PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

ARTICLE TYPE www. rsc.org/xxxxxx | XXXXXXXX

DNA spontaneous mutation and its role in the evolution of GC-content: assessing the impact of the genetic sequence

José P. Cerón-Carrasco[*,a] **and Denis Jacquemin**[b,c]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 20XX
DOI: 10.1039/b000000x

The structure of DNA is not constantly at its equilibrium point but evolves with time. It is generally accepted that evolution induces a decrease of the guanine–cytosine (GC) content and a concomitant increase of the adenine–thymine (AT) ratio through a biased GC→AT mutation process. Unfortunately, the mechanism behind this natural alteration of the stored genetic information is not fully understood. Here, we use a hybrid QM:QM' approach to assess the link between one of the sources of the spontaneous mutation, the so-called G*C* rare tautomers that arise from a double proton exchange between the basis, and the evolution of the GC-content. Our simulations indicate that the G*C* mutation is mainly accumulated in GC-rich regions rather than randomly spread, and consequently the GC→AT error tends to locate in coding fragments. That specific preference is indirectly induced by the base pairs confining the mutated point, as they tune the structure of the first hydration-shell that solvates the reactive base pair undergoing the tautomerisation. The reorganisation of the explicit water molecules eventually modifies the energy barriers as well as the stability of the genetic error during the process.

1 Introduction

Extraordinary advances obtained during the last decade have uncoded the secret of the human genome, increasing our knowledge regarding the biochemical reactions that govern the life.[1,2] Indeed, it is know understood that characteristic double helix architecture of DNA proposed by Watson and Crick[3] is strengthened by the hydrogen-bonds (H-bonds) and the π-stacking interactions established between guanine–cytosine (GC) and adenine–thymine (AT) base pairs.[4] The resulting non-covalent network is tight enough to maintain the two strands bounded together while still allowing their opening during the replication process.[5] The stored genetic information is correctly transmitted from one generation to the next if both GC and AT pairs are always preserved in their canonical Watson and Crick forms. However, if other structures are present during DNA's unwind, an error (mutation) might appear in the code.[6] One of the most plausible sources for the mutation of DNA is the so-called rare tautomers, the products arising from the proton transfer (PT) reactions between base pairs, as originally suggested by Löwdin fifty years ago.[7] More specifically, Löwdin used a simple chemical model to propose the PT-related tautomeric equilibria in AT and GC as one of the driving forces behind the spontaneous mutations. That mechanism has been extensively studied by several theoretical groups because rare tautomers can hardly be detected by experimental means.[8–25] It has been demonstrated that only non-canonical GC structures can lead to permanent mutations,[8,9] and one can therefore refine Löwdin's hypothesis: tautomeric equilibria can yield the rare tautomeric form of GC, hereafter denoted as G*C* (Figure 1).

In a recent work, Wang, Schaefer-III and co-workers have shown that the genetic sequence affects the PT in damaged DNA by adding $H^{\bullet}$, H^{+} and H^{-} entities at several positions in the GC base pairs.[26] These authors built up a series of hydrogenated-GC to subsequently explore the induced tautomeric equilibria. Their simulations revealed that the PT mechanisms depend on the sequence in these hydrogenated-DNA structures.[26] There is however an important unanswered question: does the genetic sequence impacts on the stability of the spontaneous mutation? In other words, is there a relationship between the sequence and the GC→G*C* conversion rate in undamaged-DNA? We address this question here, aiming to assess if PT reactions take part in the evolution. As schematised in Figure 1, the non-canonical G*C* form would yield single brand errors (G*T and C*A) in the first generation and the complete base pairs-swap (G*C*→AT) in the granddaughter strands. In that process, the initial G*T and C*A mispairs appear due to the reverse interbase H-bond pattern induced by the mutation: $O_6\cdots H_4$–N_4/N_1–$H_1\cdots N_3$ for the canonical GC, O_6–$H_4\cdots N_4/N_1\cdots H_1$–$N_3$ for the G*C* mutation. The biological impact of such mechanism can therefore be analysed according to the energetic profile of the

[a] *Departemento de Química Física, Universidad de Murcia, 30100 Murcia, Spain.; Tel: +34 868 88 7434; E-mail: jpceron@um.es.*
[b] *CEISAM, UMR CNRS 6230, BP 92208, Université de Nantes, 2, Rue de la Houssinière, 44322 Nantes, Cedex 3, France.*
[c] *Institut Universitaire de France, 103 bd St Michel, 75005 Paris Cedex 5, France.*

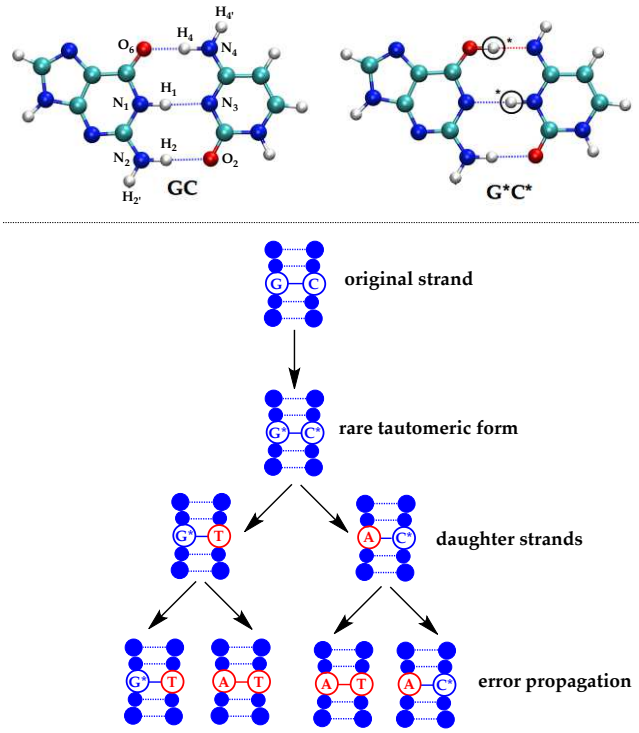


Fig. 1 Top: chemical structures of the canonical GC base pairs compared to its G*C* rare tautomeric form, including atomic numbering. The transferred protons are circled and marked with asterisks. Bottom: logical tree for genetic errors induced by G*C*. Adapted with permission from Acc. Chem. Res., **47**, 2467 (2014) Copyright of American Chemical Society.

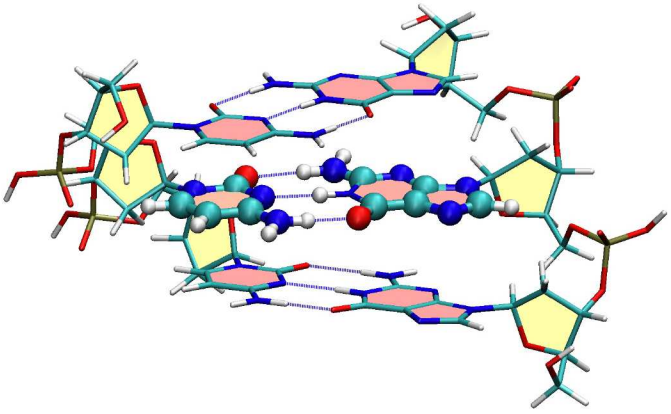
GC→G*C* equilibrium. Based on the DNA's reproduction period and base-pair opening times, Florián and Leszczyński[8] estimated that the G*C* mutation could lead to a permanent error if the energetic barrier for the forward proton transfer does not exceed ca. 28 kcal.mol^{-1} while the barrier for the reverse transfer should be larger than ca. 3 kcal.mol^{-1}. Consequently, the investigation of such process could at least partly explain the GC→AT conversion[27] and might help to clarify if the GC-content has reached equilibrium or, on the contrary, is still decreasing.[28]

In this contribution, the possible effects of genetic sequence on the GC→G*C* conversion have been systematically investigated by combining two quantum methods (QM), e.g., density functional theory (DFT) and semiempirical approaches, in a hybrid QM:QM' scheme. A three base pairs DNA fragment is used, as this model system is the minimum fragment allowing the accurate simulation of the π-stacking interactions in DNA.[29–31] The relative energies of the canonical and rare tautomeric forms as well as the transition states that connect them have been computed by a solvation model coupling discrete and implicit approaches, a suitable representation of the hydration effect in biological conditions as it allows to explore the possible catalytic role played by the first hydration shell (water-assisted tautomerisation)[32] while long-range solvent effects are described by a continuum model. Our calculations provide new insights on the impact of genetic sequence in the spontaneous mutation of DNA.

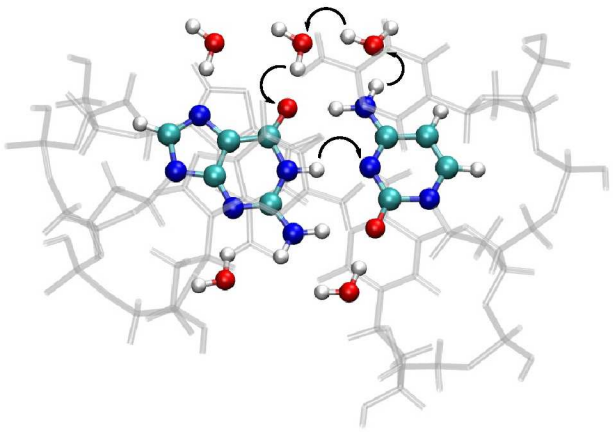
2 Methodology

The choice of the chemical model is critical when simulating large biomolecules. As stated above, the model must reproduce the main chemical interactions that govern the double helix structure, e.g., the interbase H-bonds connecting the base pair as well as the characteristic π-stacking between two consecutive bases located in the same strand.[31,33–35] We have designed a series of chemical models in which the canonical GC structure is sandwiched in a DNA-trimer that mimics all possible d(5'-XGX-3') sequences, i.e., d(5'-AGA-3'), d(5'-AGC-3'), d(5'-AGG-3'), d(5'-AGT-3'), d(5'-CGA-3'), d(5'-CGC-3'), d(5'-CGT-3'), d(5'-CGG-3'), d(5'-GGA-3'), d(5'-GGC-3'), d(5'-GGG-3'), d(5'-GGT-3'), d(5'-TGA-3'), d(5'-TGC-3'), d(5'-TGG-3') and d(5'-TGT-3'). As any model, these chemical systems are simplified representations of the reality, but they allow to accurately predict the frequency at which rare tautomers appear during cell replication if a quantum mechanical treatment is applied.[26] All d(5'-XGX-3') sequences were generated with the X3DNA software package.[36] More specifically, we use X3DNA to built up the double-stranded B-form DNA trimers applying the default parameters (twist = 36°; rise = 3.375 Å). The resulting DNA fragment, which contains ca. 200 atoms, is schematised in Figure 2. The mutation GC→G*C* is subsequently induced in the central base pair.

In order to maintain a reasonable computational cost, we decided to use a hybrid QM:QM' approach in which the geometry of the central GC base pair is optimized at the M06-2X/6-31G(d) level[37], successfully used for modelling DNA-trimer models[26] and that correctly reproduce π-stacking effects,[38] while the rest of the system was frozen in space and described at the same level of theory (border basis pairs) or with PM6[39] (sugar-phosphate backbone). Such computational strategy correctly mimics the characteristic double helix form and simultaneously allows one to relax the central base during the mutagenic process.[26] The vibrational analysis performed in the middle base pair confirmed the nature of the stationary structures: absence of imaginary frequency for minima (both canonical and mutated geometries), and one single imaginary vibration for transition states. Single-point calculations have been performed in the aqueous medium characteristic of living cells through the well-known polarisable continuum model (PCM)[40] with a larger basis set (including diffuse functions)



(a) DNA trimer model with no explicit water molecules



(b) Microhydrated model for the GC embedded DNA trimer

Fig. 2 (a) Side view of one of the trimers used here: the d(5'-GGG-3') sequence. The embedded GC base pair that undergoes the mutation is plotted with balls-and-sticks, while both the border base pairs confining the central GC (also located at the high layer but frozen in the space) and the lateral sugar-phosphate backbone (low layer) are represented in wireframe. (b) Top view of the microhydrated model the central GC base pair in the same DNA trimer. Arrows show the movements of the protons during the water-assisted tautomeric reaction. Optimized atoms at the high layer (middle GC base pair and explicit water molecules) are represented with balls-and-sticks. For the sake of clarity, the rest of the sequence is displayed as shaded-tubes.

to provide more reliable results. The used level of theory can be shortened as: PCM-[M06-2X/6-311++G(d,p):PM6]//M06-2X/6-31G(d):PM6.

To provide a more realistic model for the influence of the aqueous environment into DNA mutation mechanism, the PCM solvent model has been further improved by surrounding the middle GC base with explicit water molecules. More specifically, five water molecules are placed in the vicinity of the solvent-exposed heteroatoms of the GC base pair to model the observed hydration shell around GC base pairs embedded in DNA.[41] The resulting microhydrated-model provides a more refined picture of the biological media as it accounts for both specific GC-water interactions,[42,43] and also allows us to explore the possible catalytic role of the water molecules during the tautomeric equilibria. As illustrated by arrows in Figure 2, the latter mechanism involves the exchange of the proton $H_{4'}$ rather than the H_4 as in the direct tautomeric equilibria (see atomic numbering in Figure 1), so that the mutation arising from the former is hereafter labeled as $G^*C^*_W$ to indicate the catalytic role played by water. All QM simulations have been carried out with the ONIOM scheme[44] implemented in Gaussian09.[45]

3 Results and discussions

Aiming to determine the actual impact of the rare tautomers on the GC-content evolution, the relative electronic energy (ΔE) as well as the relative Gibbs free energies (ΔG) of GC and G*C* forms have been first computed for each DNA sequence with the model showed in the top of Figure 2. The probability of the GC→G*C* mutation is given by the equilibrium constants (K_{eq}):

$$K_{eq} = e^{-\Delta G/RT} \quad (1)$$

where R is the ideal gas constant and T is the temperature (298.15 K). K_{eq} can be compared to the total spontaneous mutation range measured by Topal and Fresco,[46] which lies in the 10^{-8}–10^{-10} range. The simulations for the DNA trimer in absence of explicit water molecules are discussed in next section, and these first results are next compared to the effects induced by the first hydration shell.

3.1 DNA-trimer with no explicit solvation

Let us first discuss the frequency of the mutation when using the model showed in the top of Figure 2, which is given by the difference between the canonical GC and the rare tautomeric G*C* energies. According to the data listed in Table 1, the impact of the sequence on the stability of the G*C* mutation is trifling: all ΔE energies lie in the 9.12–10.54 kcal.mol^{-1} range and similar figures are obtained for ΔG values of 9.14–10.68 kcal.mol^{-1}. Consequently, the formation rate of G*C* is ca. 10^{-8} in most of the cases irrespective of the considered sequence, with the exception of the CGT, CGG and GGC sequence ($K_{eq} \sim 10^{-07}$). This latter value indeed slightly exceeds the upper limit of the spontaneous mutation range measured by Topal and Fresco.[46] Nevertheless, the observed difference is probably too small (ca. 1 kcal.mol^{-1}) to draw general conclusions. Since external base pairs are frozen

Table 1 Relative electronic energy (ΔE/kcal.mol^{-1}), relative Gibbs free energies at 298.15 K (ΔG/kcal.mol^{-1}), and equilibrium constants (K_{eq}) for the G*C* mutation. Solvent effects are described with PCM (no explicit water molecules are added).

Sequence	ΔE	ΔG	K_{eq}
AGA	10.54	10.05	4.23×10^{-08}
AGC	9.99	10.04	4.32×10^{-08}
AGG	10.12	10.68	1.46×10^{-08}
AGT	9.99	10.39	2.38×10^{-08}
CGA	9.92	9.68	8.00×10^{-08}
CGC	9.79	9.58	9.46×10^{-08}
CGT	9.80	9.53	1.03×10^{-07}
CGG	9.64	9.14	1.97×10^{-07}
GGA	9.98	9.83	6.18×10^{-08}
GGC	9.12	9.10	2.11×10^{-07}
GGG	9.74	9.85	5.91×10^{-08}
GGT	9.92	10.25	3.03×10^{-08}
TGA	10.00	9.78	6.76×10^{-08}
TGC	9.93	9.90	5.49×10^{-08}
TGG	9.60	9.69	7.80×10^{-08}
TGT	9.83	10.54	1.87×10^{-08}

Table 2 Computed interbase H-bond distances (in Å) and Mulliken atomic charges (in $|e|$) for the canonical GC DNA-embedded base pair with no explicit water molecules.

	Distance			Charge		
Sequence	O_6–N_4	N_1–N_3	N_2–O_2	H_4	H_1	H_2
AGA	2.797	2.934	2.908	0.33	0.34	0.32
AGC	2.783	2.926	2.963	0.32	0.34	0.32
AGG	2.783	2.917	2.974	0.33	0.34	0.33
AGT	2.808	2.930	2.906	0.32	0.34	0.33
CGA	2.808	2.913	2.849	0.32	0.33	0.34
CGC	2.833	2.913	2.845	0.32	0.34	0.33
CGT	2.820	2.911	2.848	0.32	0.35	0.33
CGG	2.810	2.929	2.872	0.32	0.34	0.33
GGA	2.793	2.934	2.913	0.32	0.34	0.32
GGC	2.803	2.931	2.907	0.32	0.34	0.33
GGG	2.772	2.919	2.895	0.32	0.34	0.33
GGT	2.810	2.935	2.907	0.32	0.34	0.33
TGA	2.808	2.914	2.852	0.32	0.34	0.33
TGC	2.831	2.919	2.851	0.32	0.34	0.33
TGG	2.794	2.916	2.871	0.32	0.34	0.33
TGT	2.821	2.918	2.853	0.32	0.35	0.33

during the mutation, additional calculations have been performed to discard any possible artefact arising from an over-restricted structure. More specifically, we recomputed the energy profile for the tautomeric equilibria of GGG fully optimising the three bases pairs located in the high layer. The obtained results agree with the partial optimisation procedure ($K_{eq} = 2.55 \times 10^{-08}$), and the same conclusion is reached: the mutation rate lies in the order of the spontaneous mutation. To further illustrate the actual effect of the sequence on the structural parameters of GC, Table 2 lists the interbase H-bond distances as well as the partial atomic charges of the interbase protons. An inspection of the geometries reveals only small differences (ca. 0.1 Å) in the H-bond lengths. That homogeneity in the base pair distances is consistent with the nearly constant partial atomic charges for the three protons involved in the GC interbase H-bonds (H_4, H_1 and H_2). Consequently, changing the confining base pairs does not significantly alter the structure of the central GC.

These first data apparently indicate that the G*C* mutation is sequence-independent, so that it is formed with the same probability along all the genetic code. In absence of a preferred location for the GC→G*C* equilibrium, the genetic code would naturally evolve to a complete AT sequence, that is, to the AT catastrophe by following recursively the logical error propagation tree shown in Figure 1. However, it is important to underline that the final impact of the mutation depends on the lifetime of the induced error and not only on thermodynamical parameters. For this reason we have to evaluate the energetic profile of the process, that is, the energetic barrier that interconnects the canonical GC structure to its C*C* counterpart. This requires the use of a discrete solvent model,[47,48] because as noted by Leszczyński and coworkers,[49] the role of water in nucleic acid bases stability should not be underestimated. In addition, it is well-known that the specific interaction of DNA with the surrounding water molecules is crucial to understand the DNA's stability and functionality,[42,43,50–52] and they might also catalyse proton transfer reactions in biological systems.[53,54]

3.2 Microhydrated DNA-trimer

The DNA model is next refined by combining the PCM solvent model with five explicit water molecules around the central GC bas pair. That discrete/continuum solvent model is applied to the sequence with the most stable rare tautomer, GGC (ΔE=9.12 kcal.mol^{-1}; ΔG=9.10 kcal.mol^{-1}) and to the unfavored AGA sequence (ΔE=10.54 kcal.mol^{-1}; ΔG=10.05 kcal.mol^{-1}), as well as to the intermediate GGG (ΔE=9.74 kcal.mol^{-1}; ΔG=9.85 kcal.mol^{-1}). Obviously, this model allows one to explore the possible catalytic role of water molecules during the tautomeric equilibrium (see Figure 2). This alternative mechanism involves the exchange of the proton $H_{4'}$ leading to G*C*$_W$ rather than the H_4 as in the direct tautomeric equilibrium (see atomic numbering in Figure 1).

The results obtained with the hybrid discrete/continuum solvent model are summarised in Tables 3 and 4. The comparison of the relative energies of the microhydrated (Table 3) and the non-microhydrated (Table 1) models demonstrates that the first hydration shell unstabilises the G*C* form, as evidenced

Table 3 Relative electronic energy (ΔE/kcal.mol^{-1}), relative Gibbs free energies at 298.15 K (ΔG/kcal.mol^{-1}), and equilibrium constants (K_{eq}) for the G*C* mutation using selected sequences. Solvent effects are described with a discrete/continuum solvent model.

	G*C*					G*C*$_W$				
Sequence	ΔE_f	ΔE_r[a]	ΔE	ΔG	K_{eq}	ΔE_f	ΔE_r	ΔE	ΔG	K_{eq}
AGA	13.56	0.22	13.34	12.73	4.57×10^{-10}	20.24	2.46	17.78	17.37	1.83×10^{-13}
GGC	14.68	0.19	14.49	11.74	1.74×10^{-09}	25.79	9.99	15.80	15.62	3.50×10^{-12}
GGG	11.14	–	11.46	9.88	5.70×10^{-08}	19.47	5.67	13.80	14.17	4.06×10^{-11}

[a] GG←G*C* is a barrierless reverse reaction in GGG sequence.

Table 4 Computed interbase H-bond distances (in Å) and Mulliken atomic charges (in $|e|$) for the canonical GC DNA-embedded base pair with five explicit water molecules.

	distance			charge		
Sequence	O_6–N_4	N_1–N_3	N_2–O_2	H_4	H_1	H_2
AGA	2.826	2.907	2.854	0.31	0.35	0.32
GGC	2.813	2.918	2.879	0.31	0.35	0.33
GGG	2.783	2.909	2.900	0.32	0.35	0.32

by the larger changes on free energies (and smaller K_{eq}) values. Notably, one can see a difference in the impact of the explicit water molecules in ΔG values. More specifically, the shift of the tautomeric equilibrium towards the canonical GC form is ca. 2 kcal.mol^{-1} for GGC, but only 0.03 kcal.mol^{-1} for the GGG sequence. By comparing the stability of G*C* and G*C*$_W$ within the discrete/continuum model framework we observe that the relative energies of the products related to the water-assisted mechanism are larger by three orders of magnitude than their G*C* counterparts in all cases. This finding hints that imino/enol form generated by the transfer of proton $H_{4'}$ through the water loop is energetically less favourable than the product originating from the direct transfer of H_4. However, we notice a difference in the increase of ΔG depending on the sequence: 4.64 kcal.mol^{-1}, 3.88 kcal.mol^{-1} and 4.29 kcal.mol^{-1} for AGA, GGC and GGG sequences, respectively. In spite of such differences, both G*C* and G*C*$_W$ series follow the same stability order: GGG > GGC > AGA.

Since explicit solvent model has a minor impact on the interbase H-bond lengths (the explicit water molecules weaken the O_6–N_4 bond while reducing the N_1–N_3 distance, but such effect is limited to ca. 0.03 Å), the dissimilarity in the energies reported in Tables 1 and 3 can be attributed to the specific interactions with the water molecules in contact with the GC base pair. As illustrated in Figure 3, surrounding water molecules undergo a large change in their position depending on the sequence. For instance, the water molecule bound to the O_2 atom is significantly displaced in the GGC trimer compared to the two other sequences. More interesting are the changes in the hydration pattern around $H_{4'}$, the proton involved in the water-assisted tautomerisation (front part of

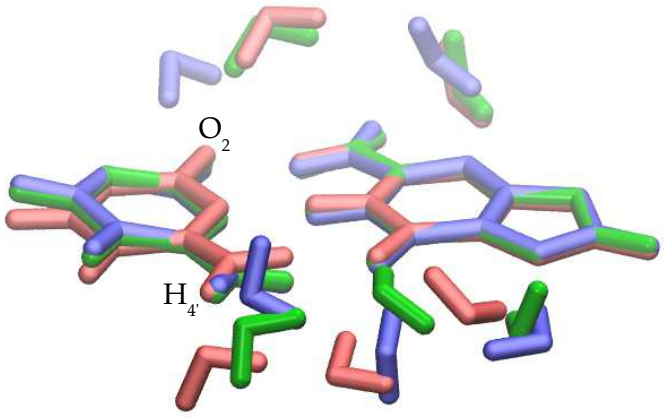


Fig. 3 Overlay the central microhydrated-GC structure embedded in the AGA (red), GGC (blue), and GGG (green) sequences.

Figure 3). In all cases the water molecule close to that position orientates the oxygen towards $H_{4'}$ and make accessible the subsequent PT reaction, though only the water loop for GGG has an optimal structure to facilitate the reaction as the water molecules remain close to the plane of the base pair. This might explain why the GGG sequence leads to the lowest forward barrier (ΔE_f = 19.47 kcal.mol^{-1}) and the most stable G*C*$_W$ mutation ($K_{eq} = 4.06 \times 10^{-11}$). It should be underlined that the genomic sequence does not directly alter the tautomeric equilibrium but the structure of the hydration-shell around the base pair that undergoes the PT reaction, which in turn affects the mutation rate.

We finally analyse the possible macroscopic effect of G*C* rare tautomer by determining the lifetime of the error in

DNA. As discussed in Introduction, Florián and Leszczyński[8] demonstrated that a structural change in the GC base pair would yield to a permanent mutation if the forward barrier (ΔE_f) is less than 28 kcal.mol^{-1} and the reverse (ΔE_r) larger than 3 kcal.mol^{-1}. Table 3 lists the barriers corresponding to the relative energies of the transition states along the GC→G*C* and GC→G*C*$_W$ equilibria. These transition states are characterised by a single imaginary frequency corresponding to the vibration of the protons H_4 and $H_{4'}$, respectively. According to the computed values, all ΔE_f are less 28 kcal.mol^{-1}, and consequently all process satisfy the first requisite. The ΔE_r values show however a different picture. In the particular case of the direct tautomerisation process, the reverse GC→G*C* is a barrierless process ($\Delta E_r < 0.22$ kcal.mol^{-1}), so G*C* does not reach the required lifetime to produce a permanent mutation: it quickly reverts to the canonical GC. In contrast, the GC→G*C*$_W$ equilibrium is associated to a much larger barrier. This is the logical consequence water-assisted mechanism, as the proton is transferred from the N_4 atom in cytosine to the O_6 site of guanine through a water loop (see Figure 2). G*C*$_W$ is consequently located in a deeper minimum in the potential energy surface for both GGC and GGG with a back barrier of 9.99 and 5.67 kcal.mol^{-1}, respectively. This is not the case for AGA, with a $\Delta E_r = 2.46$ kcal.mol^{-1}. Consequently, G*C*$_W$ fulfils all Florián and Leszczyński's kinetic prerequisites in GGC and GGG sequences, but not in AGA. The G*C*$_W$ emerges as the most prone rare tautomer to be present in DNA during its replication in specific sequences. Let us focus on the K_{eq} for GC→G*C*$_W$ reaction to extract biological conclusions. The computed K_{eq} values lie bellow the total rate of spontaneous mutation established by Topal and Fresco (10^{-8} and 10^{-10}).[46] As expected, this result indicates that the water-assisted tautomeric equilibrium partially contributes to the global measured mutation, but it is not the only source of genetic errors. A close inspection of Table 3 shows that the mutation rate in the GGG ($\sim 10^{-11}$) sequence is significantly larger compared to GGC ($\sim 10^{-12}$) and AGA ($\sim 10^{-13}$): the G*C*$_W$ tautomeric mutation is mainly located in GC-rich regions, which are also the gene-rich regions in the genome compared to GC-poor regions, that are mainly deserts of genes.[27] Consequently, although the amount of G*C*$_W$ seems *a priori* small, it might have important consequences as it is concentrated in genomic fragments of prime relevance. The reported theoretical evidences allow us to conclude that the decreasing of the GC-content along with the genome will be mainly located in coding regions. It is quite remarkable that a slight structural change such as the exchange of two protons could be crucial in DNA functionality.

4 Overview and conclusions

The present contribution brings the attention to one of the plausible causes of the evolution of GC-content in DNA: the proton exchange in the GC base pair. We used a series of genetic sequences within realistic chemical models accounting for the combined effects of stacking and hydration and relied on a hybrid QM:QM' approach [M06-2X/6-311++G(d,p):PM6]. Our calculations reveal that the GC→G*C* reaction is not equally probable along DNA, as the reorganisation of water molecules enhancing the stability of the mutagenic process in GGG sequence, and consequently the GC→AT mutations are mainly located at GC-rich regions. The used computational protocol is accurate enough to shed light into causal factors in the evolution of the GC-content, one of the remaining questions to fully understand the structural changes in DNA. However, further work is clearly required to explore the dynamics of the tautomeric equilibria, for instance by means of *ab initio* molecular dynamics.[55–58] We hope that the reported data could be used to guide other simulations while helping into the debate related to the equilibrium point for the GC-content.

5 Conflict of interest

The authors declare no competing financial interest.

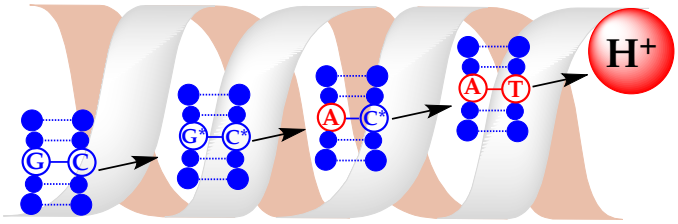
6 Acknowledgement

J.P. C.-C. acknowledges the support from the FP7 EU Marie Curie Actions through the Campus Mare Nostrum 37/38 CMN UMU Incoming Mobility Programme ACTion (U-IMPACT), and the support provided by Université de Nantes through its invited professor programme. D.J. acknowledges the European Research Council (ERC) and the *Région des Pays de la Loire* for financial support in the framework of a Starting Grant (Marches - 278845) and a *recrutement sur poste stratégique*, respectively. This research used resources of the CEN1 in Spain and the CINES and the CCIPL in France.

References

1 Lander, E. et al. *Nature* **2001**, *6822*, 860–921.
2 Stein, L. D. *Nature* **2004**, *431*, 915–916.
3 Watson, J. D.; Crick, F. H. C. *Nature* **1953**, *171*, 737–738.
4 Manalo, M. N.; Pérez, L. M.; LiWang, A. *J. Am. Chem. Soc.* **2007**, *129*, 11298–11299.
5 Frank-Kamenetskii, M. D. *Nature* **1987**, *328*, 17–18.
6 Villani, G. *J. Phys. Chem. B* **2010**, *114*, 9653–9662.
7 Löwdin, P. O. *Rev. Mod. Phys.* **1963**, *35*, 724–732.
8 Florián, J.; Leszczyński, J. *J. Am. Chem. Soc.* **1996**, *118*, 3010–3017.
9 Gorb, L.; Podolyan, Y.; Dziekonski, P.; Sokalski, W. A.; Leszczyński, J. *J. Am. Chem. Soc.* **2004**, *126*, 10119–10129.

10 Kryachko, E. S. *Int. J. Quantum Chem.* **2002**, *90*, 910–923.
11 Bertran, J.; Blancafort, L.; Noguera, M.; Sodupe, M. In *Computational Studies of RNA and DNA*; Šponer, J., Lankas, F., Eds.; Springer, 2006; pp 411–432.
12 Kumar, A.; Sevilla, M. D. *Chem. Rev.* **2010**, *110*, 7002–7023.
13 Herrera, B.; Toro-Labbé, A. *J. Phys. Chem. A* **2007**, *111*, 5921–5926.
14 Cerón-Carrasco, J. P.; Requena, A.; Michaux, C.; Perpète, E. A.; Jacquemin, D. *J. Phys. Chem. A* **2009**, *113*, 7892–7898.
15 Cerón-Carrasco, J. P.; Requena, A.; Zúñiga, J.; Michaux, C.; Perpète, E. A.; Jacquemin, D. *J. Phys. Chem. A* **2009**, *113*, 10549–10556.
16 Gu, J.; Wong, N.-B.; Xie, Y.; Schaefer III, F. H. *Chem. Eur. J.* **2010**, *16*, 13155–13162.
17 Liu, H.; Li, G.; Zhang, L.; Li, J.; Wang, M.; Bu, Y. *J. Chem. Phys.* **2011**, *135*, 134315.
18 Lin, Y.; Wang, H.; Gao, S.; Schaefer III, H. F. *J. Phys. Chem. B* **2011**, *115*, 11746–11756.
19 Chen, H.-Y.; Yeh, S.-W.; Hsu, S. C. N.; Kao, C.-L.; Dong, T.-Y. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2674–2681.
20 Cooper, W. G. *Int. J. Quantum Chem.* **2012**, *112*, 2301–2323.
21 Lin, Y.; Wang, H.; Gao, S.; Li, R.; Schaefer III, H. F. *J. Phys. Chem. B* **2012**, *116*, 8908–8915.
22 Barnett, R. N.; Joseph, J.; Landman, U.; Schuster, G. B. *J. Am. Chem. Soc.* **2013**, *135*, 3904–3914.
23 Hsu, S. C. N.; Wang, T.-P.; Kao, C.-L.; Chen, H.-F.; Yang, P.-Y.; Chen, H.-Y. *J. Phys. Chem. B* **2013**, *117*, 2096–2105.
24 Brovarets, O. O.; Zhurakivsky, R. O.; Hovorun, D. M. *Phys. Chem. Chem. Phys.* **2014**, *16*, 3715–3725.
25 Villani, G. *J. Chem. Phys. B* **2014**, *118*, 5439–5452.
26 Lin, Y.; Wang, H.; Wu, Y.; Gao, S.; Schaefer III, H. F. *Phys. Chem. Chem. Phys.* **2014**, *16*, 6717–6725.
27 Galtier, N.; Piganeau, G.; Mouchiroud, D.; Duret, L. *Genetics* **2001**, *159*, 907–911.
28 Karro, J. E.; Peifer, M.; Hardison, R. C.; Kollmann, M.; von Günberg, H. H. *Mol. Biol. Evol.* **2007**, *25*, 362–374.
29 Chen, H.-Y.; Kao, C.-L.; Hsu, S. C. N. *J. Am. Chem. Soc.* **2009**, *131*, 15930–15938.
30 Kobylecka, M.; Leszczyński, J.; Rak, J. *J. Chem. Phys.* **2009**, *131*, 085103–1.
31 Galano, A.; Alvarez-Idaboy, J. R. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12476–12484.
32 Kumar, A.; Pottiboyina, V.; Sevilla, M. D. *J. Phys. Chem. B* **2011**, *115*, 15129–15137.
33 Gil, A.; Branchadell, V.; Bertran, J.; Oliva, A. *J. Phys. Chem. B* **2009**, *113*, 4907–4914.
34 Acosta-Silva, C.; Branchadell, V.; Bertran, J.; Oliva, A. *J. Phys. Chem. B* **2010**, *114*, 10217–10227.
35 J. P. Cerón-Carrasco, J. Zúñiga, A. Requena, E. A. Perpète, C. Michaux and D. Jacquemin, *Phys. Chem. Chem. Phys.*, 2011, **13**, 14584–14589.
36 Lu, X.-J.; Olson, W. K. *Nat. Protoc.* **2008**, *3*, 1213–1227.
37 Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
38 Capobianco, A.; Peluso, A. *RSC Adv.* **2014**, *4*, 47887–47893.
39 Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173–1213.
40 Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.
41 Schneider, B.; Berman, H. M. *Biophys. J.* **1995**, *69*, 2661–2669.
42 Auffinger, P.; Westhof, E. *J. Mol. Biol.* **2000**, *300*, 1113–1131.
43 Makarov, V.; Pettitt, B.; Feig, M. *Acc. Chem. Res.* **2002**, *35*, 376–384.
44 Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (Theochem)* **1999**, *462*, 1–21.
45 Frisch, M. J. et al. Gaussian 09 Revision D.01. 2009; Gaussian Inc. Wallingford CT.
46 Topal, M. D.; Fresco, J. R. *Nature* **1976**, *263*, 285–289.
47 Alemán, C. *Chem. Phys.* **2000**, *253*, 13–19.
48 Alemán, C. *Chem. Phys.* **1999**, *244*, 151–162.
49 Furmanchuk, A.; Shishkin, O. V.; Isayev, O.; Gorb, L.; Leszczyński, J. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9945–9954.
50 Rueda, M.; Kalko, S. G.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **2003**, *125*, 8007–8014.
51 Kabelac, M.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 903–917.
52 Kumar, A.; Sevilla, M.; Suhai, S. *J. Phys. Chem. B* **2008**, *112*, 5189–5198.
53 Przybylski, J. L.; Wetmore, S. D. *J. Phys. Chem. B* **2010**, *114*, 1104–1113.
54 Alagona, G.; Ghio, C.; Nagy, P. I. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10173–10188.
55 Xiao, S.; Wang, L.; Liu, Y.; Lin, X.; Liang, H. *J. Chem. Phys.* **2012**, *137*, 195101.
56 Loos, P. F.; Dumont, E.; Laurent, A. D.; Assfeld, X. *Chem. Phys. Lett.* **2009**, *475*, 120–123.
57 Dupont, C.; Patel, C.; Dumont, E. *J. Phys. Chem. B* **2011**, *115*, 15138–15144.
58 Garrec, J.; Patel, C.; Rothlisberger, U.; Dumont, E. *J. Am. Chem. Soc.* **2012**, *134*, 2111–2119.



Graphical abstract: We use theoretical tools to investigate the possible role played by DNA sequence in the base pair tautomerization phenomena.