Bioinformatics, 34(1), 2018, 41–48
doi: 10.1093/bioinformatics/btx516
Advance Access Publication Date: 25 September 2017
Original Paper

OXFORD

Structural bioinformatics

In silico structural modeling of multiple epigenetic marks on DNA

Konrad Krawczyk[1,2], Samuel Demharter[1], Bernhard Knapp[2,3], Charlotte M. Deane[2] and Peter Minary[1,*]

[1]Department of Computer Science, Oxford University, OX1 3QD Oxford, UK, [2]Department of Statistics, Oxford University, OX1 3LB Oxford, UK and [3]Faculty of Medicine and Health Sciences, International University of Catalonia, Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 1, 2017; revised on July 11, 2017; editorial decision on August 8, 2017; accepted on September 22, 2017

Abstract

There are four known epigenetic cytosine modifications in mammals: methylation (5mC), hydroxymethylation (5hmC), formylation (5fC) and carboxylation (5caC). The biological effects of 5mC are well understood but the roles of the remaining modifications remain elusive. Experimental and computational studies suggest that a single epigenetic mark has little structural effect but six of them can radically change the structure of DNA to a new form, F-DNA. Investigating the collective effect of multiple epigenetic marks requires the ability to interrogate all possible combinations of epigenetic states (e.g. methylated/non-methylated) along a stretch of DNA. Experiments on such complex systems are only feasible on small, isolated examples and there currently exist no systematic computational solutions to this problem. We address this issue by extending the use of Natural Move Monte Carlo to simulate the conformations of epigenetic marks. We validate our protocol by reproducing in silico experimental observations from two recently published high-resolution crystal structures that contain epigenetic marks 5hmC and 5fC. We further demonstrate that our protocol correctly finds either the F-DNA or the B-DNA states more energetically favorable depending on the configuration of the epigenetic marks. We hope that the computational efficiency and ease of use of this novel simulation framework would form the basis for future protocols and facilitate our ability to rapidly interrogate diverse epigenetic systems.

Availability and implementation: The code together with examples and tutorials are available from http://www.cs.ox.ac.uk/mosaics

Contact: peter.minary@cs.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The mammalian genome has four known epigenetic modifications of cytosine: methylation (5mC), hydroxymethylation (5hmC), formylation (5fC) and carboxylation (5caC) (Bhutani *et al.*, 2011; Wu and Zhang, 2015) (see Supplementary Fig. S1). Transcription regulation by cytosine methylation has been studied extensively and it is known to reduce gene expression by preventing transcription factor binding (Siegfried *et al.*, 1999). The specific roles of the remaining three oxidized epigenetic marks are still mostly unknown (Shen *et al.*, 2013; Sun *et al.*, 2014). It has been hypothesized that the oxidized epigenetic marks might be the intermediate steps in the process of demethylation (Guo *et al.*, 2011) (Supplementary Fig. S1). However it has also been suggested that oxidized epigenetic marks might perform regulatory roles in their own right (Song *et al.*, 2013) or even be implicated in disease (Sun *et al.*, 2014; Kroeze *et al.*, 2015). It was demonstrated that in some scenarios 5hmC reverses

the stabilizing effects of 5mC, which might be a factor in its ability to modulate transcription (Thalhammer *et al.*, 2011). There is also a host of regulatory proteins, which bind to 5fC selectively, suggesting a role of its own in chromatin organization and transcriptional regulation (Iurlaro *et al.*, 2013; Muenzel *et al.*, 2010; Spruijt *et al.*, 2013).

Epigenetic marks are often located in CpG islands, which are long-stretches of DNA with repeated CpG dinucleotides (Song *et al.*, 2013; Wang *et al.*, 2014). Such localization of multiple epigenetic marks might have an effect on chromatin structure (Raiber *et al.*, 2015). Recently three high-resolution structures have been released to the Protein Data Bank (PDB) (Berman *et al.*, 2000), which demonstrate the effect of two 5mC or two 5hmC or six 5fC modifications on DNA dodecamers (Drew *et al.*, 1981). The 5mC and 5hmC modifications did not affect the geometry of B-DNA significantly (Lercher *et al.*, 2014; Renciuk *et al.*, 2013). However, a recent crystallographic (X-ray) study demonstrated that the presence of a formylated CpG island (5fCpG)3 can lead to significant helical underwinding and change in curvature, giving rise to a new structural form of DNA, F-DNA (Supplementary Fig. S2) (Raiber *et al.*, 2015).

As demonstrated by Raiber *et al.* (2015), in order to fully discern the structural roles of 5mC, 5hmC and 5fC, one should study collections of epigenetic marks. We refer to such a collection as 'epigenetic makeup', which is a specific arrangement of epigenetic marks on a particular stretch of DNA. Interrogating all possible epigenetic makeups on a given piece of DNA for thermodynamic and structural properties would require synthesizing all such epigenetic makeup variants under identical conditions. Therefore, such investigation of all possible epigenetic makeups on a given structural template (B-DNA, F-DNA) is very resource-intensive. Computational approaches to study epigenetic systems have focused on computationally expensive Density Functional Theory (DFT) (Acosta-Silva *et al.*, 2010; Carvalho *et al.*, 2015; Yusufaly *et al.*, 2013) or Molecular Dynamics (MDs) simulations (Carvalho *et al.*, 2014, 2015; Carson *et al.*, 2016 ; Derreumaux *et al.*, 2001; Frauer *et al.*, 2011; Severin *et al.*, 2011; Wanunu *et al.*, 2011). Carefully studying all possible combinations of epigenetic makeups via DFT is computationally prohibitive; however, doing so with MD is becoming increasingly possible given right computational resources which will surely benefit the field in the future.

We propose a formalized framework for addressing the epigenetic problem and use a streamlined computational approach to gain a structural insight as to how variability in epigenetic makeup affects the thermodynamic preference for different DNA structures (e.g. B-DNA or F-DNA). The first goal of developing such a computational framework is to provide an accessible way to rapidly simulate different epigenetic makeups on an arbitrary DNA structure. Our second goal is to provide a versatile framework which allows for use of different sampling methodologies and force fields (Dans *et al.*, 2017; Ivani *et al.*, 2016; Sponer *et al.*, 2013) in order to provide a unified platform to assess different parametrizations and computational methodologies relating to DNA/RNA simulations with or without (multiple) epigenetic marks. For this purpose, we employed the molecular simulation software MOSAICS. It has been demonstrated previously that natural move Monte Carlo (NMMC) technology as implemented in MOSAICS can efficiently map the structural space of nucleic acids (Minary and Levitt, 2010; Sim *et al.*, 2012). Using NMMC we also predicted nucleosome occupancy in the context of DNA methylation (Minary and Levitt, 2014). Here, we extend NMMC to be able to simulate a variety (e.g. 5mC, 5hmC and 5fC) of epigenetic marks. We perform simulations on recently released high-resolution structures of the DNA dodecamers with 5mC, 5hmC and 5fC modifications. We validate our protocol by reproducing a variety of experimental observations (Lercher *et al.*, 2014; Raiber *et al.*, 2015). These include configurations of 5hmC and 5fC as well as the role of 5fC in the stability of the F-DNA form. Furthermore, we show that in our simulations a formylated CpG island is more energetically favorable in the F-DNA form than in the B-DNA form.

These proofs of concept results demonstrate the utility of our simulation framework, specifically in relation to the novel field of structural epigenetics. Given the scale of the structural space introduced by the epigenetic marks, the force-fields and sampling methodologies would require future refinement to correctly address new DNA forms (Dans *et al.*, 2017; Raiber *et al.*, 2015). MOSAICS is capable of simulating DNA/RNA structures with or without epigenetic marks, and a large variety of conformational sampling methodologies. Therefore, we hope that it can provide a stable software framework to gain immediate insight into molecular systems and to benchmark different force-fields and sampling methodologies used to model epigenetic marks on nucleic acid structures.

2 Materials and methods

2.1 MOSAICS framework

MOSAICS is a suite of simulation protocols for nucleic acids and proteins. MOSAICS is a versatile framework that allows the testing of diverse protocols involving different combinations of force fields and sampling methodologies. As opposed to methods such as MDs, methods available in the MOSAICS framework allow for using reduced degrees of freedom to sample the structural space efficiently. Users can choose from a suite of sampling techniques [e.g. Parallel Tempering (Hansmann, 1997), Normal Modes (Hayward and de Groot, 2008)] and different force fields [e.g. Amber (Pérez *et al.*, 2007), CHARMM (MacKerel *et al.*, 1998)]. Simulations can be performed with explicit waters or using implicit solvent (Case, 2001). Simulations using MOSAICS have been demonstrated to accurately model large RNA ensembles (Sim *et al.*, 2012), molecular machines such as chaperonines (Zhang *et al.*, 2012), to agree with previous MD and NMR studies (Krawczyk *et al.*, 2016), simulate MHC peptide detachment (Knapp *et al.*, 2016), simulate diabodies (Moraga *et al.*, 2015) and model the functional motions of biological systems (Demharter *et al.*, 2016). Recently, it was shown that using MOSAICS one can reliably model nucleosome occupancy *in vitro* by explicitly simulating nucleosomal DNA accommodating a large number (tens of thousands) of sequences with and without 5mC epigenetic marks (Minary and Levitt, 2014). Following on from this study, the current MOSAICS setup was extended with other epigenetic marks to allow comprehensive simulations of structural epigenetic systems.

2.2 Extending MOSAICS to include 5hmC and 5fC marks

Previous versions of MOSAICS were not adapted towards epigenetic marks. In version 3.9.2 we have now added new topology definitions to allow for the simulations of 5hmC and 5fC marks. The relevant files can be found at MOSAICS Web site (www.cs.ox.ac.uk/mosaics).

2.3 DNA Dodecamer model generation

For the comparative study of B-DNA vs F-DNA stability it was necessary to obtain a hydrated B-DNA model. The Dodecamer

sequence in this case was this of 4QKK: d(CTACGCGCGTAG). The B-DNA models were created using the Make-NA service (http://structure.usc.edu/make-na/server.html.). Epigenetic marks were added by threading atoms to structure as described previously (Minary and Levitt, 2014) and detailed instructions are available through our website (www.cs.ox.ac.uk/mosaics). The orientation of the formyl in the model was identical to the orientation of formyl present in the structure of 4QKK.

2.4 Natural move Monte Carlo

The Simulations were carried out at 300 K temperature using MOSAICS software package (www.cs.ox.ac.uk/mosaics). We used the Amber ParmBSC0 (37) potential as this energy function in combination with a distance dependent dielectric model for bulk solvent effect showed good correlation with experimental results in a previous MOSAICS application to epigenetics (34). A newer version of Amber is now available, ParmBSC1 (Ivani *et al.*, 2016). ParmBSC0 and ParmBSC1 share the same non-bonded interactions (Dans *et al.*, 2017), which are primarily leading to the results we obtain. To model water mediated H-bond interactions (Case, 2001) we used explicit waters, which were either present as crystal waters (F-DNA) or added using VMD (B-DNA). Simulations were performed in the all-atom representation. The majority of the simulations were performed on the Hartree center (Science & Technology Facilities Council) as part of the first Xeon Phi Access Programme.

2.5 Epigenetic makeup

The inverse epigenetic problem depends on the set of considered types of epigenetic modifications, Ω_e and the set of possible locations, Ω_l that may host a modification. For example, assuming that F-DNA is the template one may choose two 'types' of epigenetic modifications, $\Omega_e = \{f, \varnothing\}$ (formyl and lack of modification) at all 6 native locations in the CpG island, $(CpG)_3$. Therefore, $|\Omega_l| = 6$ and $|\Omega_e| = 2$ ($|.|$ refers to the size of the set), giving rise to 2^6 possibilities. Unfortunately, energies related to individual epigenetic makeups may not be directly comparable because each epigenetic makeup defines a new system. Instead, each epigenetic makeup can also be threaded to a reference structure (of the same system) that assumes a conformation distinct from the template. For F-DNA the reference structure can naturally be chosen as a straight B-DNA form. The energy difference between the template and reference structures accommodating the same epigenetic makeup is a good indicator of how much the epigenetic makeup stabilizes the template. In our example we can calculate the energies for all 2^6 epigenetic makeups (out of which 24 are unique if we disregard symmetry) on both F-DNA (template) and B-DNA.

2.6 Energy minimization

We performed the energy comparison between B-DNA and F-DNA structures using the conjugate gradient minimization available in MOSAICS and the same force-field described above (see Section 2.4). For a given makeup of 5fC marks on B-DNA and F-DNA we generated a 5 Å hydration shell around them using VMD (Humphrey *et al.*, 1996). We removed waters at random to arrive at the same number of water molecules in both B-DNA and F-DNA variants (removed from the structure that had more waters). The biologically inherent hydration of B-DNA and F-DNA structures would be clearly different, and the size of the hydration shell assures reliable coverage of the surfaces of the DNA molecules. Equating the numbers of waters assured that the numbers of atoms between F-DNA and B-DNA systems are the same and thus their energies would be comparable. This operation was repeated 15 times.

Supplementary Figure S3 suggests that a particular epigenetic makeup (Γ) can be associated with an energy difference, $\Delta E_\Gamma = E_\Gamma$ (F-DNA) $- E_\Gamma$ (B-DNA), which represents its stabilizing effect on the template (F-DNA) with respect to the reference (B-DNA). Quantifying the effect (losing or gaining more stability) of changing an epigenetic makeup from Γ to Γ' can be given by $\Delta\Delta E_{\Gamma\rightarrow\Gamma'} = \Delta E_{\Gamma'} - \Delta E_\Gamma$. Assuming that Γ' is the native epigenetic makeup found in F-DNA crystal structure then according to Supplementary Figure S3, it has a stabilizing effect since $E_{\Gamma'}$ (F-DNA) $< E_{\Gamma'}$ (B-DNA), thus $\Delta E_{\Gamma'} < 0$. If Γ refers to the case when all modifications are removed then as Figure 4 indicates it has a destabilizing effect on F-DNA as E_Γ (B-DNA) $< E_\Gamma$ (F-DNA). Therefore, changing Γ to Γ' is a favorable (stabilizing F-DNA) move (in epigenetic makeup space) as indicated by the negative value of $\Delta\Delta E_{\Gamma\rightarrow\Gamma'}$.

3 Results

The molecular simulation software MOSAICS was previously used to predict nucleosome occupancy by simulating 5mC marks on long stretches of DNA (Minary and Levitt, 2014). Here we extend this particular MOSAICS protocol to perform simulations of 5hmC and 5fC epigenetic marks. We validate our protocol using four high-resolution crystal structures containing 5hmC, 5mC, 5fC as well as the unmodified version (C) of a DNA Dodecamer. First, we studied the effect of isolated epigenetic marks (5hmC and 5mC) on base pairing. Then we proceed to perform simulations of the CpG island with six 5fC marks. Our simulations agreed with experimental observations in that we correctly identified the energetically favorable structures for different epigenetic marks. On this basis we propose epigenetic makeups that present possible transition points between B-DNA and F-DNA forms of a DNA Dodecamer.

3.1 Simulating configurations of 5hmC and 5fC marks

The authors of the crystal structure containing the 5hmC mark (4C5X, Supplementary Fig. S2) observed two main configurations of the hydroxymethyl modification. Figure 1A shows the major configuration, called S_A, with 70% occupancy, which has the hydroxymethyl –OH group pointed towards the O6 oxygen of the neighboring guanine. The figure also depicts the other configuration, called S_B, which has 30% occupancy and has the –OH group pointing towards the phosphate backbone. We tested whether the distribution of the configurations was reflected in our simulations.

We performed NMMC simulations (see Section 2) of the 5hmC mark using the recently published high-resolution structure as a starting point for the protocol. We started simulations with either the S_A or S_B configuration. We noted the dihedral angle of the C4-C5-OH-HO atoms in each trajectory frame as a description of the 5hmc–OH angle. To estimate the distribution of the 5hmC–OH angle, we combined the distributions from both starting points. If there was no bias towards configuration S_A or S_B in our simulations, or there was no switching between the two, we would see a 50:50 distribution. Instead we obtained a distribution of 60% configuration S_A and 40% configuration S_B, similar to the proportion reported in the original article (Fig. 1B). The percentage discrepancy between our simulation and the experimentally observed values can stem from the presence of essential thermal fluctuations, which are captured by our simulations (conducted at a physiological temperature) but not captured by the crystallographic study. Our simulations did not produce any new configurations of the epigenetic

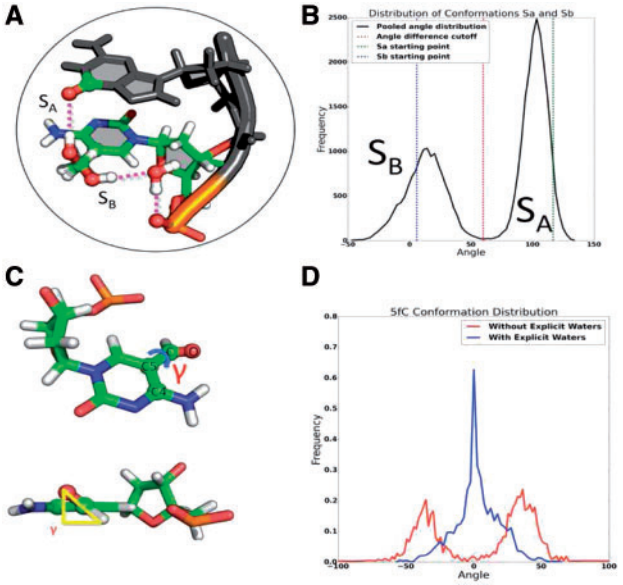

Fig. 1. Simulating configurations of 5hmC and 5fC. (**A**) Authors of 4C5X noted that there exist two major configurations of 5hmC, one where –OH points towards the O6 of guanine (configuration S_A) and one where hydroxymethyl –OH points towards the phosphate backbone, forming water-mediated hydrogen bonds (configuration S_B). In the crystal structure the two configurations had 70 and 30% occupancy respectively. (**B**) Distribution of S_A and S_B of 5hmC configurations in terms of distributions of the dihedral angle HO-OH-C5M-C5. The cutoff between configurations S_A and S_B was taken as the boundary between the two distributions, at 60 degrees. In our simulations, we observe a greater proportion of configuration S_A (60% in favor of S_A) as noted by the authors of 4C5X. **(C)** We defined the γ angle to describe the configuration of 5fC. Gamma was defined as the dihedral angle along C4-C5-C-O. **(D)** Distribution of the configurations of 5fC. We collected the γ angle statistics in a similar fashion as for the configurations of 5hmC and simulations were performed with and without structural waters (present in the crystal). Simulations with structural waters exhibited the same distribution that was observed in the crystal structure. Simulations without structural waters had a tendency to be centered around -40 and 40 degrees

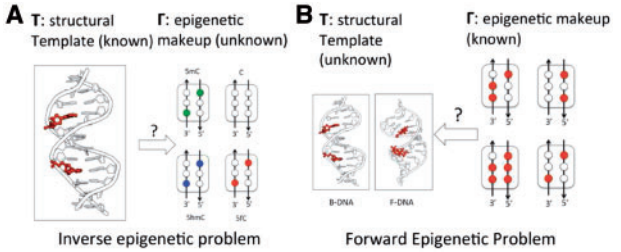

Fig. 2. In structural epigenetics we identify the inverse and forward epigenetic problems. (**A**) Inverse epigenetic problem. Given a structural template T, we would like to identify an epigenetic makeup Γ, which is most energetically suitable for the template. (**B**) Forward epigenetic problem. Given an epigenetic makeup Γ, predict the most suitable 3D structure among alternative templates (or the structural change Γ causes with respect to a reference structure without epigenetic modifications)

marks, but generated populations around the experimentally determined configurations S_A and S_B (Fig. 1B). We have performed the simulations both with explicit waters as well as with implicit solvent (Fig. 1B shows the distributions with the explicit waters present). Performing simulations with explicit waters yielded the same distribution as with the implicit water, however after a longer simulation time.

This demonstrates that we are able to reliably simulate the configurations of 5hmC. We used a similar approach on the 5fC system (4QKK). The authors of this structure noted that the carbonyl O of the 5fC should be in the same plane as the nucleotide rings (γ angle being 0° as in Fig. 1C). We started NMMC simulations of 4QKK orienting the 5fC at random γ angles between −30 and 30 degrees (as suggested by the authors of the original structure). We observed that the distribution was centered around $\gamma = 0°$, only if we used explicit waters in our simulations rather than implicit solvent alone (Fig. 1D). When there were no explicit waters in our simulations, the angle distribution split in a bimodal fashion, with the modes being centered around −40 and 40 degrees, respectively; occasionally the formyl flipped by 180 degrees altogether. Thus, it appears that in our 5fC simulations, it is necessary to use explicit waters in order to keep the stability of the experimentally seen configurations of 5fC. Hydration in general has a profound effect on our ability to simulate 5fC as the F-DNA structure maintains a hydration network in our simulations (see Supplementary Section S1). This is in line with previous computational (Carvalho *et al.*, 2014) and experimental (Raiber *et al.*, 2015) studies that implicated the hydration is affected by the epigenetic makeup and it is important in maintaining F-DNA state, respectively.

3.2 The structure of F-DNA energetically favors multiple 5fC epigenetic marks

Having demonstrated that we can use our protocol to reliably reproduce the experimental configurations of epigenetic marks in our two test cases, we tested the effects of varying the epigenetic makeup on the overall structure of the B-DNA and F-DNA.

First we modeled the effects of isolated epigenetic marks as observed in the high-resolution structures [5hmC (4C5X), 5mC (4C63)]. In line with the experimental observations (Lercher *et al.*, 2014; Renciuk *et al.*, 2013), such isolated epigenetic marks do not appear to have a profound effect on the structure (helical parameters or curvature) of B-DNA. Nevertheless we can confirm that there might be context-dependent effects, which could be enhanced if multiple epigenetic marks are present (see Supplementary Section S2). Given that epigenetic marks often cluster along the DNA strands, it is necessary to consider different epigenetic makeup variants together with the structural and thermodynamic effects these might have. For this purpose, we defined the inverse and forward epigenetic problems (Fig. 2).

In the forward epigenetic problem, given an epigenetic makeup, we aim to find the optimal structure the system can assume out of a set of templates. In the inverse epigenetic problem, given a structural template, we aim to find which epigenetic makeup would be the most optimal for it. Defining those two problems formalizes the way we think about the different epigenetic marks and their configurations, offering an organized framework in which to address them.

As an application of our computational framework, we tested the tendency of the DNA to remain in the F-DNA state depending on the epigenetic mark type (inverse epigenetic problem, Fig. 2A) (Raiber *et al.*, 2015). For this purpose, we defined four statistics, which quantify the curvature difference between F-DNA and B-DNA structures. These were defined as the distances between C5 atoms and distances between phosphates in (epigenetically modified and unmodified) CYT residues on opposite strands (see Fig. 3). When the DNA is fully extended, as in B-DNA, these residues are at maximum separation. These distances are much smaller in the F-DNA state.

Since the formylated makeup is more likely to fold in the F-DNA state [as demonstrated by F-DNA to B-DNA transition after 5fC to

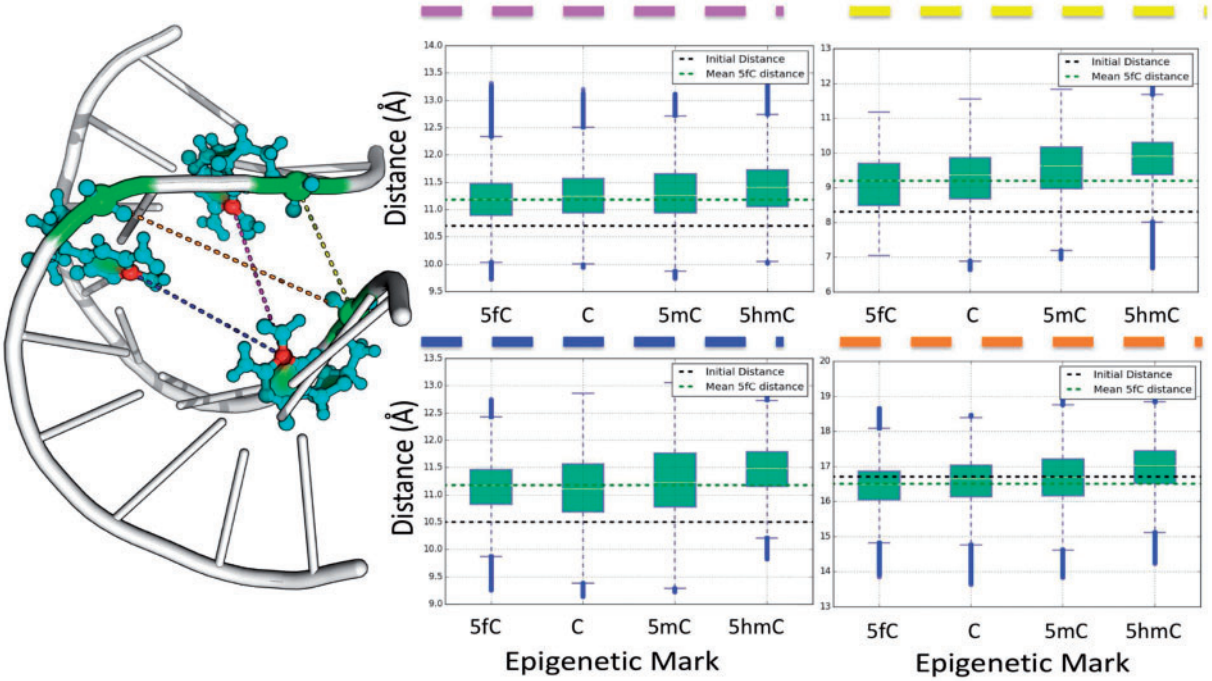

Fig. 3. We have assessed our ability to solve the inverse epigenetics problem by quantifying the preference to remain in F-DNA form depending on the epigenetic modification present (5fC, 5hmC, 5mC or C). A DNA Dodecamer is unlikely to exist in F-DNA forms without multiple 5fC modifications (e.g. 1VE8 and findings of Raiber *et al.*). Therefore, a simulation framework addressing the inverse epigenetic problem should be able to predict this. As an indicator of being in the F-DNA state, we take the statistics of distances between C5 atoms as well as phosphate atoms in the (epigenetically modified and unmodified) CYT residues on opposite strands. These distances are shorter in the F-DNA form than in the B-DNA form. We plot the statistics for these distances after simulating the F-DNA dodecamer [d(CTACGCGCGTAG) as in 4QKK] with six 5fC, six 5mC, six 5hmC marks or non-methylated altogether. The box-plots indicate the statistics of those distances from our simulations, sorted by upper quartile. The distances are the closest for 5fC (green line), indicating that in this case our simulations have a higher tendency to maintain the folded shape of F-DNA than the other epigenetic marks

5hmC conversion (Raiber *et al.*, 2015)], a protocol solving the inverse epigenetic problem should be able to identify 5fC as the preferred epigenetic mark for F-DNA. To generate our starting structures, we mutated the F-DNA dodecamer to have C, 5mC or 5hmC in the place of 5fC. We then performed NMMC simulations on each of the four epigenetic makeups. In each simulation, we recorded the statistics for the four distances we selected as measures for approximating the bending of the F-DNA structure. In our simulations, 5fC on average maintains the closest distances, i.e. the most bending. Only in one case unmodified variety achieves a smaller average distance. In this case however, the distance spread is higher for the unmodified residue than the 5fC variety and thus we argue that based on the latter measure the structure appears to be more conserved in the presence of formyl modifications. This correctly suggests that the 5fC is the most optimal epigenetic mark for maintaining the F-DNA state.

Next, we attempted to address the forward epigenetic problem (Fig. 2B) in cases when epigenetic marks may appear at alternative locations. First we assessed whether our simulations correctly reproduce the energetic favorability for a structure, given a particular formyl epigenetic makeup. For example, experimental evidence suggests that having only two epigenetic marks at opposite ends as in 1VE8 should be more favorable energetically in the B-DNA state (PDB ID 1VE8 and findings of Raiber *et al.*, 2015)

For a given number of formyl modifications, we set up simulations of both B-DNA and F-DNA forms (dodecamer sequence as in 4QKK), having the same number of atoms so as to be able to compare the energy of the systems in the two states. The energy difference ($E_{F\text{-}DNA} - E_{B\text{-}DNA}$) for the same epigenetic makeup between F-DNA and B-DNA templates indicates which one is more favorable (for F-DNA or B-DNA) according to our current protocol. The unformylated and 2-formylated makeups were more energetically favorable in the B-DNA state. However, the 6-formylated epigenetic makeup was more energetically favorable in the F-DNA state. Therefore in each of the three cases, our protocol correctly identifies the experimental observation and suggests the native solution to the forward epigenetic problem (Fig. 4A) is in fact the most stable. This provides an indication that our protocol is suitable to study solutions to the forward epigenetic problem in general. This in turn might provide hints as to where the transition between the B-DNA and F-DNA forms happens, providing new insights into the F-DNA to B-DNA transition as studied by Raiber *et al.* (2015).

3.3 Epigenetic designability of DNA

Given that we appear to be able to predict energetic favorability of a given epigenetic makeup towards either the F-DNA or B-DNA state, we performed a study into the epigenetic configurations at which the B-DNA/F-DNA transition might occur.

We enumerated all the 24 possible epigenetic makeups where each of the six available positions can be either formylated or not formylated (2^6 possible combinations, ignoring mirrored cases due to symmetry of structure). We performed energy minimization on each of the makeups for the B-DNA and F-DNA forms, noting the energy difference between the two states ($E_{F\text{-}DNA} - E_{B\text{-}DNA}$). The results of this procedure are given in Figure 4B.

We observed very high variability in the stabilizing effects (on F-DNA with respect to B-DNA) of epigenetic makeups even if they

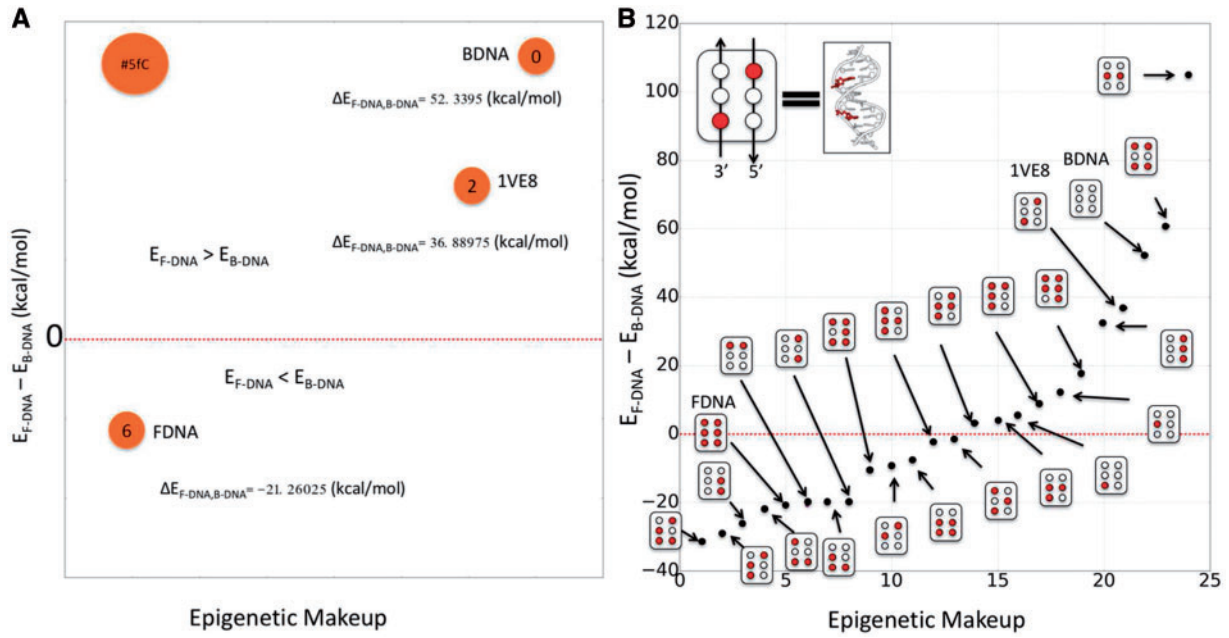

Fig. 4. (**A**) Difference between the energy of the F-DNA and B-DNA for a different number of 5fC modifications present. The configuration with two epigenetic marks was chosen as in the structure of 1VE8. (**B**) The same as in (A) but for all individual epigenetic makeups possible (on the CpG island), which are schematically annotated for each point. The filled circles correspond to positions of 5fC marks, the empty circles indicate the unmodified C

had the same number of formylated sites. On the other hand, such variability is not surprising as certain partially formylated makeups are unlikely to lead to perfect B-DNA or F-DNA states as shown here. Rather, these might correspond to transition states between the two and could provide insights into how designable a CpG island is. Therefore, we propose that results in Figure 4B be used as a reference to perform detailed MD simulations or for prioritizing epigenetic makeups to be studied in an experimental setup.

We do not expect that the main concepts drawn from our results presented by Figure 4 are sensitive to using one of the currently available AMBER force-fields (Krepl *et al.*, 2012; Pérez *et al.*, 2007; Zgarbová *et al.*, 2013, 2015) because the focus of this study is to provide insight into the quantities such as $\Delta\Delta E_{\Gamma\to\Gamma'}$ (see above) for every possible pair (Γ, Γ') rather than predict the absolute values of quantities such as ΔE_{Γ} and $\Delta E_{\Gamma'}$. Although the absolute values of ΔE_{Γ} $\Delta E_{\Gamma'}$, may partially depend on differences in the template structures (e.g. different sugar puckering, ε, ζ and χ torsions), the quantities $\Delta\Delta E_{\Gamma\to\Gamma'} = \Delta E_{\Gamma'} - \Delta E_{\Gamma}$ are mostly determined by the non-bonded interactions, which are the same for all recent generation of AMBER DNA force-fields (Ivani *et al.*, 2016; Krepl *et al.*, 2012; Pérez *et al.*, 2007; Zgarbová *et al.*, 2013, 2015).

4 Discussion

Recent studies of epigenetic marks, in particular 5hmC and 5fC, have expanded our understanding of the effects they have on the stability and structure of DNA (Carvalho *et al.*, 2014; Lercher *et al.*, 2014; Raiber *et al.*, 2015; Renciuk *et al.*, 2013; Szulik *et al.*, 2015; Wanunu *et al.*, 2011). Evidence is accumulating that these modifications are not only intermediate demethylation steps but also have functional roles of their own (Brazauskas and Kriaucionis, 2014; Breiling and Lyko, 2015). Studying all the epigenetic makeups experimentally in relevant sequence contexts, such as CpG islands, would be challenging due to the large number of variations that would have to be tested.

We have extended the molecular simulation software MOSAICS to be able to model 5hmC and 5fC modifications. The purpose of this was to provide an accessible platform to study diverse epigenetic makeups and to provide a unified framework to benchmark different sampling methodologies and force fields especially relating to epigenetics. In order to demonstrate that the results, which can be obtained from MOSAICS, can be used to draw conclusions relating to epigenetic systems, we use a previous training-free protocol which had been shown to provide reliable results for 5mC marks (Minary and Levitt, 2014). We show that the extension of this protocol to 5fC and 5hmC marks, reproduces observed distributions of the experimentally derived 5hmC and 5fC configurations (Lercher *et al.*, 2014; Raiber *et al.*, 2015), which increases the confidence of the conclusions drawn from our simulations. We studied the effects of isolated epigenetic marks. It appears that there are only minor structural effects a single epigenetic mark might have (Supplementary Section S2), which is in accordance with experimental studies (Lercher *et al.*, 2014; Raiber *et al.*, 2015; Renciuk *et al.*, 2013). Nevertheless, we note that there can exist weak context-dependent interactions, which could exert an effect if more epigenetic marks were present, such as in the structure of F-DNA. Such context-dependent interactions might be brought about by water-mediated hydrogen bonds. Authors of the F-DNA structure postulated that hydration might be an important factor in maintaining the F-DNA structure. Our results appear to corroborate this suggestion. Water appears to play a major role in our ability to simulate 5fC in particular. The F-DNA structure, because of its helical underwinding, traps water molecules at its center and facilitates access in the exposed grooves, which contributes to its maintaining of a water network through much of our simulation (Supplementary Section S1).

Nevertheless, the role of the F-DNA structure remains elusive. Studying the F-DNA structure in the context of different epigenetic makeups might shed light on its potential role in chromatin rearrangement and the ability to recruit transcription factors. For instance, it would be worthwhile to check if this structure facilitates

the enzymatic activity of TET upon CpG islands. A similarly bent structure can be observed in 5hmC DNA-ngTET complex and it appears to facilitate the access of the enzyme to the 5hmC mark (Supplementary Fig. S7) (Hashimoto *et al.*, 2015). Studying such effects requires enumerating multiple variants of the epigenetic makeup and assessing their energetic favorability. Our protocol correctly determines that multiple 5fC modifications are more energetically favorable in the F-DNA rather than B-DNA structure.

Our general framework correctly identifies solutions to the inverse and forward epigenetic problems, as tested on the high-resolution crystal structures of the oxidized epigenetic marks. On the basis of this, we have enumerated all the possible epigenetic makeups in a representative structure (4QKK) and assessed their energetic favorability. The results might provide insights into the transition between the B-DNA and F-DNA structures and act as a reference for prioritizing partial epigenetic makeups to be studied experimentally.

In conclusion, we defined and addressed the forward and inverse epigenetic problems by developing a robust computational simulation framework that is ready-to-use (see Section 2). Furthermore, epigenetic marks as well as new forms of DNA such as F-DNA might benefit from development of more reliable parametrizations for force-fields (Dans *et al.*, 2017; Ivani *et al.*, 2016; Sponer *et al.*, 2013). There have not been many studies benchmarking dynamics of DNA with respect to different force fields and methodologies (Dans *et al.*, 2016, 2017; Galindo-Murillo *et al.*, 2016; Ivani *et al.*, 2016). Force fields receive parametrization updates which run the risk of reflecting the current (sometimes biased) state of the PDB. Just like with any other experiments, to conduct reliable benchmarking of force-fields and sampling methodologies, uniform testing conditions should be assured. In the field of molecular simulations, this would mean a platform, which is capable of performing simulations using an arbitrary molecule, arbitrary force field and an arbitrary sampling methodology. This will be particularly important as new parametrizations would surely be needed for the new epigenetic marks and diverse DNA forms, such as F-DNA. MOSAICS allows for performing simulations with different force fields and sampling methodologies and as such can be employed as a benchmarking platform. We hope that our platform would form a basis for consolidating the efforts in simulating and studying the structural effects of diverse epigenetic makeups. In particular, as a future directions one may also use our platform to model conformational transitions triggered by epigenetic modifications.

Acknowledgements

We acknowledge use of Hartree Centre resources as part of the first Xeon Phi Access Programme for this work. The STFC Hartree Centre is in association with IBM providing High Performance Computing platforms funded by the UK's investment in e-Infrastructure. The Centre aims to develop and demonstrate next generation software, optimized to take advantage of the move towards exa-scale computing.

Funding

This work was supported by the 2020 programme (Engineering and Physical Sciences Research Council Cross Discipline Interface Programme, EP/I017909/1) to C.M.D.

Conflict of Interest: none declared.

References

Acosta-Silva,C. *et al.* (2010) Mutual relationship between stacking and hydrogen bonding in DNA. Theoretical study of guanine-cytosine, guanine-5-methylcytosine, and their dimers. *J. Phys. Chem. B*, **114**, 10217–10227.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhutani, N. *et al.* (2011) DNA demethylation dynamics. *Cell*, **146**, 866–872.

Brazauskas,P. and Kriaucionis,S. (2014) DNA modifications: Another stable base in DNA. *Nat. Chem.*, **6**, 1031–1033.

Breiling,A. and Lyko,F. (2015) Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin*, **8**, 24.

Carson,S. *et al.* (2016) Hydroxymethyluracil modifications enhance the flexibility and hydrophilicity of double-stranded DNA. *Nucleic Acids Res.*, **44**, 2085–2092.

Carvalho,A.T.P. *et al.* (2015) Theoretical modelling of epigenetically modified DNA sequences. *F1000Res.*, **24**, 52.

Carvalho,A.T.P. *et al.* (2014) Understanding the structural and dynamic consequences of DNA epigenetic modifications: computational insights into cytosine methylation and hydroxymethylation. *Epigenetics*, **9**, 1604–1612.

Dans, P.D. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **20**, 4217–4230.

Dans, P.D. *et al.* (2016) Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res*, **44**, 4052–4066.

Demharter, S. *et al.* (2016) Modeling functional motions of biological systems by customized natural moves. *Biophys. J*, **111**, 710–721.

Derreumaux, S. *et al.* (2001) Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Res*, **29**, 2314–2326.

Drew, H.R. *et al.* (1981) Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A*, **78**, 2179–2183.

Frauer,C. *et al.* (2011) Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. *PLoS One*, **6**, e21306.

Galindo-Murillo,R. *et al.* (2016) Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.*, **12**, 4114–4127.

Guo,J.U. *et al.* (2011) Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. *Cell Cycle*, **10**, 2662–2668.

Hansmann,U.H.E. (1997) Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, **281**, 140–150.

Hashimoto,H. *et al.* (2015) Structure of Naegleria Tet-like dioxygenase (NgTet1) in complexes with a reaction intermediate 5-hydroxymethylcytosine DNA. *Nucleic Acids Res.*, **43**, 10713–10721.

Hayward,S. and de Groot,B.L. (2008) Normal modes and essential dynamics. *Methods Mol. Biol.*, **443**, 89–106.

Humphrey,W. *et al.* (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.

Iurlaro,M. *et al.* (2013) A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.*, **14**, R119.

Ivani,I. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.

Knapp,B. *et al.* (2016) Exploring peptide/MHC detachment processes using Hierarchical Natural Move Monte Carlo. *Bioinformatics*, **32**, 181–186.

Krawczyk,K. *et al.* (2016) Tertiary element interaction in HIV-1 TAR. *J. Chem. Inf. Model*, **56**, 1746–1754.

Krepl,M. *et al.* (2012) Reference simulations of noncanonical nucleic acids with different χ variants of the AMBER FORCE Field: quadruplex DNA, quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.*, **8**, 2506–2520.

Kroeze, L.I. *et al.* (2015) 5-Hydroxymethylcytosine: An epigenetic mark frequently deregulated in cancer. *Biochim. Biophys. Acta*, **1855**, 144–154.

Lercher,L. *et al.* (2014) Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem. Commun. (Camb.)*, **50**, 1794–1796.

MacKerel,A.D., Jr. *et al.* (1998) CHARMM: the biomolecular simulation program. In: Brooks BR *et al.* (eds) *J Comput Chem.*, **30**, 1545–1614.

Minary,P. and Levitt,M. (2010) Conformational optimization with natural degrees of freedom: a novel stochastic chain closure algorithm. *J. Comput. Biol.*, **17**, 993–1010.

Minary,P. and Levitt,M. (2014) Training-free atomistic prediction of nucleosome occupancy. *Proc. Natl. Acad. Sci. USA*, **111**, 6293–6298.

Moraga,I. *et al.* (2015) Tuning cytokine receptor signaling by re-orienting dimer geometry with surrogate ligands. *Cell*, **160**, 1196–1208.

Muenzel,M. *et al*. (2010) Chemical discrimination between dC and 5MedC via their hydroxylamine adducts. *Nucleic Acids Res*., **38**, e192.

Pérez,A. *et al*. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J*., **92**, 3817–3829.

Raiber,E.-A. *et al*. (2015) 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol*, **22**, 44–49.

Renciuk,D. *et al*. (2013) Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Res*., **41**, 9891–9900.

Severin, P.M.D. *et al*. (2011) Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Res*, **39**, 8740–8751.

Shen,L. *et al*. (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, **153**, 692–706.

Siegfried,Z. *et al*. (1999) DNA methylation represses transcription in vivo. *Nat. Genet*., **22**, 203–206.

Sim, A.Y.L. *et al*. (2012) Modeling and design by hierarchical natural moves. *Proc. Natl. Acad. Sci. USA*, **109**, 2890–2895.

Song,C.X. *et al*. (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, **153**, 678–691.

Sponer,J. *et al*. (2013) Relative stability of different DNA guanine quadruplex stem topologies derived using large-scale quantum-chemical computations. *J. Am. Chem. Soc*., **135**, 9785–9796.

Spruijt,C.G. *et al*. (2013) Dynamic readers for 5-(Hydroxy)methylcytosine and its oxidized derivatives. *Cell*, **152**, 1146–1159.

Sun,W. *et al*. (2014) From development to diseases: The role of 5hmC in brain. *Genomics*, **104**, 347–351.

Szulik,M.W. *et al*. (2015) Differential stabilities and sequence-dependent base pair opening dynamics of Watson-crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. *Biochemistry*, **54**, 1294–1305.

Thalhammer,A. *et al*. (2011) Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chem. Commun. (Camb.)*, **47**, 5325–5327.

Tsui,V. and Case,D.A. (2001) Theory and Applications of the Generalized Born Solvation Model in macromolecular simulations. *Biopolymers*, **56**, 275–291.

Wang,L. *et al*. (2014) Programming and inheritance of parental DNA methylomes in mammals. *Cell*, **157**, 979–991.

Wanunu,M. *et al*. (2011) Discrimination of methylcytosine from hydroxymethylcytosine in DNA molecules. *J. Am. Chem. Soc*., **133**, 486–492.

Wu,H. and Zhang,Y. (2015) Charting oxidized methylcytosines at base resolution. *Nat. Struct. Mol. Biol*., **22**, 656–661.

Yusufaly,T.I. *et al*. (2013) 5-Methylation of cytosine in CG: CG base-pair steps: A physicochemical mechanism for the epigenetic control of DNA nanomechanics. *J. Phys. Chem. B*., **117**, 16436–16442.

Zgarbová,M. *et al*. (2015) Refinement of the sugar-phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput*., **11**, 5723–5736.

Zgarbová,M. *et al*. (2013) Toward improved description of DNA backbone: Revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput*., **9**, 2339–2354.

Zhang,J. *et al*. (2012) Multiscale natural moves refine macromolecules using single-particle electron microscopy projection images. *Proc. Natl. Acad. Sci*., **109**, 9845–9850.