



## UvA-DARE (Digital Academic Repository)

### Atomistic insight into the kinetic pathways for Watson-Crick to Hoogsteen transitions in DNA

Vreede, J.; Pérez de Alba Ortíz, A.; Bolhuis, P.G.; Swenson, D.W.H.

**DOI**

[10.1093/nar/gkz837](https://doi.org/10.1093/nar/gkz837)

**Publication date**

2019

**Document Version**

Final published version

**Published in**

Nucleic Acids Research

**License**

CC BY

[Link to publication](#)

**Citation for published version (APA):**

Vreede, J., Pérez de Alba Ortíz, A., Bolhuis, P. G., & Swenson, D. W. H. (2019). Atomistic insight into the kinetic pathways for Watson-Crick to Hoogsteen transitions in DNA. *Nucleic Acids Research*, 47(21), 11069-11076. <https://doi.org/10.1093/nar/gkz837>

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)*

# Atomistic insight into the kinetic pathways for Watson–Crick to Hoogsteen transitions in DNA

Jocelyne Vreede, Alberto Pérez de Alba Ortíz, Peter G. Bolhuis and David W.H. Swenson\*

Computational Chemistry, Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Received January 15, 2019; Revised August 16, 2019; Editorial Decision September 09, 2019; Accepted September 29, 2019

## ABSTRACT

**DNA predominantly contains Watson–Crick (WC) base pairs, but a non-negligible fraction of base pairs are in the Hoogsteen (HG) hydrogen bonding motif at any time. In HG, the purine is rotated  $\sim 180^\circ$  relative to the WC motif. The transitions between WC and HG may play a role in recognition and replication, but are difficult to investigate experimentally because they occur quickly, but only rarely. To gain insight into the mechanisms for this process, we performed transition path sampling simulations on a model nucleotide sequence in which an AT pair changes from WC to HG. This transition can occur in two ways, both starting with loss of hydrogen bonds in the base pair, followed by rotation around the glycosidic bond. In one route the adenine base converts from WC to HG geometry while remaining entirely within the double helix. The other route involves the adenine leaving the confines of the double helix and interacting with water. Our results indicate that this outside route is more probable. We used transition interface sampling to compute rate constants and relative free energies for the transitions between WC and HG. Our results agree with experiments, and provide highly detailed insights into the mechanisms of this important process.**

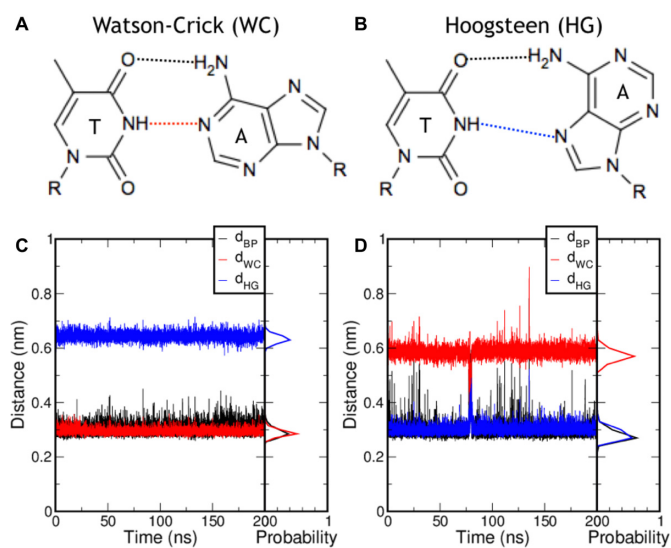
## INTRODUCTION

Six years after Watson and Crick published their model for the structure of DNA (1), Karst Hoogsteen suggested an alternative way for nucleotide bases to form hydrogen bonded pairs (2). This alternative geometry has the purine flipped ‘upside-down’, such that the 5-ring of the purine forms a hydrogen bond to the pyrimidine, rather than the 6-ring (Figure 1 A and B). In DNA, going from the Watson–Crick (WC) to Hoogsteen (HG) geometry, the flip of the purine (from *anti* in WC to *syn* in HG) requires a  $\sim 180^\circ$  rota-

tion of the base along the bond connecting the base to the sugar, known as the glycosidic bond. For more detail on the structural aspects see the Supporting Information, including Supplementary Figure S1. Raman spectroscopy of a Hoogsteen AT base pair in a crystalline environment revealed altered vibrations for atoms participating in the hydrogen bonding interactions between the base pairs (3). Advanced dispersion NMR experiments uncovered the transient presence of HG geometries in naked duplex DNA (4), suggesting that a non-negligible amount of DNA exists in the HG configuration. Recently, the sequence d(ATTAAT)<sub>2</sub> was shown to crystallize in a double helix with all the bases in HG geometry, indicating that any AT-rich sequence may contain HG base pairs (5). A survey of known protein–DNA crystal structures revealed that HG base pairing occurs in several protein–DNA complexes (6), including the human DNA polymerase hPol- $\iota$  (7), involved in replication. This particular polymerase replicates DNA via HG base pairing (7). Biochemical assays show that hPol- $\iota$  is severely inhibited by modified nucleotides that cannot adopt the HG conformation (8), thus elevating HG base pairing to a position so far only reserved for WC base pairs: they provide a basis for duplicating DNA (9). In addition, Hoogsteen base pairing plays a role in cell division and differentiation, as indicated by crystal structures of the anti-tumor protein p53 (10) and the cell differentiation regulating MAT  $\alpha 2$  homeodomain (11), both in complex with DNA containing HG base pairs. These structures provide evidence for the biological relevance of Hoogsteen base pairs, yet little is known on the mechanisms of the conversion between WC and HG base pairs.

In naked DNA, one out of every  $\sim 250$  AT base pairs occurs in the HG configuration at any given time, with a residence time of several hundreds of  $\mu\text{s}$  (4). This transient nature of the HG configuration hampers thorough experimental investigation of the mechanism of WC to HG conversion. Molecular simulation can complement experiments by providing atomistic insights at high spatial and temporal resolution. Previous work involved quantum chemical calculations which predicted that the stability of AT base pairs

\*To whom correspondence should be addressed. Tel: +33 4 72 72 88 46; Fax: +33 4 72 72 89 50; Email: dwhs@hyperblazer.net  
Present address: David W.H. Swenson, Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique and Centre Blaise Pascal, F-69342 Lyon, France.



**Figure 1.** Hydrogen bond patterns for an adenine–thymine (AT) base pair. (A) and (B) show chemical drawings in (A) Watson–Crick (WC) and (B) Hoogsteen (HG) configuration. (C) and (D) show time traces for the hydrogen bond distances of a 200 ns MD simulation in (C) WC and (D) HG, with probability histograms showing the accumulated data for (C) 9 and (D) 5 individual 200 ns MD runs. The hydrogen bond distances  $d_{BP}$ ,  $d_{WC}$  and  $d_{HG}$  are colored black, red and blue, respectively.

in HG configuration would be comparable to that of WC pairing in DNA duplexes (12). Recent discrete path sampling computations suggest that the transition of a DNA duplex containing only WC base pairs to a duplex exclusively containing HG base pairs involves several intermediates related to the conversion of single base pairs (13). However, the mechanism of single WC to HG conversions was not considered in that study (13).

Currently, a brute-force all-atom molecular dynamics (MD) investigation of the mechanism and kinetic aspects of the WC to HG transition is impossible, as the long timescales, in the order of several milliseconds, are prohibitive for such studies. Such long timescales are caused by high free energy barriers separating the stable states. One way of overcoming these barriers is by employing biasing potentials that drive the system towards the barrier region along a predefined reaction coordinate. Two earlier computational studies using biasing methods suggested that the WC to HG transition in a single base pair can occur in various ways. Using conjugate peak refinement (14) several mechanisms for the WC to HG transition were identified, yet this study did not include explicit water molecules, thus providing little evidence as to what the most likely mechanism would be (4). The existence of multiple possibilities for the WC to HG conversion was also suggested by umbrella sampling simulations (15), yet required the backbone of the DNA to be restrained to prevent distortion by the biasing potentials. While methods employing additional potentials are well suited for computing free energy barriers and other thermodynamic properties, they often fail to yield mechanistic insight at ambient conditions, as a poor choice of reaction coordinate may lead to a wrong reaction mechanism, bad sampling and a poor estimation of the rate constants.

The transition path sampling (TPS) (16) algorithm is another way to solve the timescale problem which avoids these drawbacks. TPS collects an ensemble of short reactive trajectories connecting a predefined initial and final state, without prior knowledge of the transition state region. The definitions of the initial and final states can be very simple. For instance, the WC and HG states only require order parameters related to the hydrogen bond patterns between the bases in the base pair undergoing the transition. By defining interfaces along an order parameter, which tracks the progress of the transition, transition interface sampling (TIS) enables the calculation of rates and relative free energies (17). More importantly, the speed-up gained by using TPS and related techniques is tremendous. Assuming a transition rate in the order of  $10 \text{ s}^{-1}$ , observing a single transition would require on average 100 ms of molecular dynamics. For a reasonably accurate estimation of the rate constant many transitions are required, thus scaling up the required simulation time to seconds. In contrast, when using TPS, the barrier region is sampled using MD trajectories of only tens of nanoseconds, thus providing a speed-up in the order of several million to a billion.

TPS enables analysis of the mechanism, the transition state ensemble and the reaction coordinate in terms of order parameters (18). It has been used to study the base ‘flipping’ transition in a small DNA oligomer (19), and here we apply TPS to study the transitions between the WC and HG base-pairing motifs. Our results reveal that WC to HG conversion can proceed along several mechanistic routes with varying degrees of exposure of the purine to solvent. The most prevalent route involves the adenine base flipping out of the double helix into the solvent, followed by rotation along the glycosidic bond and re-entry into the double helix. This finding contrasts with earlier computational studies, which found the route with the base remaining inside the helical confines to be more likely. These studies used either an implicit water model (4,13) or restrained the movement of the DNA backbone (15). Using TIS (17), we computed rates for the WC to HG transition, in agreement with experimental data (4). Our results provide a basis to expand the investigation of WC to HG transitions in different nucleotide sequences (20,21), including modified nucleotides (22,23), and in the presence of various other factors, such as DNA binding proteins, ions and compounds affecting DNA, such as formaldehyde (24) and triostin A (25). Furthermore, our models may aid in the interpretation of experimental data by identifying intermediate states in spectroscopic studies (26).

## MATERIALS AND METHODS

From (4), we selected nucleotide sequence 5'-CGATTTTTTGGC-3' in which the ninth pair (T9-A4') undergoes the WC to HG transition. Preparation of the system and all MD simulations were performed with GROMACS, version 4.5.3 (27), employing the AMBER03 force field (28) in combination with the TIP3P water model (29). Transition path sampling (TPS) (16,18) carries out an unbiased sampling of MD trajectories that connect predefined initial and final states, in the current case the WC and HG configurations as defined by different hydrogen bonding

patterns. TPS is a random walk through trajectory space, where new trajectories are generated from old trajectories by a shooting algorithm, and accepted with a Metropolis rule (16,18). Here, we use the one-way, flexible path length TPS algorithm as previously implemented and applied to protein systems (30). Rate constants can be computed with transition interface sampling (TIS) (17). Like TPS, this method is a Monte Carlo procedure in path space. TIS involves ‘interfaces’ or hypersurfaces for a fixed value of some order parameter  $\lambda$ , where the order parameter should be a reasonable approximation of the progress of the transition. As  $\lambda$  we used  $\arctan2(d_{WC}, d_{HG})$ , with  $d_{WC}$  and  $d_{HG}$  specific hydrogen bond distances for WC and HG respectively, to sample both the WC to HG and the HG to WC transitions. Further details on these methods can be found in the Supporting Information, including Supplementary Figure S2. Data and custom scripts are available upon request.

## RESULTS AND DISCUSSION

### Preferred mechanism for the Watson–Crick to Hoogsteen transition

The transition from Watson–Crick to Hoogsteen base pairing can occur for the pair at T9 in the sequence 5'-CGATTTTTTGGC-3' as observed in relaxation dispersion NMR experiments (4). We performed nine independent 200 ns MD simulations on this nucleotide sequence with all base pairs in Watson–Crick (WC) configuration. Figure 1C shows the time traces of the hydrogen bond distances  $d$  of one of these simulations, with  $d_{BP}$  the hydrogen bond that is formed in both WC and HG,  $d_{WC}$  the hydrogen bond that forms when the base pair is in WC configuration and  $d_{HG}$  the hydrogen bond that is formed when the base pair is in Hoogsteen (HG) configuration; see Figure 1 A and B for a molecular drawing. Figure 1 C also shows the probability histograms for the  $d_{WC}$  and  $d_{HG}$  distances of the 1.8  $\mu$ s aggregate simulation time. A similar observation can be made for the system with the T9-A4' base pair in the HG configuration, see Figure 1D for hydrogen bond distance time traces of a 200 ns MD run and probability histograms of five independent 200 ns MD simulations. Though the Hoogsteen state exhibits larger fluctuations in the hydrogen bond distances, no transitions occur within the 1  $\mu$ s aggregate simulation time. As no transitions between WC and HG are observed in the  $\mu$ s aggregate simulation time, such transitions can be considered rare and occur on timescales several orders of magnitude separated from the molecular timescale.

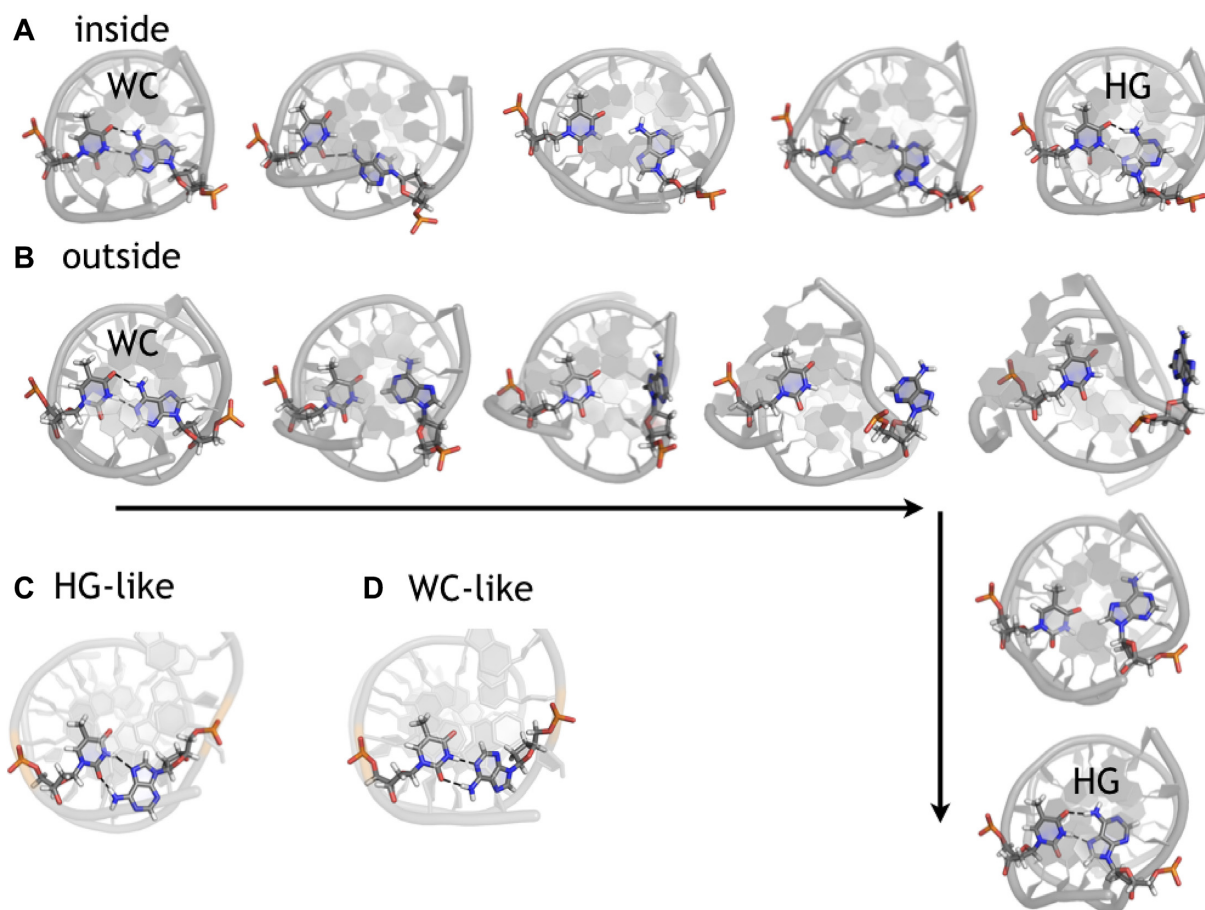
The transition path sampling (TPS) approach avoids long waiting times in stable states by focusing on the actual transition between those states. Starting from an initial trajectory that samples a WC to HG transition, TPS collects reactive trajectories connecting WC and HG by monitoring MD simulations. These are started from randomly chosen snapshots along a reactive trajectory called shooting points. If the MD simulation reaches a stable state, the trial trajectory can be accepted if it connects WC to HG. By first forcing a transition using metadynamics (31), followed by equilibration of this transition to unbiased dynamics (see the Supporting Information), we were able to investigate two types

of transitions occurring for the WC to HG conversion: *inside* and *outside*. In *inside* transitions, the adenine base converts from WC to HG while remaining inside the double helix, while in *outside* transitions the base flips outward into the solvent before entering the HG state. Figure 2A and B shows representative snapshots of an *inside* and an *outside* path respectively. Using both transition types as initial paths, we initiated several independent path sampling runs, labeled *inTPS* when started from the *inside* transition and *outTPS* when started from the *outside* transition, resulting in 1880 accepted paths for the *inTPS* simulations and 1716 accepted paths for the *outTPS* simulations. The acceptance ratio for both sets was around 30%. For both sets, the most common path length is  $\sim 1.8$  ns. For details on the path sampling statistics, see Supporting Information, Supplementary Table S1 and Supplementary Figure S4. A path tree, such as shown in the Supporting Information, Supplementary Figure S5, shows that distribution of shooting points in either WC (green) or HG (red) direction is balanced, indicating good sampling, and that paths are sufficiently decorrelated. The total number of decorrelated paths is  $\sim 125$ .

For all trajectories several geometric parameters were calculated, including the hydrogen bond distances between the bases  $d_{WC}$  and  $d_{HG}$ , the number of water oxygen atoms within 0.6 nm of atom N6 of residue 4DA'  $N_{water}$ , the base opening angle  $\theta$  and the base rolling angle  $\phi$ , see Figure 3. The latter turned out to distinguish WC from HG better than the glycosidic angle  $\chi$ , as in  $\phi$  only the orientation of the base is considered, see the Supporting Information, Supplementary Figure S3. The hydrogen bond distances  $d_{WC}$  and  $d_{HG}$  can be plotted as a single coordinate in the form of  $\lambda = \arctan2(d_{WC}, d_{HG})$ . The Supporting Information contains a detailed description of all the order parameters. The distance  $d_{WC}$  is 0.3 nm in the WC state, and 0.6 nm in HG, and vice versa for  $d_{HG}$ , which corresponds to  $\lambda = 0.46$  and 1.11 respectively.

To obtain insights into the mechanistic aspects of the WC to HG conversion, the path ensembles are projected as a path density onto two-dimensional planes defined by order parameters, such as described in the previous paragraph. See the Supporting Information for further details on how the path density is obtained. For both sets of simulations we plotted the path density projected onto  $\lambda$ , the base rolling angle  $\phi$  and the base opening angle  $\theta$  in Figure 4.  $\theta$  is close to  $0^\circ$  in both WC and HG states, indicating the adenine base is within the double helix, and pairing with thymine T9 in the opposite strand. The base rolling angle  $\phi$  has a value of around  $10^\circ$  in the WC state and ranges between  $135^\circ$  and  $-135^\circ$  including the periodic boundary at  $\pm 180^\circ$ . The path densities in the  $\lambda, \theta$  plane show striking differences for the *inTPS* and the *outTPS* simulations, see Figure 4 A and B. The path density obtained for the *outTPS* simulations shows that  $\theta$  lowers to  $-30^\circ$  before an increase occurs at  $\lambda = 0.5$ , indicating departure from the WC state. Similarly, upon leaving the HG state as indicated by  $\lambda < 1.0$ ,  $\theta$  has reached  $-20^\circ$ . When  $\lambda$  is in between  $[0.5, 0.9]$ , the base opening angle is never close to zero, but  $-30^\circ$  or lower, with the majority of paths reaching  $-60^\circ$  to  $-90^\circ$ . These values for the base opening angle represent paths with the adenine leaving the confines of the double helix. For the *inTPS* simulations, a similar pattern emerges, with the ad-



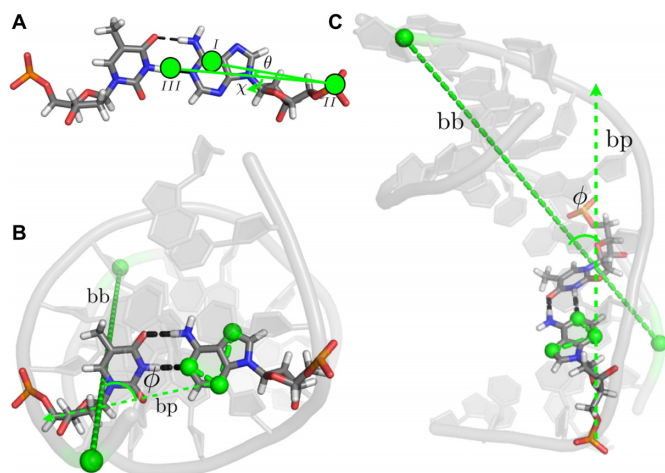


**Figure 2.** Snapshots of transition paths. Snapshot series showing (A) the *inside* and (B) the *outside* routes. Snapshots of the additional states (C) WC-like and (D) HG-like. The DNA is rendered as a grey cartoon model, with the A-T base pair undergoing the transition in sticks with carbon in gray, nitrogen in blue, oxygen in red, phosphorus in orange and hydrogen in white. The black dashed lines indicate hydrogen bonds.

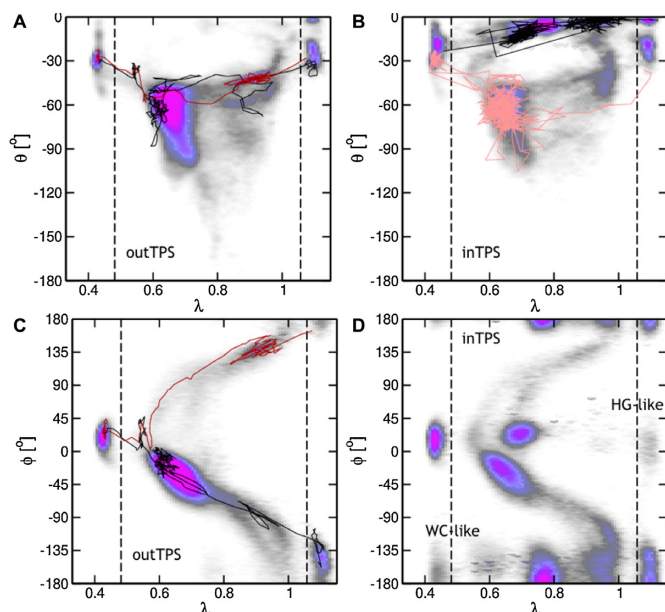
dition of density occurring at  $\theta > -30^\circ$ . This density indicates paths with the adenine staying within the double helix. Comparing the two density plots shows that the two TPS simulations have not converged to the same profile. Both the *outTPS* and the *inTPS* simulations show density indicating *outside* paths, however, the *outTPS* does not sample any *inside* paths. We conclude that the WC to HG conversion can occur in two ways, either *inside* or *outside*, and as a first qualitative assessment *outside* seems to be more likely.

The path density plots in the  $\lambda, \phi$  plane indicate that the adenine can rotate in two directions, see Figure 4C and D. An increase or a decrease in  $\phi$  indicates clockwise or counter-clockwise rotation looking along the axis of the glycosidic bond from the base to the deoxyribose group. The lines in Figure 4C indicate representative paths showing both types of rotation. An intermediate conformation at around  $\lambda = 0.7$ ,  $\phi = [-60^\circ, 0^\circ]$  occurs, indicating the counter-clockwise route (black line in Figure 4C) seems to be favored over the clockwise rotation, or that a partially rotated intermediate state is visited. The *outTPS* simulations sample only *outside* paths. The *inTPS* simulations must sample *inside* paths, but the path densities show striking similarities to the *outTPS* density. However, additional

density occurs at  $\lambda = 0.7$ ,  $\phi = [0^\circ, +45^\circ]$  and  $\lambda = 0.75$ ,  $\phi = [-180^\circ, -135^\circ]$  and  $[135^\circ, 180^\circ]$  representing rotation within the double helix. No configurations in between these ranges were observed, indicating that this rotation occurs faster than the time between consecutive frames in the trajectories, which is 5 ps. Therefore, we could not identify *inside* transitions as clockwise or counter clockwise. However, the path length distributions of both sets of TPS simulations peak at 1.5 ns, which is much longer than the time required for rotation of the base, see Supplementary Figure S4 in the Supporting Information. The rotation of the base seems to be the fastest step in the *inside* route, preceded by hydrogen bond breakage and followed by hydrogen bond reformation, both of which are slower. Finally, additional density seems to occur in Figure 4D at  $\lambda < 0.48$  and at  $\lambda > 1.1$ , indicating conformations that are either WC-like, or HG-like, respectively. Snapshots of these conformations are shown in Figure 2C and D for HG-like and WC-like respectively. The HG-like conformation shows the adenine in WC orientation, but forming a different hydrogen bonding pattern to the thymine, involving oxygen atom O2 rather than O4 of the thymine. Similar for the WC-like conformation, the adenine is in HG conformation, but forming a hydrogen bond to the oxygen atom O2.



**Figure 3.** Schematic illustration of relevant order parameters. (A) The glycosidic angle  $\chi$  and the base opening angle  $\theta$ . (B, C) The base rolling angle  $\phi$  from the top (B) and side (C) of the DNA.  $\chi$  is the dihedral angle between atoms O4', C1', N9, C4 in the rolling base 4DA'.  $\theta$  is the angle between I, the center of mass of the rolling base, II, the center of mass of the phosphate groups in nucleotides 4DA' and 5DA', directly neighboring the nucleotide containing the rolling base and III, the center of mass of the base pairs 8DT-5DA' and 10DG-3DC', directly neighboring the base pair containing the rolling base. This definition is taken from Ref. (32).  $\phi$  is the angle between the vectors  $bp$  and  $bb$ , which is a proxy for the long axis of the DNA. Vector  $bb$  connects the phosphorus atoms in residues DG11 and DC11'. Vector  $bp$  is the vector resulting from the cross product of the vectors connecting atoms N3 and N1 and N3 and N7 in the rolling base.



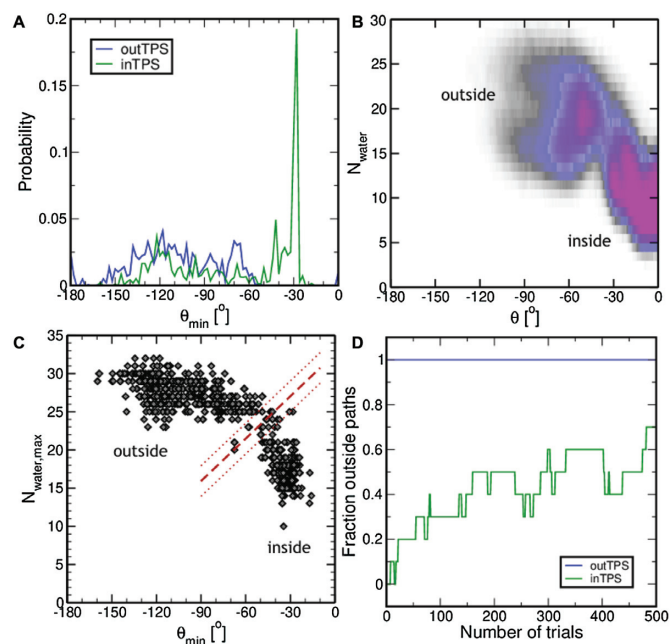
**Figure 4.** Path density plots in the  $\lambda$ ,  $\theta$  plane (A, B) and the  $\lambda$ ,  $\phi$  plane (C, D) for the (A, C) *outTPS* and (B, D) the *inTPS* simulations. The color of the path density ranges from white (no paths) to magenta (highest path density). The black, red and pink lines indicate representative paths as a guide to the eye. The dashed lines indicate the boundaries of the stable states as projected onto  $\lambda$ .

To distinguish between *inside* and *outside* paths we plotted histograms of  $\theta_{\min}$ , the minimum value of  $\theta$ , in each path for the *inTPS* and the *outTPS* simulations, see Fig-

ure 5A. The paths in the *outTPS* simulations do not sample  $\theta_{\min} > -50^\circ$ , while the paths in the *inTPS* simulations go up to  $\theta_{\min} = -22^\circ$ . In the *outside* route, the adenine residue undergoing the transition becomes exposed to solvent. An alternative way of distinguishing *inside* and *outside* paths could be provided by counting the number of water molecules around the rolling base. The Supporting Information, in particular Supplementary Figures S6, S7 and related text, includes a discussion about tuning the calculation of  $N_{\text{water}}$  such as to give a sufficiently clear separation of *inside* and *outside* channels. Figure 5B shows the path density of the *inTPS* simulations projected onto  $\theta$  and the number of water oxygen atoms within a radius of 0.6 nm of atom N6 of the rolling adenine,  $N_{\text{water}}$ . The path density shows two channels, one centered at  $N_{\text{water}} = [10]$  and  $\theta > -34^\circ$ , and one centered at  $N_{\text{water}} = 20$  and  $\theta = -50^\circ$ . These channels are not fully separated. We then plotted  $\theta_{\min}$  and the maximum value for  $N_{\text{water}}$ ,  $N_{\text{water,max}}$  in Figure 5C. This scatter plot suggests that there are two sets of paths, which overlap. These sets can be separated by a linear function  $N_{\text{water,max}} = 0.185\theta_{\min} + 32$ , as indicated by the red line in Figure 5C. Paths below this line are *inside* paths, and paths that are above the dividing line are *outside* paths. This dividing line was fitted to two points in the region of the lowest density in between *inside* and *outside*. We also defined a region in which paths are classified as neither *inside* nor *outside* with a margin of  $N_{\text{water,max}} = 2$  above and below the dividing line, see Figure 5C. The Supporting Information contains a discussion on why a margin of  $N_{\text{water,max}} = 2$  is sufficient. With this definition for *inside* and *outside*, we can now track whether the TPS simulations switch from sampling *inside* paths to *outside* paths and vice versa. Note that if a path is classified as neither *inside* nor *outside* it is assigned the label of the previously accepted path. The path length distribution, shown in Supplementary Figure S4 in the Supporting Information, shows that *inside* paths are slightly shorter than *outside* paths. Both the *inTPS* and *outTPS* simulations consisted of 10 independent runs. Figure 5D shows the fraction of *outside* paths as a function of the number of trials. The *outTPS* simulations starts with all runs containing *outside* paths and never switches to sampling *inside* paths. For the *inTPS* simulations the fraction of *outside* paths is zero at the beginning, and increases to 0.7 after 500 trials, indicating that most runs switch to sampling *outside* paths. This observation strongly suggests that the *outside* path is the predominant mechanism for the WC to HG transition.

We can quantify this statement by a rough statistical analysis. We assume that the switching process is a Markov process with a probability  $p(\text{inside} \rightarrow \text{outside} | \text{TP}_{\text{in}})$  to undergo the switching from *inside* to *outside*, and the probability  $p(\text{outside} \rightarrow \text{inside} | \text{TP}_{\text{out}})$  for the reverse process. When starting in the *inside* channel, this process relaxes to the equilibrium distribution which would be dominated by the *outside* paths. We can estimate the relaxation time (in terms of trials) by fitting the green curve in Figure 5D to an exponential function,  $p_{\text{outside}} = 1 - e^{-t/\tau}$ , which roughly gives a relaxation time of  $\tau = 350$  trials moves. Thus we can conclude that  $p(\text{inside} \rightarrow \text{outside} | \text{TP}_{\text{in}}) \approx 1/350$ .

To show that the *outside* channel is more likely than the *inside* one, we need to show that  $p(\text{inside} \rightarrow \text{outside} | \text{TP}_{\text{in}}) > p(\text{outside} \rightarrow \text{inside} | \text{TP}_{\text{out}})$ .

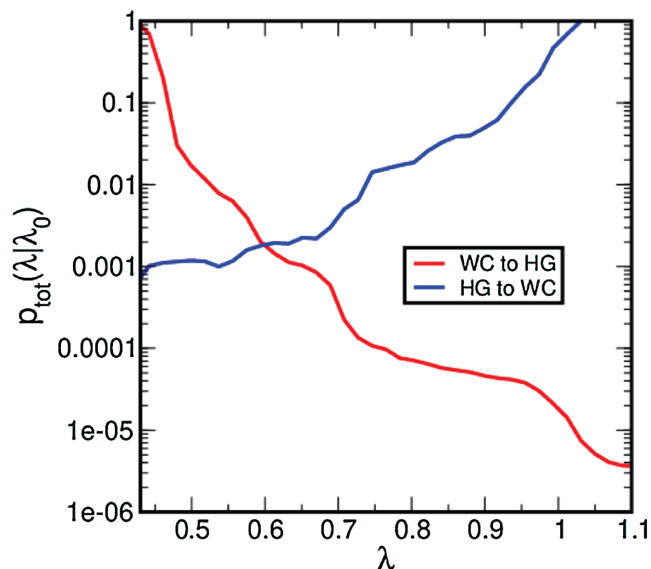


**Figure 5.** Analysis of outside and inside paths. (A) Path-weighted probability histogram of the minimum base opening angle  $\theta_{\min}$  for the *inTPS* and *outTPS* simulations. (B) Path density plot in the  $\theta$ ,  $N_{\text{water}}$  plane for the *inTPS* simulations. The color of the path density ranges from white (no paths) to magenta (highest path density). (C) Scatter plot of  $\theta_{\min}$  and  $N_{\text{water,max}}$  for each accepted path in the *inTPS* simulations. The dashed red line indicates the dividing line between *inside* and *outside* paths. The dotted red lines indicate the region in which paths are considered to be neither *inside* nor *outside*. (D) Fraction of *outside* paths per trial.

Given the evidence that we have observed seven out of ten switches from *inside* to *outside*, while at the same time the number of observed *outside* to *inside* switches remains zero, a simple Bayesian analysis yields the probability that this is true equals  $\sim 0.99948$ , see the Supporting Information. Thus, the odds in favor of the hypothesis that the *outside* channel is more stable than the *inside* channel, are  $\sim 2000$  against 1.

### Rate constants for the Watson–Crick to Hoogsteen transition

Transition Interface Sampling (TIS) is a path sampling technique to calculate reaction rates. TIS requires the definition of interfaces along an order parameter that is a reasonable estimate of the reaction coordinate (17). To calculate the rates of the Watson–Crick to Hoogsteen transition using TIS, the same stable states definitions as used in the TPS simulations can be used. However, neither the glycosidic angle  $\chi$  nor the base rolling angle  $\phi$  can be used as an order parameter to track the transition, even though both angles have distinctly different values in the WC and the HG states. In particular for the outside transition,  $\chi$  and  $\phi$  can take any value as the adenine group can rotate freely once it has left the interior of the double helix. Instead we used the  $\lambda = \arctan2(d_{\text{WC}}, d_{\text{HG}})$  parameter, as this provides distinction between the stable states and a sufficient description of the progression of the transition. To determine whether  $\lambda$  is a reasonable order parameter to track the progression of the WC to HG conversion, and vice versa, we plotted path den-



**Figure 6.** Total crossing probability for the WC to HG and HG to WC transitions as a function of  $\lambda$ . These curves are obtained by combining the crossing probabilities computed for each interface using the weighted histogram analysis method (33). More details are given in the Supporting Information.

sity profiles for each interface in the  $\lambda$ ,  $\theta$  and the  $\lambda$ ,  $\phi$  planes, in the Supporting Information, Supplementary Figures S9–S12. Interfaces close to the final state show  $\theta$  far from zero, indicating that those interfaces sample *outside* transitions. Interfaces close to the initial states do not sample such large ranges for  $\theta$ , indicating that the distinction between *inside* and *outside* becomes relevant only after a certain progression of the transition. The path density profiles suggest that the choice for going via the *inside* or the *outside* channel lies at  $0.65 < \lambda < 0.75$ .

To obtain the rate, we determined the crossing probability at  $\lambda = 1.0$  for WC to HG, and at  $\lambda = 0.5$  for HG to WC, as well as the probability to reach the HG or the WC state, respectively, see Figure 6. The slope of the log of the crossing probability  $p_{\text{tot}}(\lambda_i|\lambda_0)$  as a function of  $\lambda$  for the WC to HG transition changes several times, indicating that even though it is possible to sample the WC to HG transition along the  $\arctan2(d_{\text{WC}}, d_{\text{HG}})$  order parameter, there are regions in which the system tends to stay longer. These regions occur where the slope is flatter, at  $\lambda < 0.5$ ,  $\lambda = 0.7$  and  $\lambda = 0.95$ . These regions correspond to the high density regions in the TPS path density plots, see Figure 4.

At  $\lambda = 1.0$ ,  $p_{\text{tot}}(\lambda_i|\lambda_0) = 1.5 \times 10^{-5}$ . Multiplying this with the probability to reach the HG state  $p(\text{HG}|\lambda_i) = 0.09$  and by the flux out of the WC state  $\Phi_{\text{WC}} = 5.5 \times 10^8 \text{ s}^{-1}$ , we computed the rate for the WC to HG transition to be  $k_{\text{WC} \rightarrow \text{HG}} = 742 \text{ s}^{-1}$ . Similarly we computed the rate for the HG to WC transition, using  $p_{\text{tot}}(\lambda_i|\lambda_0) = 1.2 \times 10^{-3}$  at  $\lambda = 0.5$ , the probability to reach the WC state  $p(\text{WC}|\lambda_i) = 0.94$  and the flux out of the HG state  $\Phi_{\text{HG}} = 1.45 \times 10^9 \text{ s}^{-1}$ , resulting in a rate of  $k_{\text{HG} \rightarrow \text{WC}} = 1.6 \times 10^6 \text{ s}^{-1}$ . Table 1 shows the rates as obtained from TIS simulations and from carbon dispersion relaxation NMR experiments (4). The rates computed by TIS are consistently 50 times faster than the experimentally determined rates. Taking the ratio of the two



**Table 1.** Rate constants

	experiment	TIS
$k_{WC \rightarrow HG}$ ( $s^{-1}$ )	$14.2 \pm 1.03$	742
$k_{HG \rightarrow WC}$ ( $s^{-1}$ )	3670	$1.6 \times 10^5$
$\Delta G$ ( $k_B T$ )	5.5	5.4

The free energy difference  $\Delta G$  is calculated as the ratio of the rate constants  $k_{WC \rightarrow HG}/k_{HG \rightarrow WC}$ .

rates, which is the same as the Gibbs free energy difference between the WC and HG states, shows agreement between the computed and experimental rates.

One explanation for the fact that the rates as computed by TIS are faster than the experimental data could be related to the force field we have used. To check this we performed TPS simulations of the *outside* transition with a force field containing improved parameters for nucleic acids, parmbsc1 (34); see the Supporting Information, in particular Supplementary Figure S8. These simulations give similar results as the *outTPS* simulations, indicating that the force field used in this study provides a sufficiently accurate description. The dihedral potential for the glycosidic angle is stiffer in the parmbsc1 force field, which results in higher barriers for the rotation along this bond, and may result in lower rates. Another issue with these force fields in general is the description of the interactions between the atoms. We have used a fixed-charge pairwise additive force field to keep the simulations with an aggregate simulation time of several  $\mu s$  tractable. However, this simplified description may have resulted in an overestimation of the rate constants.

## CONCLUSION

We have presented TPS simulations describing the transitions between the well-known Watson–Crick base-pairing motif and the Hoogsteen base-pairing motif. Recent experiments have shown that HG pairing, in which the purine is ‘upside-down’ relative to the WC motif, makes up a non-negligible fraction of DNA at physiological conditions. Using TPS, we obtained over 3500 samples of the transition between WC and HG, providing valuable insight into the mechanism of this transition. There are two main ways that the WC to HG transition can occur: either the purine stays in the double helix and rolls over directly, the *inside* route, or it goes *outside*, becomes exposed to solvent, rolls over, and re-enters the double helix. Thorough analysis of the relevant order parameters and a rough Bayesian analysis showed that this *outside* mechanism is preferred. We computed the rate constants for the WC to HG conversion and vice versa using TIS, as well as the relative free energies for the WC and HG states. Our results are in agreement with experimental data.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

Sander Roet provided background literature on the relevance of HG base pairs.

## FUNDING

All simulations are performed on the Dutch National Supercomputer cartesius [SH-013-15 and SH-217-12]; JV received funding from Netherlands Organization for Scientific Research (NWO) via Athena Grant [700.58.802]; A.P.A.O. received funding from the Mexican National Council for Science and Technology (CONACYT); P.G.B. and D.W.H.S. acknowledge support from the European Union’s Horizon 2020 research and innovation program [676531] (project E-CAM). Funding for open access charge: University of Amsterdam.

*Conflict of interest statement.* None declared.

## REFERENCES

- Watson, J.D. and Crick, F.H.C. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
- Hoogsteen, K. (1959) The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, **12**, 822–823.
- Strobel, J.L. and Scovell, W.M. (1980) Laser Raman spectroscopy of a complementary base pair in the Hoogsteen configuration. *Biochim. Biophys. Acta (BBA)*, **608**, 201–214.
- 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**, 498–502.
- Acosta-Reyes, F.J., Alechaga, E., Subirana, J.A. and Campos, J.L. (2015) Structure of the DNA duplex d(ATTAAT)<sub>2</sub> with Hoogsteen hydrogen bonds. *PLoS One*, **10**, e0120241.
- 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**, 3420–3433.
- Nair, D.T., Johnson, R.E., Prakash, S., Prakash, L. and Aggarwal, A.K. (2004) Replication by human DNA polymerase- $\alpha$  occurs by Hoogsteen base-pairing. *Nature*, **430**, 377–380.
- Johnson, R., Prakash, L. and Prakash, S. (2005) Biochemical evidence for the requirement of Hoogsteen base pairing for replication by human DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10466–10471.
- 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**, 955–968.
- 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. *Nat. Struct. Mol. Biol.*, **17**, 423–429.
- Aishima, J. (2002) A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.*, **30**, 5244–5252.
- Gould, I.R. and Kollman, P.A. (1994) Theoretical investigation of the hydrogen bond strengths in guanine-cytosine and adenine-thymine base pairs. *J. Am. Chem. Soc.*, **116**, 2493–2499.
- Chakraborty, D. and Wales, D.J. (2017) Energy landscape and pathways for transitions between Watson–Crick and Hoogsteen base pairing in DNA. *J. Phys. Chem. Lett.*, **9**, 229–241.
- Fischer, S. and Karplus, M. (1992) Conjugate peak refinement: An algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem. Phys. Lett.*, **194**, 252–261.
- 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**, 7769–7778.
- Dellago, C., Bolhuis, P.G., Csajka, F.S. and Chandler, D. (1998) Transition path sampling and the calculation of rate constants. *J. Chem. Phys.*, **108**, 1964–1977.
- 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**, 7762–7774.
- Dellago, C., Bolhuis, P.G. and Geissler, P.L. (2002) Transition path sampling. *Adv. Chem. Phys.*, **123**, 1–78.



19. Hagan, M.F., Dinner, A.R., Chandler, D. and Chakraborty, A.K. (2003) Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 13922–13927.
20. Abrescia, N.G.A., Thompson, A., Huynh-Dinh, T. and Subirana, J.A. (2002) Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 2806–2811.
21. 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.
22. Isaksson, J., Zamaratski, E., Maltseva, T.V., Agback, P., Kumar, A. and Chattopadhyaya, J. (2001) The first example of a Hoogsteen basepaired DNA duplex in dynamic equilibrium with a Watson–Crick basepaired duplex—a structural (NMR), kinetic and thermodynamic study. *J. Biomol. Struct. Dyn.*, **18**, 783–806.
23. Sathyamoorthy, B., Shi, H., Zhou, H., Xue, Y., Rangadurai, A., Merriman, D.K. and Al-Hashimi, H.M. (2017) Insights into Watson–Crick/Hoogsteen breathing dynamics and damage repair from the solution structure and dynamic ensemble of DNA duplexes containing m1A. *Nucleic Acids Res.*, **45**, 5586–5601.
24. Bohnuud, T., Beglov, D., Ngan, C., Zerbe, B., Hall, D., Brenke, R., Vajda, S., Frank-Kamenetski, M. and Kozakov, D. (2012) Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic Acids Res.*, **40**, 7644–7652.
25. Singh, U.C., Pattabiraman, N., Langridge, R. and Kollman, P.A. (1986) Molecular mechanical studies of d(CGTACG)<sub>2</sub>: complex of triostin A with the middle A - T base pairs in either Hoogsteen or Watson–Crick pairing. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 6402–6406.
26. Stelling, A.L., Xu, Y., Zhou, H., Choi, S.H., Clay, M.C., Merriman, D.K. and Al-Hashimi, H.M. (2017) Robust IR-based detection of stable and fractionally populated G-C+ and A-T Hoogsteen base pairs in duplex DNA. *FEBS Lett.*, **591**, 1770–1784.
27. Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D. *et al.* (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845–854.
28. Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T. *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**, 1999–2012.
29. 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.
30. Vreede, J., Juraszek, J. and Bolhuis, P.G. (2010) Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. *Biophys. J.*, **98**, 174a.
31. Laio, A. and Parrinello, M. (2002) Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 12562–12566.
32. 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. *J. Chem. Theory Comput.*, **5**, 3105–3113.
33. Swenson, D.W.H., Prinz, J.H., Noe, F., Chodera, J.D. and Bolhuis, P.G. (2019) OpenPathSampling: a Python framework for path sampling simulations. 1. Basics. *J. Chem. Theory Comput.*, **15**, 813–836.
34. Ivani, I., Dans, P.D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A. *et al.* (2015) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.