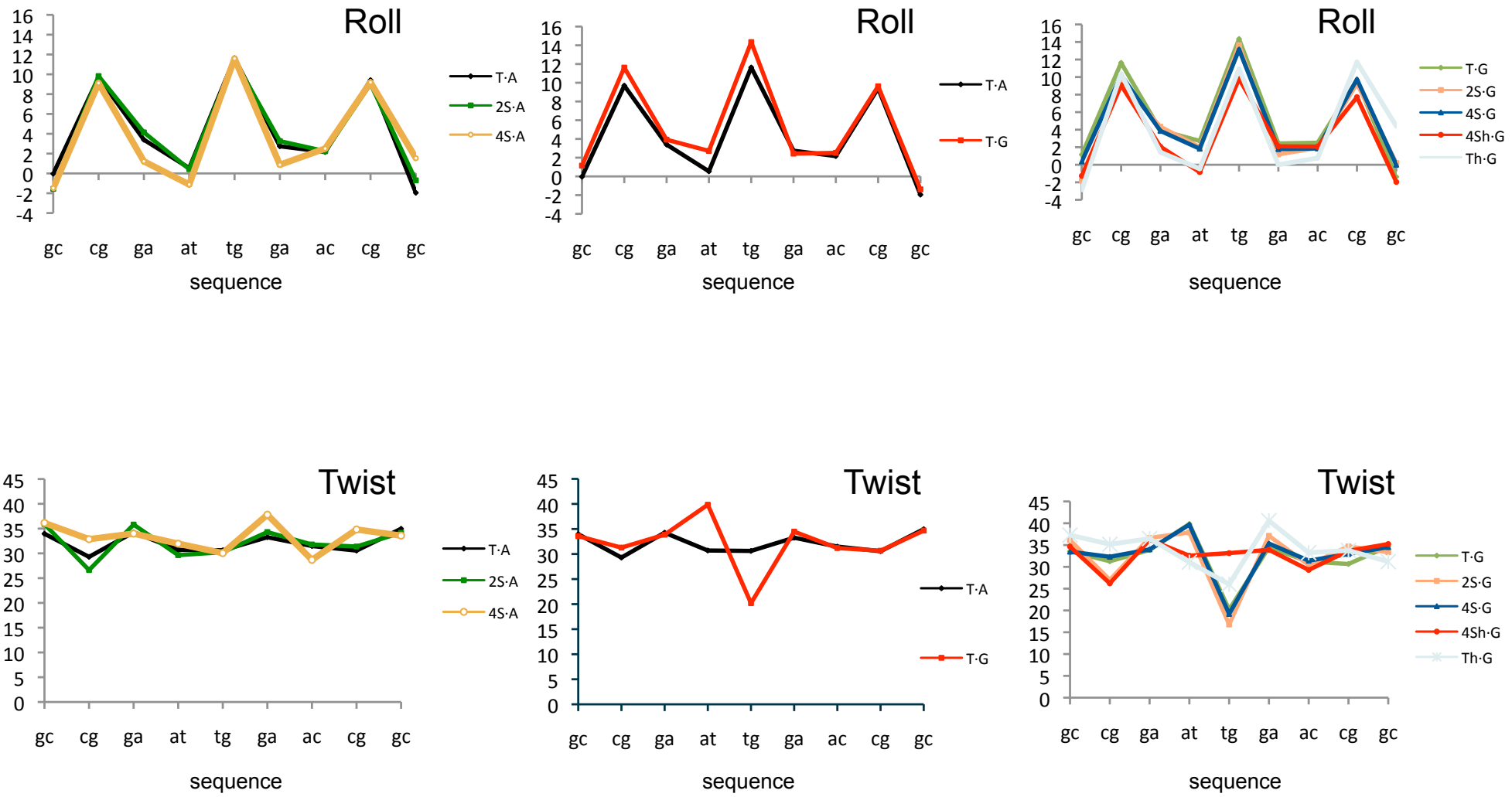
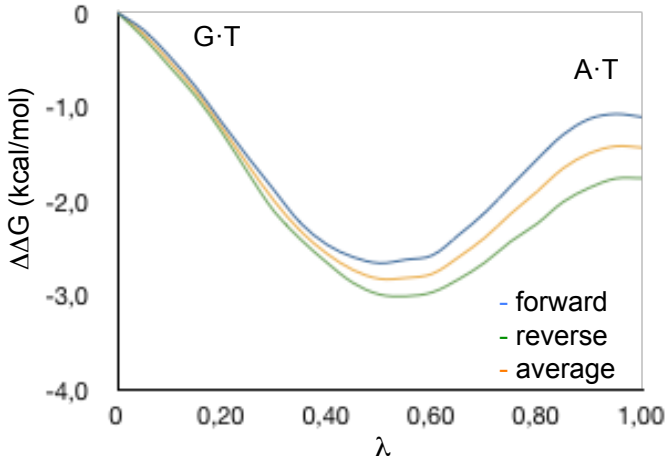
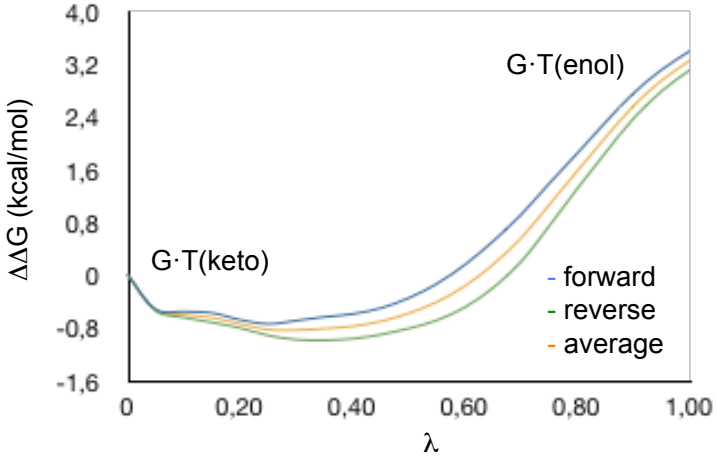

Roll
Twist
T·A
2S·A
4S·A
T·G
2S·G
4S·G
4Sh·G
Th·G
gc cg ga at tg ga ac cg gc
sequence

Figure S5


G·T
A·T
- forward
- reverse
- average
ΔΔG (kcal/mol)
λ
0
-1,0
-2,0
-3,0
-4,0
0
0,20
0,40
0,60
0,80
1,00


G·T(enol)
G·T(keto)
- forward
- reverse
- average
ΔΔG (kcal/mol)
λ
4,0
3,2
2,4
1,6
0,8
0
-0,8
-1,6
0
0,20
0,40
0,60
0,80
1,00

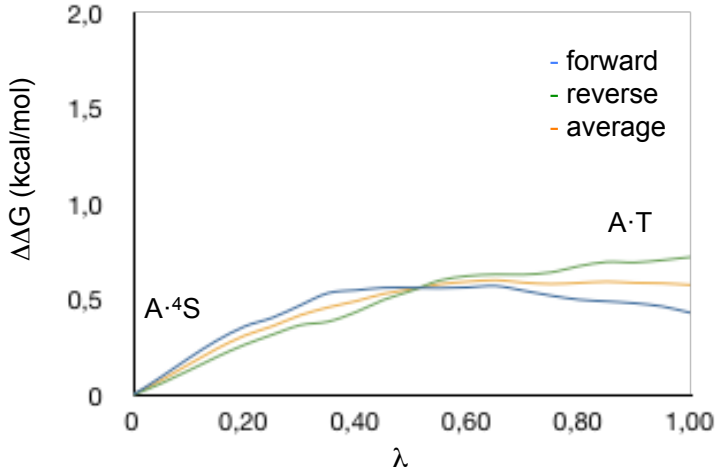
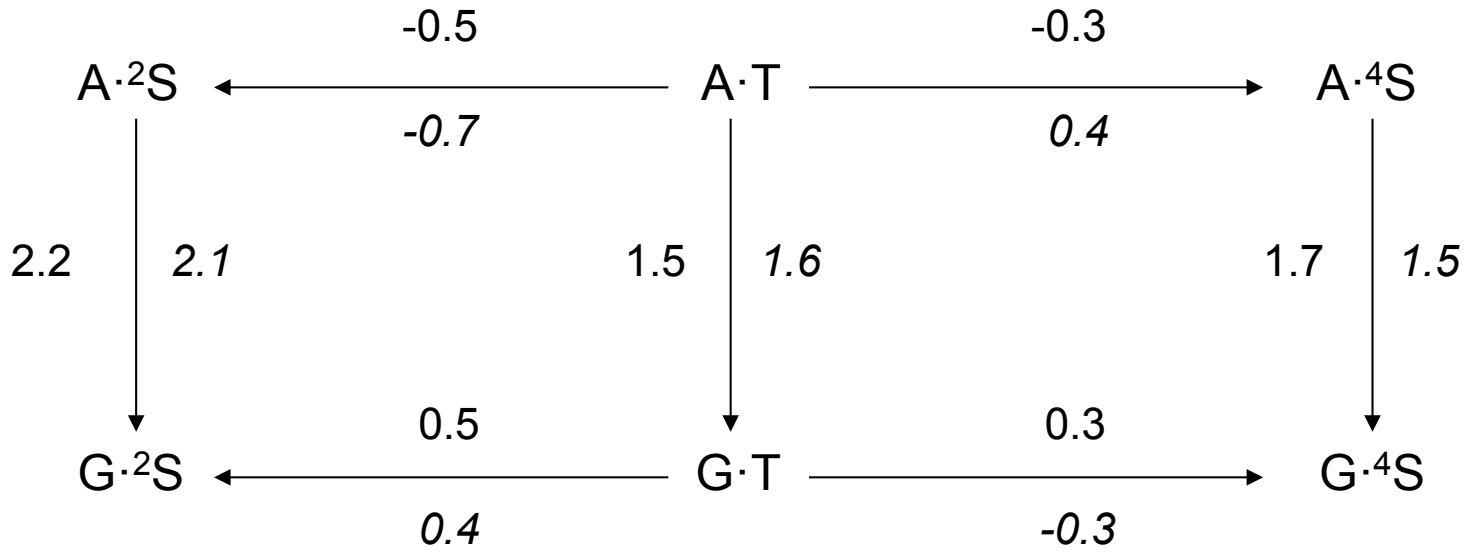

- forward
- reverse
- average
A·T
A·⁴S
ΔΔG (kcal/mol)
λ
2,0
1,5
1,0
0,5
0
0
0,20
0,40
0,60
0,80
1,00

Figure S6



(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 S1. 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.