Sequence dependence of transient Hoogsteen base-pairing in DNA

Alberto Pérez de Alba Ortíz [1,2], Jocelyne Vreede [1] and Bernd Ensing [1,3*]

[1] Van 't Hoff Institute for Molecular Sciences and Amsterdam Center for Multiscale Modeling, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.

[2] Soft Condensed Matter, Debye Institute for Nanomaterials Science, Department of Physics, Utrecht University, Princetonplein 1, 3584 CC Utrecht, The Netherlands.

[3] AI4Science Laboratory, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.

August 4, 2021

Abstract

Hoogsteen (HG) base-pairing is characterized by a 180° rotation of the purine base with respect to the Watson-Crick-Franklin (WCF) motif. Recently, it has been found that both conformations coexist in a dynamical equilibrium and that several biological functions require HG pairs. This relevance has motivated experimental and computational investigations of the base-pairing transition. However, a systematic simulation of sequence variations has remained out of reach. Here, we employ advanced path-based methods to perform unprecedented free-energy calculations. Our methodology enables us to study the different mechanisms of purine rotation, either remaining inside or after flipping outside of the double helix. We study seven different sequences, which are neighbor variations of a well-studied A·T pair in A_6-DNA. We observe the known effect of A·T steps favoring HG stability, and find evidence of triple-hydrogen-bonded neighbors hindering the inside transition. More importantly, we identify a dominant factor: the direction of the A rotation, with the 6-ring pointing either towards the longer or shorter segment of the chain, respectively relating to a lower or higher barrier. This highlights the role of DNA's relative flexibility as a modulator of the WCF/HG dynamic equilibrium. Additionally, we provide a robust methodology for future HG proclivity studies.

1 INTRODUCTION

In 1953, emblematic studies by Watson and Crick [1], and Franklin and Gosling [2] defined the structure of DNA for the first time. The discovery of a specific pairing of purine and pyrimidine bases via hydrogen bonds revealed not only DNA's double-helical shape, but also the basis for its replication, which is essential for a genetic information carrier. In Fig. 1A we depict a Watson-Crick-Franklin (WCF) A·T base pair, with the characteristic hydrogen-bonding pattern. Six years later, in 1959, Hoogsteen proposed an alternative base-pairing motif based on crystallographic data from A-T crystals, in which the purine base *rolls* 180° around the glycosidic bond [3], i.e. from *anti* to *syn*, with respect to the WCF geometry, as depicted in Fig. 1B. In the Hoogsteen (HG) configuration, the pyrimidine forms hydrogen bonds with the 5-ring of the purine, rather than with its 6-ring, causing the opposite backbone C1' atoms of the bases to be at a somewhat shorter distance, as well as some degree of twisting and bending of the double helix in the vicinity of the base pair [3, 4]. In the last decades, a number of studies have shown that the abundance of the HG conformation is non-negligible in canonical duplex DNA, and that its biological implications are very relevant. In 2011, Nikolova and coworkers reported the transient presence of HG base pairs in specific steps inside canonical duplex DNA using nuclear magnetic resonance (NMR) relaxation dispersion spectroscopy [5]. They reported populations of A·T and G·C HG base pairs of around 0.5%, with residence times of up to 1.5 ms. Recent studies have measured even larger HG populations, of 1.2%, in an A·T rich segment [6]. A few years later, Alvey and colleagues, also using NMR relaxation dispersion, demonstrated that HG base pairs appear in more diverse sequences than previously expected [7]. These studies shifted the perspective on HG base pairs, which were initially thought to appear mainly in distorted or damaged DNA, but are now considered to coexist with the WCF motif in a dynamic equilibrium [8]. A recent survey of structures of DNA-protein complexes in the Protein Data Bank (PDB) showed 140 HG base pairs out of a total of around 50,000 [4]. Some of these complexes are of particular biological relevance. For example, the human DNA polymerase-ι performs replication exclusively via HG base pairing [9, 10], a function that was previously thought to be unique of WCF base pairs. Other examples of the involvement of HG in DNA-protein complexes include the p53 tumor suppressor protein [11], the TATA-box binding protein involved in transcription [12], and the MATα2 homeodomain which regulates transcription in cells [13].

The newly discovered relevance of HG base pairs, and the dynamic equilibrium in which they coexist with WCF base pairs, demands a more detailed understanding of the transition mechanisms between both conformations. Given the difficulty of observing

*Email: b.ensing@uva.nl

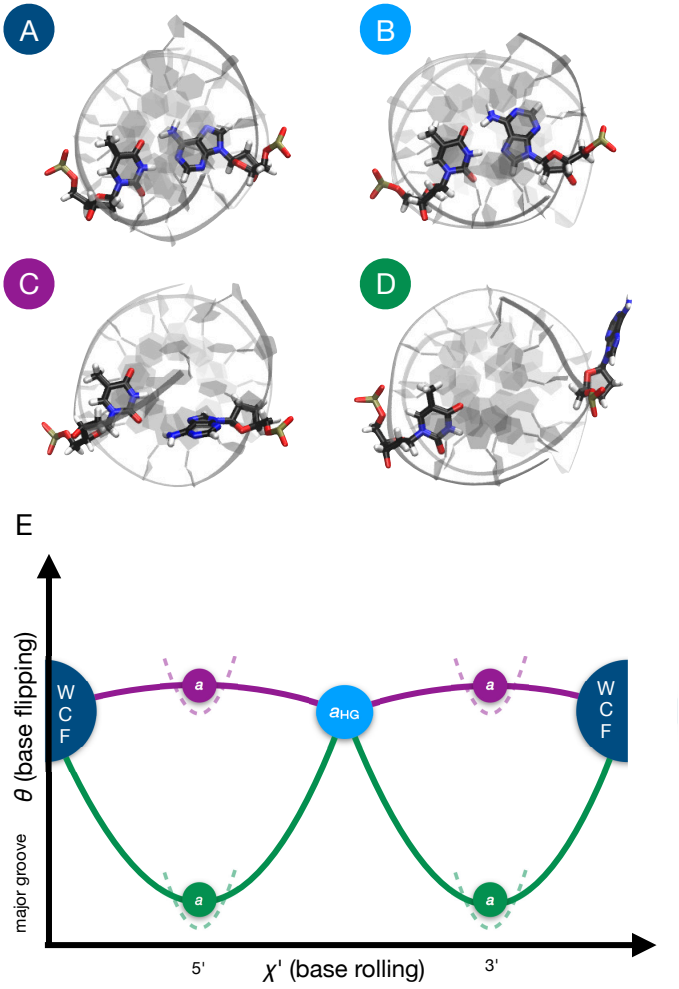

Figure 1: **(A)** WCF conformation of the A16·T9 base pair in A_6-DNA. **(B)** HG conformation of the same base pair. **(C)** Intermediate state of the WCF-to-HG transition via the *inside* pathway, i.e. with the rolling of A16 occurring within the double helix. **(D)** Intermediate state of the WCF-to-HG transition via the *outside* pathway. i.e. with A16 flipped out of the double helix via the major groove **(E)** Two cyclic paths in the space of the collective variables (CVs) χ' and θ, which describe the rolling and the flipping of A16, respectively. Both paths start and end at the WCF state, describing a full revolution of A16 with intermediate states in which the 6-ring of A16 points toward the 5' or the 3' direction of DNA. The HG state is marked by an attractor, represented by the circle labeled a_{HG}. The purple path samples the inside mechanism, while the green path explores the outside mechanism. Additional attractors, represented by the circles labeled a, are restrained between the WCF and the HG states along their respective paths and are also restrained along θ to prevent switching between the inside and the outside mechanisms. Restraints are represented by dashed lines.

the short-lived intermediate states experimentally, computational approaches based on molecular dynamics (MD) simulations, boosted with enhanced sampling, have provided valuable insights. In their original work from 2011, Nikolova and co-workers complemented their NMR results with conjugate peak refinement (CPR) simulations. They simulated the DNA sequence 5'-CGATTTTTTGGC-3' (A_6-DNA) in vacuo using the CHARMM27 force field [14], and studied the transition of the A16·T9 pair. Two collective variables (CVs) were used to steer the transition: (1) the χ glycosidic angle, which describes the rolling of A16 around the glycosidic bond and is defined by the atoms O4'-C1'-N9-C4; and (2) the θ base opening angle, which describes the flipping of A16 outside of the double helix towards the major groove and is defined in [15]. The CPR simulations revealed two types of pathways: one in which the rolling of the adenine, either clockwise or counterclockwise, occurs within the double helix, i.e. with a small base opening angle; and another in which the rolling occurs outside of the double helix, i.e. with a large base opening angle. We refer to these two types of pathways as *inside* and *outside*, respectively. Examples of intermediate states of the two types of pathways are shown in Fig. 1C and D. The (χ, θ) pseudo-free-energy landscapes of Nikolova and coworkers showed an overwhelming preference for the inside mechanism, likely influenced by the lack of solvent to stabilize the flipped conformations. In 2015, Yang and colleagues used umbrella sampling (US) to obtain a (χ, θ) free-energy landscape of the same

A16·T9 pair in A_6-DNA [16]. They employed a version of the AMBER99-BSC0 force field [17] with modified parameters for the glycosidic torsion, and used explicit TIP3P water [18]. Their vast US calculations, with over 300 windows, revealed multiple pathways, including inside ones, and outside ones opening either towards the major or the minor groove. Yang and coworkers reported a free-energy difference from the WCF to the HG state of 4.4 kcal/mol, close to Nikolova and coworkers' NMR result of 3.0 kcal/mol. The free-energy barrier was of 10-11 kcal/mol for the inside pathways, and up to 14 kcal/mol for the outside ones, again close to NMR results of 16 kcal/mol. A few years later, Yang and coworkers reported another (χ, θ) free-energy surface, this time for a base pair with an N1-methylated adenine [19]. Similarly, Ray and Andricioaei calculated a (χ, θ) free-energy surface for DNA and RNA [20]. While such 2D free-energy landscapes are rich in insight, they also demand significant computational expense to obtain several μs-long MD runs; hindering the systematic study of multiple DNA variations. In particular, investigations of the effect of the DNA sequence on HG base-pairing proclivity have remained beyond computational reach, even though there is experimental evidence that specific sequence patterns, e.g. A·T steps, can favor HG base pairing [21].

Path-based approaches offer an efficient alternative to study the WCF-to-HG transition. Techniques like path-metadynamics (PMD) [22]—i.e. metadynamics using as a unique CV the progress along an adaptive path connecting two known states in a multidimensional CV-space—focus sampling on the transition channels, rather than on the entire conformational landscape, e.g. the entire (χ, θ) plane. Recently, we performed a PMD mechanistic study [23], where we modeled the known A_6-DNA with the state-of-the-art AMBER99-BSC1 force field [24] in explicit solvent. We focused on one direction of rotation, with the A16 6-ring pointing toward the A17 neighbor in the 3' direction of DNA. We analyzed additional CVs, such as: the donor-acceptor distances of the characteristic hydrogen bonds of each type of base pairing, the distance between the backbone C1' atoms, and the distance between the neighboring bases of the rolling A16 (C15 and A17). The latter CV displayed an increase at the mid-rotated intermediate state of inside pathways, indicating that the neighboring bases are displaced in order to accommodate the rotation of the adenine. In contrast, the distance between bases neighboring the rolling A16 remained unaffected in outside transitions, indicating that the inside pathways might be more subject to sequence dependence than the outside ones. We also found that, while the θ pseudo-dihedral as defined in [15] is effective at biasing base opening, the χ dihedral can induce rotations of the sugar ring, rather than of the adenine. This can be addressed by defining a less locally dependent pseudo-dihedral based on centers of mass, χ', as done in [25]. Most importantly, and contrary to previous studies, we found a less pronounced preference for the inside pathway [23]. Together with other coauthors, we recently reported another pathway-focused study [25]. We used transition path sampling (TPS) [26] to generate unbiased trajectories, and transition interface sampling (TIS) [27] to calculate a WCF-to-HG free-energy difference of 3.2 kcal/mol, which compares well with Nikolova and coworker's previous experiments [5]. Most importantly, TPS showed trajectories spontaneously changing from the inside to the outside mechanism, evidencing a preference for the latter one. This raises a debate with the previous results from free-energy surfaces, which favored inside pathways, although having either no explicit solvent [5] or restrained neighboring bases in their setup [16]. A predominance of outside pathways could explain why sequence variation does not seem to impact significantly the WCF-to-HG free-energy barrier, according to Alvey and coworkers' NMR results [7]. Markov state modeling has also confirmed a dominant outside pathway [20].

In this work, we apply recent advances in PMD—which allow to treat multiple pathways in parallel—to study the WCF-to-HG inside and outside mechanisms in diverse DNA sequences with high efficiency. Unlike standard metadynamics [28], PMD does not suffer from exponential performance scaling with the number of CVs [29]. This is because the sampling is effectively done in only 1D, i.e. on the normalized progress component along the path, s, connecting the two known stable states in CV-space. This sampling along the path can be done with other well-established methods, like US, or using common algorithmic extensions, such as multiple walkers [23]. The path curve is optimized based on the restrained cumulative sampling density, which is an estimator of the free-energy gradient, such that the method converges to the closest low free-energy path from the initial guess. However, this optimization can be challenging in systems with multiple pathways, especially when the transition can switch between mechanisms, as seen for the WCF-to-HG inside and outside pathways [23, 25]. To tackle this scenario, we recently developed multiple-walker multiple-path-metadynamics (MultiPMD). In this scheme, we initialize several paths, each with an associated group of walkers. Under normal conditions, all paths would converge to the same low free-energy channel. By introducing special walkers, i.e. *repellers* or *attractors*, the paths can be forced to diverge and find alternative mechanisms connecting the known states. Thus, with MultiPMD one can calculate the free-energy profiles along multiple pathways simultaneously, with sub-exponential performance scaling with the number of CVs, and exploiting the parallelism of current supercomputing resources.

Here, we apply MultiPMD to find inside and outside pathways, and free-energy profiles, for the WCF-to-HG transition of the A16·T9 pair in seven sequence variations based on the well-studied A_6-DNA sequence. In order to verify the effect of local variations, we test four different nucleobases (A,T,G,C) as direct neighbors on the 5' side of the rolling A16. We repeat the procedure for the neighboring base pair on the 3' side. Our MultiPMD methods provide, for the first time, sufficient computational agility for a systematic evaluation of HG base pairing in multiple DNA chains. This enables us to extract trends about the structure and free energy of the base-pairing transition across sequences.

2 SIMULATION PROTOCOL

2.1 Sequence variations

From [5] we take the original A_6-DNA sequence 5'-CGATTTTTTGGC-3', with its complementary strand 3'-GCTAAAAAACCG-5', as shown in Fig. S1. We study the WCF-to-HG transition of the A16·T9 base pair, which is produced by the 180° rotation of A16 around its glycosidic bond. The original direct neighbors of A16 are C15 and A17, in the 5' and in the 3' direction respectively, which gives the local environment CAA. To test the local effect of the direct neighbor of A16 in the 3' direction, we replace A17 with T, G and C. This yields the new local environments CAT, CAG and CAC. We apply the same criteria for the direct neighbor of A16 in the 5' direction, and replace C15 with A, T and G. This gives the new local environments AAA, TAA and GAA. All sequences are generated as ideal B-DNA duplex structures using the make-na tools [30], and can be regenerated with the current W3DNA server [31]. Since make-na generates WCF base pairs, we manually rotate A16 to obtain the HG state for each sequence.

2.2 Molecular dynamics

System preparation is done with GROMACS 5.4.1 [32]. We employ the AMBER99-BSC1 force field [24] to model the DNA's interatomic interactions. Each sequence is placed in a periodic dodecaedron box, with a distance of 1 nm between the DNA sequence and the edge of the box. Each system is then solvated in water, which is modeled by the TIP3P [18] force field. We add 25 mM of NaCl to mimic physiological conditions and neutralize the charge of each system. The systems are energy minimized at each stage of the preparation process. To run MD, the canonical sampling through velocity rescaling (CSVR) thermostat [33] is set at at 300 K, and the Parrinello-Rahman barostat [34] at 1 bar. Each sequence, both at the WCF and at the HG state, is equilibrated for 100 ns with a time step of 2 fs. From the equilibrations, we analyze key structural features and their variations from sequence to sequence.

2.3 Multiple-path-metadynamics

MultiPMD production runs are performed with GROMACS 5.4.1 [32], patched with PLUMED 2.3.1 [35] and with the added PMD code, available at: `www.acmm.nl/ensing/software/PathCV.cpp`. We use two CVs: the base opening pseudo-dihedral angle θ, as defined in [15], which has been repeatedly proven effective to bias base flipping [5, 16, 23]; and the base rolling pseudo-dihedral angle χ', as defined in [23], which is a correction to the glycosidic torsion χ to prevent sugar rotation. The fitness of these two CVs has been discussed in [36]. A value of $\chi' = 0$ rad implies a mid-rotated A16 with its 6-ring pointing in the 3' direction, while $\chi' = \pm\pi$ implies that the A16 6-ring points toward the 5' direction. Negative values of θ imply base flipping toward the major groove. To handle both directions of rotation of A16, the χ' angle is treated with its sine and cosine. Then, the paths are curves in the CV-space spanned by $[\cos(\chi'),\ \sin(\chi'),\ \theta]$. These paths are cyclic, starting and ending at the WCF configuration, whose coordinates in the $[\cos(\chi'),\ \sin(\chi'),\ \theta]$-space are determined by averaging the CVs from the corresponding equilibration run. The initial guess for all paths describes a full revolution of A16, transitioning from WCF to HG to WCF, with no flipping. All outside paths are discretized as strings with 39 nodes, with the initial and the final node fixed at the same point in the $[\cos(\chi'),\ \sin(\chi'),\ \theta]$-space, which marks the WCF state. The inside paths are discretized as strings of 11 nodes, since they require less flexibility. The progress component along the path, s, which usually grows from $s = 0$ to $s = 1$ from the initial to the final node, is now set to grow from $s = -1$ to $s = +1$, and to be periodic in the same range, in order to match the cyclic nature of the path. This implies that the WCF state corresponds to $s = \pm 1$. on the other hand, the HG state corresponds to $s \approx 0$, but the exact value cannot be known a priori. Assigning the HG state to a specific fixed node in the middle of the path would require to make an assumption about the length of each section of the path. Instead, the HG state is marked by an attractor, i.e. a walker that does not participate in the free-energy calculation. The HG attractor is harmonically restrained, with a force constant of 50 kcal/mol, at the average value of χ' and θ during the corresponding equilibration run. Values of $-1 < s < 0$ imply a rotation with the A16 6-ring in the 5' direction, while values of $0 < s < +1$ signify a rotation in the 3' direction.

For each sequence, we initialize two paths, one for the inside, and one for the outside mechanism. We only consider the outside mechanism with opening toward the major groove, as previous work already demonstrated that opening towards the minor groove is unlikely [16, 25]. As we reported in [23], switching between the inside and the outside mechanisms can occur during a PMD calculation of base rolling. Since path-switching prevents convergence, we use a new strategy to keep the paths apart. The separation is induced by special walkers, called attractors, that do not take part in the free-energy calculation and guide the paths through specific intermediate states. We add two attractors on each path, which are steered to $s = 0.5$ and $s = -0.5$ by a moving harmonic restraint with a force constant of 5000 kcal/mol per squared path unit during the first 20 ps of the simulation, such that they are located at intermediate states in the WCF-to-HG and the HG-to-WCF sections of the cyclic path. These intermediate states between the two kinds of base pairing have the A16 mid-rotated, perpendicular to the other bases. To keep the inside path from flipping, its two attractors are restrained by a harmonic potential with a force constant of 5000 kcal/(mol rad^2) at $\theta = 0.0$,

which prevents them from leaving the confines of the double helix. In turn, the repellers of the outside path are restrained by a harmonic potential with a force constant of 5000 kcal/(mol rad^2) at $\theta = -\pi/2$, which ensures that the A16 stays flipped toward the major groove at the mid-rotation. The large values for all the force constants are chosen because the attractors are located near the top of the free-energy barriers, and we require them to remain in these high-energy, mid-rotated, conformations. Then, each path has one attractor at the HG state, and two attractors at intermediate states; one at $s = 0.5$ and one at $s = -0.5$. In Fig. 1E, we show a scheme of the inside and outside paths, the fixed nodes, the HG attractor and the intermediate-state attractors..

Nine standard walkers—for a total of 12 walkers per path—perform metadynamics on the s component of the path. The metadynamics Gaussian potentials have a width of 0.1 path progress units and a height of 0.05 kcal/mol, and are deposited every 1 ps. The paths are updated every 1 ps. The value of the half-life parameter is infinite for inside paths and 20 ps for outside paths. This parameter determines the flexibility of the path by setting the amount of simulation time that it takes for previous samples to weight only 50% of their original value for path updates. We use a *tube* potential—i.e. an upper harmonic wall at a distance of 0.0 on the component perpendicular to the path, z—with a force constant of 50 kcal/mol per path unit to maintain all walkers near the path. For all the sequences, we analyze the adapted paths after 7 ns of sampling. We obtain free-energy profiles and error bars from the average and standard deviation of the metadynamics estimation from 2 ns until 7 ns of sampling, with samples taken every 0.1 ns. There are a few exceptions to the general protocol, which we detail in the Supporting Information.

3 RESULTS AND DISCUSSION

3.1 Stable-state structures

From the 100 ns equilibration runs at the WCF and HG stable states, we analyze structural variations between the different sequences. Table S1 presents the definitions, averages and standard deviations of several CVs during the equilibrations. See [23] for more details about the CVs. First, we note that all equilibrations are stable at the respective base-pairing configurations, as evidenced by the characteristic hydrogen-bond distance, d_{WCF} or d_{HG}, and the conserved d_{HB} with average values of ~ 3 Å; and by the base rolling angle χ' close to either +1.5 rad in WCF, or to -1.5 rad in HG. However, the standard deviations of d_{HG} for the HG states, which range from 0.2 to 0.7 Å, already indicate that the stability of HG base pairing is not equal for all sequences. In contrast, the standard deviation of d_{WCF} for WCF base pairs is of ~ 0.1 Å, indicating more rigid hydrogen bonds. Another structural signature of the two kinds of base pairing is the distance between the C1' atoms of A16 and T9, d_{CC}, which shows an expected constriction from ~ 10.6 in WCF, to ~ 9.1 Å in HG conformations [4]. A key feature that varies significantly from sequence to sequence is the distance between the neighboring bases, d_{NB}, with a range from 7.4 to 8.0 Å. In the following sections, we show how this CV relates to the free-energy barriers from the WCF-to-HG transition.

3.2 Free-energy differences and barriers

Inside and outside pathways, together with their corresponding free-energy profiles, are shown in Fig. S2. The attractors successfully keep the paths separated, with the inside ones near $\theta = 0$ and the outside ones flipping toward the major groove. The free-energy profile for the CAA sequence compares well to our results from [23] and [36], which consider only the 3' direction of rotation. To simplify the analysis of the numerous free-energy profiles, in Fig. 2 we show only the barriers, in both the 3' and the 5' directions of rotation, as well the free-energy difference from WCF to HG. The free-energy differences calculated for the same sequence along inside and outside paths are consistent within ~ 2 kcal/mol; providing a sensible check for our profiles (see Fig. 2B). As expected, WCF base pairing is preferred over HG in all cases. The free-energy difference between both states ranges from ~ 0.5 to ~ 6 kcal/mol across all sequences. This range agrees with free-energy differences obtained by NMR relaxation dispersion for other sequences [7]. According to our calculations, the sequence TAA presents the most favored HG state, with a free-energy difference of 0.6 to 2 kcal/mol with respect to WCF; and with the lowest barrier overall (9.7 kcal/mol) via the outside 3' rotation. This result agrees with experimental reports of HG stability being increased by A·T steps [21]. The CAT sequence shows the most disfavored HG state, with a difference of 4.9 to 5.9 kcal/mol and the highest barrier overall (20 kcal/mol) via the 5' inside rotation.

For all sequences, the lowest free-energy barrier is that of the outside pathway with a 3' rotation (see Fig. 2C). Such a prevalent outside mechanism agrees with our previous results for the CAA sequence using TPS [25], as well as Markov state modeling by Ray and Andricioaei [20]. Most of the outside 3' rotations also show a transition state closer to HG ($s \approx 0$) than to WCF, which agrees with experiments [7]. These results imply that the conformational penalty of A16 rolling within the double helix is higher than that of it flipping out and back in.

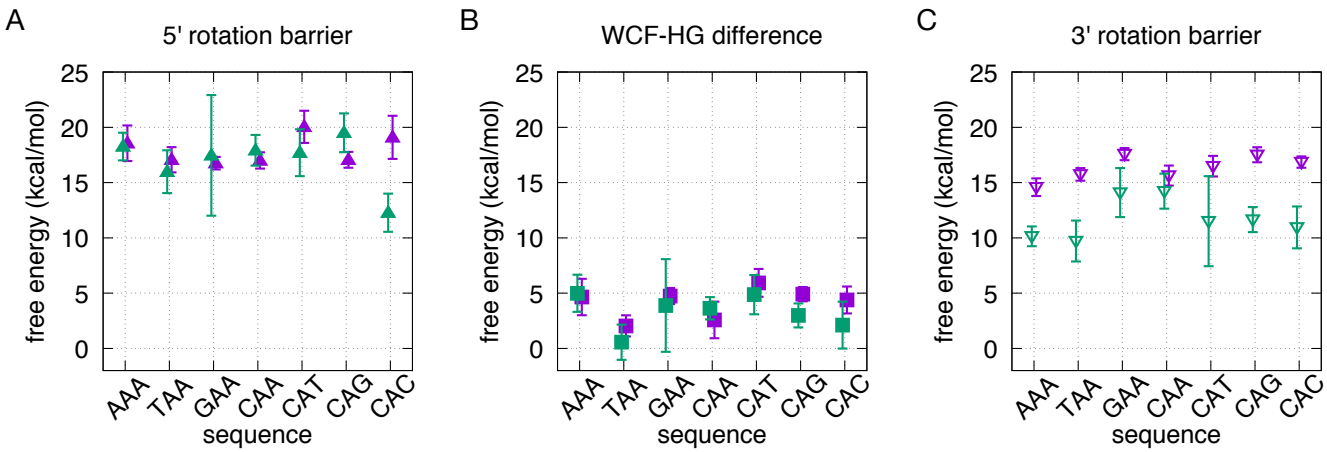

Figure 2: WCF-to-HG free-energy differences and barriers for the seven sequences shown in Fig.S1, extracted from the average free-energy profiles and error bars in Fig. S2. Inside paths are shown in purple and outside paths are shown in green. **(A)** Free-energy barriers via the rotation with the 6-ring of A16 pointing in the 5' direction. **(B)** Free-energy differences. **(C)** Free-energy barriers via the rotation with the 6-ring of A16 pointing in the 3' direction.

3.3 Sequence dependence of the free-energy barriers

In [23], we showed how the inside rotation induces an increase in the distance between the neighbors of A16, which suggests that inside pathways have a stronger neighbor-dependence. Specifically, one may hypothesize that less flexible—i.e. double-ringed (A,G) or triple-hydrogen-bonded (G,C)—neighboring bases can hinder the inside rotation of A16. We observe such a trend in the barriers for the 3' inside rotation (see Fig. 2C). The 5' neighbor variations with respect to CAA yield the following ranking of sequences, from the highest to the lowest barrier: GAA, CAA, TAA, AAA. While the 3' neighbor variations yield the following order, again for the highest to the lowest barrier: CAG, CAC, CAT, CAA. These rankings would indicate that the barriers are increased mostly by G neighbors, followed by C, T and A. However, this trend is not repeated for the inside rotations in the 5' direction (see Fig. 2A), or for the outside rotations. Instead, we notice that the most dominant influence on the free-energy barrier is the direction in which the 6-ring of A16 points during the rotation. Remarkably, the rotation in the 3' direction has a consistently and significantly lower barrier than in the 5' direction (see Fig. 2A and C). This is due to the asymmetrical length of the DNA sequence in each direction (see Fig. S1, i.e. the position of the rolling base along the sequence. Starting from the rotating base, A16, there are three base pairs in the 5' direction and eight in the 3' direction, which imply a difference in flexibility. We observe that, when the 6-ring rotates in the 5' direction, the 5-ring protrudes in the opposite direction toward the longer segment of DNA, which is less flexible, causing a higher barrier. In contrast, when the 6-ring rotates in the 3' direction, the 5-ring pushes toward the shorter, more flexible, segment of DNA, causing a lower barrier.

We analyze trends between the free-energy barriers and a few significant CVs. The average values of the CVs are taken from the attractors restrained at the intermediate states $s = -0.5$ and $s = 0.5$, which are close to the peaks of the free-energy barriers. In Fig. 3A, we show the distance between the two neighbors of the rolling A16, d_{NB}, in relation to the free-energy barrier. Several trends can be identified. First, d_{NB} remains at low values (< 8 Å) for all outside paths, close to those of the stable states (see Table S1); confirming that the transition with base flipping induces much less deformation of the neighbors. On the other hand, all the inside paths show larger values of d_{NB}, ranging from ~ 9 to ~ 14 Å. Among the inside paths, two distinctive trends can be observed. Inside rotations in the 3' direction reach much larger values of d_{NB}, with a relatively low increase in the free-energy barrier (0.3 kcal/mol per Å). In contrast, for the inside rotations in the 5' direction, the free-energy barrier quickly rises (1.7 kcal/mol per Å). This again highlights the role of the DNA segment's relative length in each direction of the rolling base, as well as the favored 3' direction of rotation.

We also analyze the influence of the water solvation on the free-energy barriers (see Fig. 3B). We measure the solvation using the parameter N_{water} [25], i.e. the number of water oxygens within 6 Å of the N6 atom of A16, which is the atom involved in the conserved hydrogen bond of both WCF and HG base pairs. The intermediate states of the inside pathways show a wide range of N_{water} values, from ~ 8 to ~ 21. For both the 3' and the 5' direction, an increase of N_{water} in the intermediate state relates with an increase in the free-energy barrier (0.2 kcal/mol per water molecule). Rather than a direct solvation of A16, the increase of N_{water} in the inside pathways is due to the separation of the neighboring nucleotides, as measured before by d_{NB}, which generates an opening accessible to water. On the other hand, outside pathway intermediates present larger and mostly constant values of $N_{\text{water}} \approx 24$, with the flipped conformations of A16 being stabilized by the water solvent. Our reported ranges of N_{water} for inside and outside mechanisms agree with those in [25].

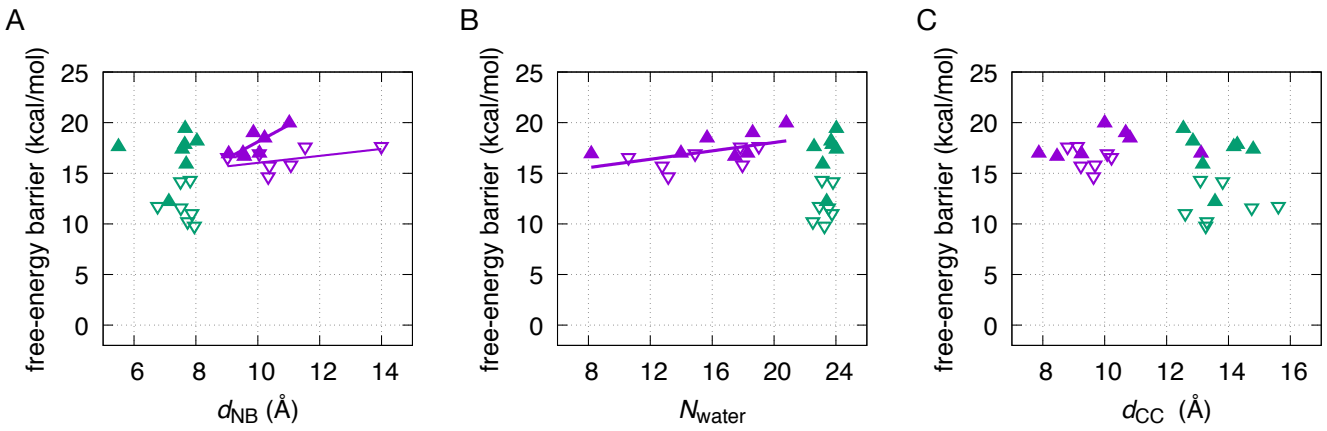

Figure 3: Trends between the WCF-to-HG free-energy barriers reported in Fig. 2 and the average value of a few selected CVs for the attractors restrained at $s = -0.5$ and $s = 0.5$. Inside paths are shown in purple and outside paths are shown in green. The barriers for the rotation with the A16 6-ring pointing in the 5' direction are represented in solid triangles pointing up, while the ones in the 3' direction are represented in outlined triangles pointing down. **(A)** Free-energy barrier vs. the distance between the neighboring bases of A16, d_{NB}. The fit of the inside 5' direction barriers is shown with a thick purple line (free-energy barrier $= 1.7$ kcal/(mol Å)$d_{\text{NB}} + 1.5$ kcal/mol). The linear fit of the inside 3' direction barriers is shown with a narrow purple line (free-energy barrier $= 0.3$ kcal/(mol Å)$d_{\text{NB}} + 12.6$ kcal/mol). **(B)** Free-energy barrier vs. the number of water oxygens within 6 Å of the N6 atom of A16, N_{water}. The linear fit of the inside 5' and 3' direction barriers is shown with a thick purple line (free-energy barrier $= 0.2$ kcal/mol $d_{\text{NB}} + 13.9$ kcal/mol). **(C)** Free-energy barrier vs. the distance between the C1' atoms of A16 and T9, d_{CC}.

Additionally, we study the relation of the free-energy barrier with the distance between the C1' atoms of A16 and T9, d_{CC} (see Fig. 3C). Inside pathway intermediates show values of d_{CC} mostly from ~ 8 to ~ 11 Å. This spans a larger range than that of the stable states (~ 9.1 to ~ 10.6 Å), indicating another possible conformational penalty for the inside pathways, but there is no clear correlation with the free-energy barrier. The intermediates of the outside pathway show larger d_{CC} values, from ~ 12 to ~ 16 Å, which are expected given that the A16·T9 base pair is completely broken. There is no clear correlation of the free-energy barriers via the outside pathways with d_{CC}, or with the other analyzed parameters.

4 CONCLUSION

Our efficient MultiPMD protocol allows, for the first time, a systematic study of HG base-pairing proclivity in diverse DNA chains. We investigate the rotation of the A16-T9 base pair in seven sequences, based on the previously studied A_6-DNA [5, 25, 23]. The seven sequences are variations of the direct neighbor of the rolling A16 in the 3' and the 5' direction of DNA (see Fig. S1). For all sequences, we study the inside and outside pathways, i.e. without and with base flipping, as depicted in Fig. 1. We obtain WCF-to-HG free-energy profiles for the inside and the outside pathway of each of the seven sequences (see Fig. S2). Our profiles are validated by: 1) the previous result for the CAA sequence reported in [36]; 2) the consistent WCF-to-HG free-energy difference between our inside and outside calculations; and 3) the relatively favored HG base pairing for the TAA sequence, which agrees with experimental evidence about the effect of A·T steps [21]. We run ~ 7 ns with twelve walkers to calculate each free-energy profile; placing our total runtime to obtain the inside and outside profiles of one sequence at ~ 168 ns. This requirement is well below the μs-long runs required for calculating previous free-energy surfaces [16, 19].

From our free-energy calculations, we observe that all sequences have a preferred outside pathway, in which the A16 rotates with its 6-ring pointing in the 3' direction (see Fig. 2). This dominant outside pathway agrees with our previous results using TPS [25], as well as with published reports using Markov state models [20]. Our result is likely to settle the debate arising for previous simulations that showed favored inside pathways, but either with no explicit solvent [5], or with a modified force field and restrained neighboring bases upon initialization [16].

We analyze the possible influence of the varied direct neighbors of A16 on the free-energy barrier for its rotation. Based on the mechanistic analysis done in [23], we expect inside paths to be more sensitive to neighbor-dependence, since they require to increase the distance between the neighbors of A16, d_{NB}, in order to accommodate the rotation. In Fig. 2C, we observe that barriers for an inside 3' rotation are increased mostly by G neighbors, followed by C, T and A. While this could point to triple-hydrogen-bonded neighbors hindering the transition, the trend is not reproduced as prominently for other pathways.

Instead, we observe that the most impactful factor for the free-energy barrier is the direction, either 3' or 5', in which the 6-ring of A16 points during the rotation. The barriers for the rotation in the 3' direction are significantly lower than their counterparts. This difference is due to the asymmetric length of the DNA chain in each direction of the transitioning base pair. The rotation of the 6-ring in the 5' direction causes the 5-ring to protrude in the 3' direction, toward the longer and more rigid side of the DNA chain, causing a higher free-energy barrier. In contrast, the 3' direction of rotation causes the 5-ring to push against the shorter and more flexible side of the DNA chain, which comes with a lower free-energy barrier. This trend is confirmed in Fig. 3A. We observe that increasing the distance between the neighbors of A16, d_{NB}, quickly raises the free-energy barrier for inside rotations in the 5' direction, while the inside rotation in the 3' direction can reach larger neighbor separations with a much lower free-energy penalty. We also observe that d_{NB} is almost undisturbed with respect to stable-state values during the outside transitions. In Fig. 3B, we analyze the number of water molecules surrounding A16, N_{water}, which is large and mostly constant for all outside intermediates. The value of N_{water} for inside intermediates is expectedly lower, and also evidences the energetically costly opening of the neighbors already described by d_{NB}. Additionally, we analyze the distance between the C1' atoms of the A16·T9 base pair, which separate significantly during the outside transitions. Nonetheless, we do not find a CV that correlates with the free-energy barrier of the outside pathways.

We believe that this work provides a robust and efficient methodology for future investigations of HG base pairing in various sequences. This includes studies with equal number of base pairs on both sides of the transitioning base, in order to elucidate the role of the neighbors exclusively, without the effect of an asymmetric length. One could also study transitions of more than one base pair, similarly to Chakraborty and Wales discrete path sampling work [37]. More importantly, our observation about the dominant influence of the relative chain length, and associated flexibility, towards the 5' and 3' directions of the rolling base, has major mechanistic implications. Protein-DNA complexes that function via HG base pairs might not only recognize, but even induce the transition by modulating the rigidity of a DNA sequence. Our simulation protocol can also enable a fast investigation of this sort of systems. Finally, the efficiency of the MultiPMD method also enables the use of higher levels of theory, such as hybrid quantum mechanics/molecular mechanics (QM/MM) studies, in which the HG transition of G·C base pairs could be simulated with flexible protonation states.

5 DATA AVAILABILITY

All the data and PLUMED input files required to reproduce the results reported in this paper are available on PLUMED-NEST (www.plumed-nest.org), the public repository of the PLUMED consortium[38], as plumID:21.033.

6 ACKNOWLEDGEMENTS

Simulations are performed on the carbon cluster at the University of Amsterdam and on the Dutch National Supercomputer cartesius [2020.015]; A.P.A.O. received funding from the Mexican National Council for Science and Technology (CONACYT).

References

[1] Watson, J. D., Crick, F. H., et al. (1953) Molecular structure of nucleic acids. *Nature,* **171**(4356), 737–738.

[2] Franklin, R. E. and Gosling, R. G. (1953) Molecular configuration in sodium thymonucleate. *Nature,* **171**(4356), 740–741.

[3] Hoogsteen, K. (1959) The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallogr.,* **12**(10), 822–823.

[4] Zhou, H., Hintze, B. J., Kimsey, I. J., Sathyamoorthy, B., Yang, S., Richardson, J. S., and Al-Hashimi, H. M. (2015) New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. *Nucleic Acids Res.,* **43**(7), 3420–3433.

[5] Nikolova, E. N., Kim, E., Wise, A. A., O'Brien, P. J., Andricioaei, I., and Al-Hashimi, H. M. (2011) Transient Hoogsteen base pairs in canonical duplex DNA. *Nature,* **470**(7335), 498–502.

[6] Imeddourene, A. B., Zargarian, L., Buckle, M., Hartmann, B., and Mauffret, O. (2020) Slow motions in A· T rich DNA sequence. *Scientific Reports,* **10**(1), 1–13.

[7] Alvey, H. S., Gottardo, F. L., Nikolova, E. N., and Al-Hashimi, H. M. (2014) Widespread transient Hoogsteen base-pairs in canonical duplex DNA with variable energetics. *Nat. Commun.,* **5**, 4786.

[8] Nikolova, E. N., Zhou, H., Gottardo, F. L., Alvey, H. S., Kimsey, I. J., and Al-Hashimi, H. M. (2013) A historical account of Hoogsteen base-pairs in duplex DNA. *Biopolymers,* **99**(12), 955–968.

[9] Nair, D. T., Johnson, R. E., Prakash, S., Prakash, L., and Aggarwal, A. K. (2004) Replication by human DNA polymerase-ι occurs by Hoogsteen base-pairing. *Nature,* **430**(6997), 377–380.

[10] Johnson, R. E., Prakash, L., and Prakash, S. (2005) Biochemical evidence for the requirement of Hoogsteen base pairing for replication by human DNA polymerase ι. *Proceedings of the National Academy of Sciences,* **102**(30), 10466–10471.

[11] Kitayner, M., Rozenberg, H., Rohs, R., Suad, O., Rabinovich, D., Honig, B., and Shakked, Z. (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nature structural & molecular biology,* **17**(4), 423.

[12] Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S.-H., Kodadek, T., and Burley, S. K. (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes & development,* **13**(24), 3217–3230.

[13] Aishima, J., Gitti, R. K., Noah, J. E., Gan, H. H., Schlick, T., and Wolberger, C. (2002) A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic acids research,* **30**(23), 5244–5252.

[14] MacKerell Jr, A. D., Banavali, N., and Foloppe, N. (2000) Development and current status of the CHARMM force field for nucleic acids. *Biopolymers: Original Research on Biomolecules,* **56**(4), 257–265.

[15] Song, K., Campbell, A. J., Bergonzo, C., de los Santos, C., Grollman, A. P., and Simmerling, C. (2009) An improved reaction coordinate for nucleic acid base flipping studies. *Journal of chemical theory and computation,* **5**(11), 3105–3113.

[16] Yang, C., Kim, E., and Pak, Y. (2015) Free energy landscape and transition pathways from Watson–Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Res.,* **43**(16), 7769–7778.

[17] Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal,* **92**(11), 3817–3829.

[18] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.,* **79**, 926–935.

[19] Yang, C., Kim, E., Lim, M., and Pak, Y. (2018) Computational probing of Watson–crick/hoogsteen breathing in a DNA duplex containing N1-methylated adenine. *Journal of chemical theory and computation,* **15**(1), 751–761.

[20] Ray, D. and Andricioaei, I. (2020) Free Energy Landscape and Conformational Kinetics of Hoogsteen Base Pairing in DNA vs. RNA. *Biophysical Journal,* **119**(8), 1568–1579.

[21] Acosta-Reyes, F. J., Alechaga, E., Subirana, J. A., and Campos, J. L. (2015) Structure of the DNA duplex d (ATTAAT) 2 with Hoogsteen hydrogen bonds. *PLoS One,* **10**(3), e0120241.

[22] Díaz Leines, G. and Ensing, B. (2012) Path finding on high-dimensional free energy landscapes. *Phys. Rev. Lett,* **109**(2), 020601.

[23] Pérez de Alba Ortíz, A., Vreede, J., and Ensing, B. (2019) The adaptive path collective variable: a versatile biasing approach to compute the average transition path and free energy of molecular transitions. In Bonomi, M. and Camilloni, C., (eds.), *Biomolecular Simulation*, chapter 11, pp. 255–290 Springer.

[24] Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., Portella, G., et al. (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods,* **13**(1), 55–58.

[25] Vreede, J., Pérez de Alba Ortíz, A., Bolhuis, P. G., and Swenson, D. W. (2019) Atomistic insight into the kinetic pathways for Watson–Crick to Hoogsteen transitions in DNA. *Nucleic acids research,* **47**(21), 11069–11076.

[26] Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.,* **53**, 291.

[27] Van Erp, T. S., Moroni, D., and Bolhuis, P. G. (2003) A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.,* **118**, 7762.

[28] Laio, A. and Parrinello, M. (2002) Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.,* **99**(20), 12562–12566.

[29] Pérez de Alba Ortíz, A., Tiwari, A., Puthenkalathil, R., and Ensing, B. (2018) Advances in enhanced sampling along adaptive paths of collective variables. *J. Chem. Phys.,* **149**(7), 072320.

[30] Macke, T. J. and Case, D. A. (1998) Modeling unusual nucleic acid structures. In Leontes, N. B. and SantaLucia, Jr., J., (eds.), *Molecular Modeling of Nucleic Acids*, pp. 379–393 American Chemical Society Washington, DC.

[31] Li, S., Olson, W. K., and Lu, X.-J. (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic acids research,* **47**(W1), W26–W34.

[32] Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.,* **91**(1-3), 43–56.

[33] Bussi, G., Donadio, D., and Parrinello, M. (2007) Canonical sampling through velocity rescaling. *J. Chem. Phys.,* **126**, 014101.

[34] Parrinello, M. and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.,* **52**, 7182–7190.

[35] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014) PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.,* **185**(2), 604–613.

[36] Hooft, F., Pérez de Alba Ortíz, A., and Ensing, B. (2021) Discovering collective variables of molecular transitions via genetic algorithms and neural networks. *J. Chem. Theo. Comput.,* **1**(1), 1–1.

[37] Chakraborty, D. and Wales, D. J. (2018) Energy landscape and pathways for transitions between Watson–Crick and Hoogsteen base pairing in DNA. *The Journal of Physical Chemistry Letters,* **9**(1), 229–241.

[38] Bonomi, M., Bussi, G., Camilloni, C., Tribello, G. A., Banáš, P., Barducci, A., Bernetti, M., Bolhuis, P. G., Bottaro, S., Branduardi, D., et al. (2019) Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods,* **16**(8), 670–673.