Rao *et al. Epigenetics & Chromatin (2018) 11:6*
https://doi.org/10.1186/s13072-018-0174-4

Epigenetics & Chromatin

RESEARCH

Open Access

Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding

Satyanarayan Rao[1], Tsu-Pei Chiu[1], Judith F. Kribelbauer[2,3,4], Richard S. Mann[3,4,5,6], Harmen J. Bussemaker[2,3*] and Remo Rohs[1,7,8,9*]

Abstract

Background: DNA shape analysis has demonstrated the potential to reveal structure-based mechanisms of protein–DNA binding. However, information about the influence of chemical modification of DNA is limited. Cytosine methylation, the most frequent modification, represents the addition of a methyl group at the major groove edge of the cytosine base. In mammalian genomes, cytosine methylation most frequently occurs at CpG dinucleotides. In addition to changing the chemical signature of C/G base pairs, cytosine methylation can affect DNA structure. Since the original discovery of DNA methylation, major efforts have been made to understand its effect from a sequence perspective. Compared to unmethylated DNA, however, little structural information is available for methylated DNA, due to the limited number of experimentally determined structures. To achieve a better mechanistic understanding of the effect of CpG methylation on local DNA structure, we developed a high-throughput method, *methyl*-DNAshape, for predicting the effect of cytosine methylation on DNA shape.

Results: Using our new method, we found that CpG methylation significantly altered local DNA shape. Four DNA shape features—helix twist, minor groove width, propeller twist, and roll—were considered in this analysis. Distinct distributions of effect size were observed for different features. Roll and propeller twist were the DNA shape features most strongly affected by CpG methylation with an effect size depending on the local sequence context. Methylation-induced changes in DNA shape were predictive of the measured rate of cleavage by DNase I and suggest a possible mechanism for some of the methylation sensitivities that were recently observed for human Pbx-Hox complexes.

Conclusions: CpG methylation is an important epigenetic mark in the mammalian genome. Understanding its role in protein–DNA recognition can further our knowledge of gene regulation. Our high-throughput *methyl*-DNAshape method can be used to predict the effect of cytosine methylation on DNA shape and its subsequent influence on protein–DNA interactions. This approach overcomes the limited availability of experimental DNA structures that contain 5-methylcytosine.

Keywords: *methyl*-DNAshape, 5-methylcytosine, DNA methylation, Epigenetics, DNA structure, DNase I cleavage sensitivity, Human Hox protein binding specificity

*Correspondence: hjb2004@columbia.edu; rohs@usc.edu
[1] Computational Biology and Bioinformatics Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
[2] Department of Biological Sciences, Columbia University, New York, NY 10027, USA
Full list of author information is available at the end of the article

Background

Cytosine methylation is the most abundant of all epigenetic marks found on DNA. At the molecular level, cytosine methylation involves the addition of a methyl (CH_3) group to the C5 atom of cytosine, yielding 5-methylcytosine (5mC). In mammalian genomes, this alteration often occurs in the context of the CpG dinucleotide and is referred to as "CpG methylation" or "DNA methylation." Ever since 5mC was proposed as a potential epigenetic factor capable of altering gene regulation and cellular differentiation [1], research in this field has been quite active. A recent review [2] highlights the complexity in the interpretation of epigenetic data and the evolution of the definition of epigenetics as the field has advanced.

Although the addition of a single methyl group at the major groove edge leads to only a subtle change in DNA structure, important functional effects have been observed at different scales. For example, methylation-induced alterations in gene expression have been observed in regulatory regions [3–5], and an increase in DNA methylation in one of the X-chromosomes in the female genome can lead to X-chromosome inactivation [6, 7]. Effects of methylation have been studied in two main contexts, genome organization and protein–DNA interactions. Owing to recent advances in technology, DNA methylation profiling can now be performed for any given genome [8–10]. Furthermore, in vitro approaches have recently been used to profile systematically the influence of methylation on DNA binding for human transcription factors (TFs) [11–14], by using variants of universal protein-binding microarray (PBM), high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX), and SELEX in combination with massively parallel sequencing (SELEX-seq). These approaches revealed that methylation affects binding across the affinity range and that the effect varies within and between TF families [13, 15–17].

To achieve mechanistic insights into these phenomena, detailed understanding of the biophysical and structural effects of DNA methylation is required. Some proteins, such as the Lac repressor, prefer having a bulky methyl group in the major groove and form hydrophobic contacts to this group [18]. By contrast, MspI, a *Moraxella* sp. restriction endonuclease, recognizes the CCGG sequence irrespective of methylation status [18]. These context-dependent effects may be explained in terms of three possible readout mechanisms: direct contacts [19], competitive binding [20, 21], and structural readout [22]. Direct contact to a methyl group allows for the possible formation or alteration of van der Waals interactions, which can either completely abolish or enhance binding [19, 23]. For example, CpG methylation of the cyclic adenosine monophosphate (cAMP) response element half-site (half-CRE) confers binding of CCAAT/enhancer-binding protein alpha (C/EBPα) and C/EBPβ and abolishes binding of CREB, c-Jun, JunD, and ATF2 [24]. In a competitive binding mechanism, the methyl-CpG binding protein (MeCP2) initially binds methylated CpG sites and then blocks sites for other proteins to bind [20, 21]. Many TFs seem to employ one of these first two mechanisms, as revealed by in vitro binding assays [23]. In the case of structure-mediated methylation sensitivity, first demonstrated for the endonuclease DNase I [25], local DNA shape changes enhance binding to target sites already preferred by particular DNA-binding proteins. While direct contacts with the methyl group confer binary effects, the shape-dependent effect is sequence context dependent and can fine-tune the binary direct contact mechanism.

Here, we introduce a methodology that enables quantitative probing of the shape-dependent methylation effect. We recently studied how DNA shape contributes to protein–DNA recognition [26–28]. However, we have not yet systematically quantified the effect of DNA methylation on protein binding [22]. Motivated by the widespread occurrence of CpG dinucleotides in TF binding motifs of different protein families [29–31], we aimed to study CpG methylation in the context of gene regulation (Fig. 1b). Understanding the protein–DNA readout of methylated cytosine requires structural insight derived from experimentally determined structures. Unfortunately, the current content of the Protein Data Bank (PDB) [32] includes only a few structures containing cytosine modifications (Fig. 1a). To close this knowledge gap, we utilized computational modeling of many DNA fragments to study the intrinsic effects induced by cytosine methylation, in a manner analogous to previous high-throughput studies of DNA shape of unmethylated genomic regions [33–35]. The resulting query tables can be utilized to analyze systematically the effect of methylation on protein–DNA interactions, as we demonstrate for DNase I cleavage and Pbx-Hox binding data.

Methods

Sequence and structure datasets

A total of 3518 DNA fragments of lengths varying from 13 to 24 base pairs (bp) were considered in all-atom Monte Carlo (MC) simulations, based on a previously published protocol (see Additional file 1 for details) [36]. Before performing simulations, we added 5-methyl groups at CpG steps to the core sequence (central regions in sequences in Additional file 2: Table S1) of every DNA fragment [25]. Sequences of these fragments were designed to capture the complete pentamer space in terms of the sequence context. Each considered sequence was defined as having at least one CpG step. For better

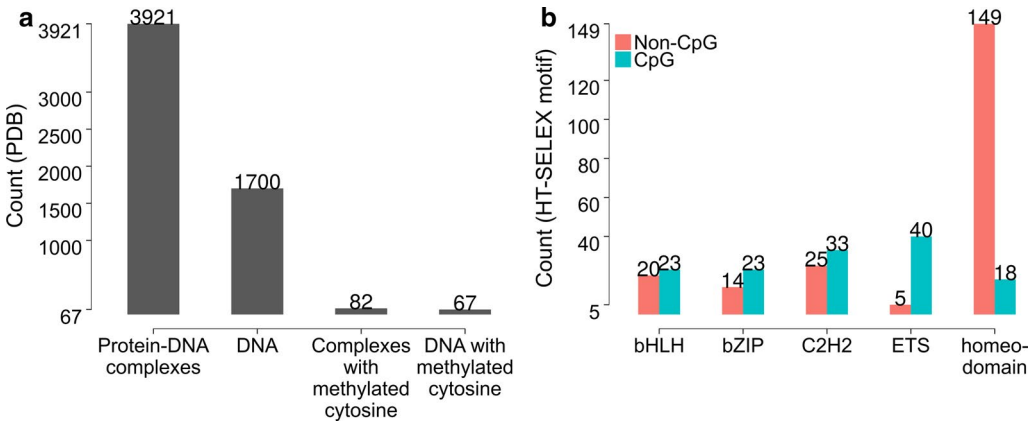


Fig. 1 Current statistics of available structures and abundance of CpG dinucleotides in TF binding sites. **a** Count statistics of protein–DNA complex and unbound DNA structures available in the PDB as of 31 May 2017. Counts of subsets of structures (right two bars) containing methylated DNA at CpG site(s) or in other sequence contexts were two orders of magnitude lower than the count of structures containing unmethylated DNA. Systematic profiling of the effect of methylation on three-dimensional DNA structure would require a substantially larger number of structures. Counts include structures solved by X-ray crystallography and NMR spectroscopy. **b** Abundance of CpG steps in TF binding motifs in HT-SELEX data for human TF datasets [29], derived using MotifDb [51]. CpG dinucleotides can be observed in binding sites irrespective of TF family. Five largest human TF families (based on number of binding sites containing at least one CpG step) are specified. Almost 90% of ETS family motifs contain CpG steps. Numbers on each bar represent counts of motifs containing CpG or no CpG steps

coverage of the sequence space, four different nucleotide combinations were used to flank each designed sequence. Canonical B-DNA structures for all DNA fragments were generated by the JUMNA program [37] and used as input for the all-atom MC simulations [36].

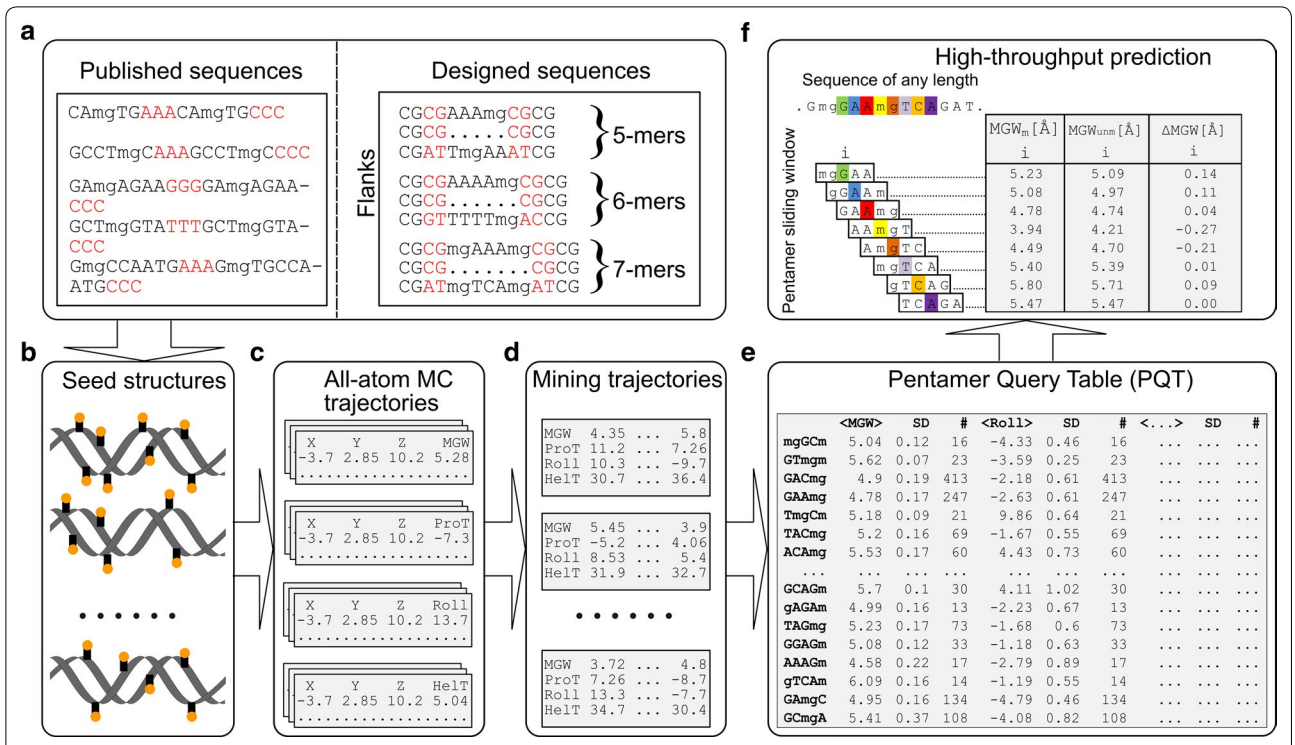
All-atom MC simulations

MC simulations (Fig. 2c) traverse the energy landscape by making random moves [38], thus combining effective sampling with fast equilibration [39]. For this study, MC sampling was expanded to include 5mC. Rotation of the 5-methyl group added one degree of freedom, whose rotation was implemented in a manner analogous to that of the thymine 5-methyl group. Partial charges for 5mC were taken from a database of AMBER force fields for naturally occurring modified nucleotides [25, 40]. For a given DNA structure, the MC simulation protocol included two million MC cycles, with each cycle attempting random variations of all degrees of freedom (Additional file 3: Table S2). After completion of the MC simulations, trajectories were analyzed by using snapshots that were stored every 100 MC cycles. After we discarded the first half-million MC cycles as an equilibration period, we mined the remaining trajectories using CURVES analysis [41] (Fig. 2d; see Additional file 1 for detailed description of methodology).

Building the *methyl*-DNAshape Pentamer Query Table

Mining of the MC trajectories generates average structural features for a given sequence. We assigned minor groove width (MGW) values to nucleotides in a strand-independent manner [33]. We adopted a pentamer sliding-window approach to record DNA shape feature values from representative structures. For a given sequence of length N, the approach profiled the shape features of $N-4$ pentamers due to end effects. For MGW and propeller twist (ProT), values were assigned to the central bp of the corresponding pentamer. For Roll and helix twist (HelT), two values were recorded for bp steps 2–3 and 3–4 of a pentamer, respectively. Shape feature values from multiple occurrences of a given pentamer in different DNA fragments were averaged and assigned as representative values for that pentamer (Additional file 4: Fig. S1).

All possible pentamers were categorized in unmethylated and methylated groups. Unmethylated pentamers contained letters from the standard DNA alphabet, {A, C, G, T}. Methylated pentamers contained letters from the expanded alphabet, {A, C, G, T, m, g}. We assigned the letter "m" to 5mC and lowercase "g" to guanine base-paired with 5mC. We considered there to be no partial methylation; thus, for a DNA fragment of length N, methylation on the forward strand at index i (5′–3′) also indicates methylation at index $i+1$ (3′–5′) on the reverse strand. The G base-paired to 5mC in a methylated 5mC/G bp cannot be treated in a similar fashion as G base-paired to unmethylated C. In addition, due to the requirement of DNA methylation at both Cs of a CpG step, each 5mC will be followed by a G base-paired to another 5mC on the opposite strand. Thus, "m" and "g" cannot be considered as independent letters.



	<MGW>	SD	#	<Roll>	SD	#	<...>	SD	#
mgGCm	5.04	0.12	16	-4.33	0.46	16	...	...	...
GTmgm	5.62	0.07	23	-3.59	0.25	23	...	...	...
GACmg	4.9	0.19	413	-2.18	0.61	413	...	...	...
GAAmg	4.78	0.17	247	-2.63	0.61	247	...	...	...
TmgCm	5.18	0.09	21	9.86	0.64	21	...	...	...
TACmg	5.2	0.16	69	-1.67	0.55	69	...	...	...
ACAmg	5.53	0.17	60	4.43	0.73	60	...	...	...
...	...	...	...	...	...	...	...	...	...
GCAGm	5.7	0.1	30	4.11	1.02	30	...	...	...
gAGAm	4.99	0.16	13	-2.23	0.67	13	...	...	...
TAGmg	5.23	0.17	73	-1.68	0.6	73	...	...	...
GGAGm	5.08	0.12	33	-1.18	0.63	33	...	...	...
AAAGm	4.58	0.22	17	-2.79	0.89	17	...	...	...
gTCAm	6.09	0.16	14	-1.19	0.55	14	...	...	...
GAmgC	4.95	0.16	134	-4.79	0.46	134	...	...	...
GCmgA	5.41	0.37	108	-4.08	0.82	108	...	...	...

Fig. 2 Workflow for high-throughput *methyl*-DNAshape method. **a** Sequence pool. DNA fragments were considered for MC simulations to capture a sequence space that includes CpG methylation. Published sequences (left rectangular box) [23] and manually designed sequences (right rectangular box) included DNA fragments comprising a variable core (containing at least one methylated CpG step, called "mg" step) and flanks (4 bp in length). Right flanks were reverse complements of left flanks. For a given length of core sequence (5, 6, or 7 bp), all possible sequences (Additional file 1) were considered for MC simulations. **b** Seed structures. Canonical B-DNA structures were generated for all selected sequences. The 5-methyl groups (orange circles) were introduced at cytosine positions with letter "m" (on Watson and Crick strand). **c** All-atom MC trajectories. Simulations were performed on seed structures for 2 million MC cycles, with snapshots recorded every 100 cycles after equilibration. **d** Mining trajectories. Recorded snapshots were analyzed for DNA shape features (see Additional file 1: Supplementary methods) associated with corresponding DNA sequences. **e** Pentamer Query Table (PQT). Pentamer sliding-window approach was applied to analyzed DNA fragments. Calculated DNA shape features (HelT, MGW, ProT, and Roll) were recorded at the center of each pentamer. Assigned value for a corresponding shape feature represents the average of all shape feature values in the sequence pool for a given pentamer in the PQT. **f** Front-end interface. Our easy-to-use *methyl*-DNAshape web server or DNAshapeR Bioconductor/R package can be used to profile shape features of any genomic region and DNA sequences of any length by using a pentamer sliding-window approach. The *methyl*-DNAshape web server, available at http://rohslab.usc.edu/methyl-DNAshape/, also outputs the effect of methylation on shape features in terms of Δshape (shown here for MGW)

Introduction of the two letters "m/g" for a 5mC/G bp increased the number of possible unique pentamers, with 475 new pentamers being added to the 512 unique pentamers representing unmethylated DNA (Additional file 5: Table S3). Here, we discuss two specific examples. In the first example, NNmgN where N ∈ {A, C, G, T} has a single methylation mark at the underlined position 3. The second example is the complex case of gmgNm. To assign shape feature values, we have to consider that 5mC precedes "g" on its 5′ flank and that "g" follows "m" on its 3′ flank (Additional file 6: Fig. S2). We ran MC simulations with these combinations of methylated CpG steps to enrich pentamers of these types of compositions (see Additional file 7 for list of all sequences studied with MC simulations).

methyl-DNAshape method for high-throughput prediction of methylated DNA shape features

The *methyl*-DNAshape method derives DNA shape features of methylated DNA at nucleotide resolution, while considering the local sequence context. In a manner analogous to our DNAshape method for unmethylated DNA [33], we used a pentamer centered at position *i* to estimate DNA shape features at that position. We adopted the equivalent approach for DNA with methylated CpG dinucleotides, to capture the methylation properties of mammalian genomes. We derived the *methyl*-DNAshape Pentamer Query Table (*m*PQT), in analogy to the DNAshape Pentamer Query Table (PQT). DNA shape features at nucleotide position *i* were determined by querying the *m*PQT based on a pentamer using two neighboring nucleotides in

both flanks ($P_i = N_{i-2}N_{i-1}N_iN_{i+1}N_{i+2}$). Ultimately, *methyl*-DNAshape calculates four feature vectors, one for each of the shape features HelT, MGW, ProT, and Roll (Fig. 2).

As in our previous work, we selected four DNA shape features that play important roles in protein–DNA recognition [33]. ProT is an intra-bp parameter that accounts for bp twisting along the base-pairing axis. Increased values of ProT lead to an opportunity to form an additional inter-bp hydrogen bond in the major groove [28]. Roll and HelT are bp step features that estimate deformation at the dinucleotide level. The MGW feature plays a pivotal role in DNA shape readout [27]. A narrow minor groove enhances negative electrostatic potential and offers favorable interactions for positively charged amino acids [27]. Although the scarcity of experimentally solved structures with CpG methylation prohibited us from performing a validation such as is possible for unmethylated structures, we compared MGW predictions using *methyl*-DNAshape with X-ray co-crystal structures (Additional file 8: Fig. S3). The *methyl*-DNAshape method is available as a web server at http://rohslab.usc.edu/methyl-DNAshape/ and as an extension to the R/Bioconductor package DNAshapeR [42] at http://bioconductor.org/packages/devel/bioc/html/DNAshapeR.html.

Results and discussion

Effect of CpG methylation on DNA shape features

To quantify the effects of cytosine methylation on DNA shape features, we compared values for all unique pentamers that contained a single CpG step, as derived from DNAshape [33] (designed for unmethylated DNA) and *methyl*-DNAshape (our high-throughput prediction method designed for methylated DNA; see "Methods" section). We considered four DNA shape features—HelT, MGW, ProT, and Roll—in this analysis.

Roll and ProT exhibited strong methylation effects (50–100% of the range observed across all unmethylated-DNA sequences). At methylated CpG steps, Roll increased by an average of 6° (range 5.1°–7.2°), representing a similar effect size as previously observed in molecular dynamics simulations [43]. In methylated C/G bp, ProT decreased by an average of 5° (range − 4.5° to − 6.0°). By contrast, we observed relatively small effects for MGW and HelT (Fig. 3). An increase in Roll caused partial unstacking of the bp step, leading to widening of the minor groove. This conformational change might affect hydrogen bond formation in the major groove by exposing amino groups of guanine bases and oxygens of cytosine bases with different relative orientations. Presence of a methylated CpG

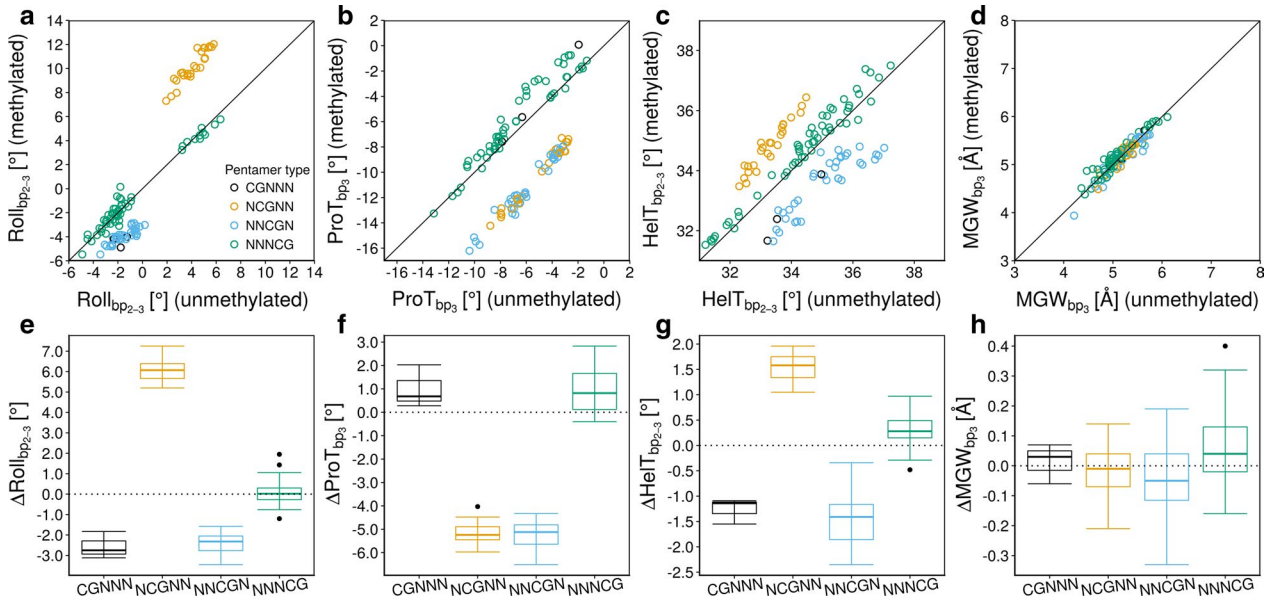


Fig. 3 Effect size of CpG methylation on DNA shape features. Methylation-induced changes were analyzed for four shape features: **a**, **e** roll, **b**, **f** propeller twist (ProT), **c**, **g** helix twist (HelT), **d**, **h** minor groove width (MGW). For each shape feature, values for pentamers from the DNAshape query table for unmethylated DNA were plotted against values for corresponding pentamers from the *methyl*-DNAshape query table for methylated DNA. For simplicity, pentamers with one and only one CpG/mpg step (where "m" represents 5-methylcytosine and "g" represents G base-paired with "m" on the reverse-complement strand) were considered, for a total of 116 occurrences (Additional file 1). For bp step features Roll and HelT, values at bp steps 2–3 of each pentamer were used. For MGW and ProT, values at the central bp of each pentamer were used. CpG methylation increased Roll by an order of magnitude (light-orange dots). The opposite was observed when methylation occurred at the immediate next bp step (light-blue dots). Presence of a methyl group at the central bp, either on the forward (light-blue dots) or reverse (light-orange dots) strand caused a decrease in ProT

step at position 1 or 3 (in the 5′–3′ direction) in pentamers resulted in a lowering of HelT by ~ 2° (Fig. 3c). Only subtle changes in MGW were observed, except for some particular sequence contexts.

Effect of CpG methylation on MGW of A-tracts

A-tracts, or poly[A/T] tracts, consist of a continuous run of at least three As or Ts without any TpA step. A-tracts, which play an important role in TF-DNA binding [44, 45], have a rigid conformation due to inter-bp hydrogen bonds in the major groove [46].

We analyzed the effect of methylation on the MGW of A-tracts flanked by CpG steps. As we derived the shape features from pentamers, we considered A-tracts of limited length of either three (e.g., AAACG; Fig. 4a) or four (e.g., AAAAC; Fig. 4b) nucleotides. For A-tracts that were three bp in length, the subsequent CpG context extended into one nucleotide position flanking the pentamer because 5mC at the fifth position of a pentamer implicitly assumes a G/5mC bp at the following position. Box plot analysis revealed that the observed narrowing or widening of the minor groove upon CpG methylation depended on the sequence composition of As and Ts in the A-tract. For example, consecutive mutation from A to T in AAAAC led to a bell-shaped MGW profile, due to the introduction of a flexible TpA "hinge" step [47]. Maximal narrowing of the minor groove upon CpG methylation was observed for AATTC (Fig. 4b). This result might be due to the fact that this particular A-tract had a narrow minor groove, an effect that was amplified through cytosine methylation in the adjacent CpG step. Effects of DNA methylation on MGW were larger and more variable for 4-bp than for 3-bp A-tracts. This result was likely due to the more distinct minor groove narrowing of longer A-tracts and suggests that the methylation effect can be amplified depending on the A-tract features of the surrounding sequence.

Bulky methyl groups introduced by CpG methylation subtly widened the major groove and, in turn, narrowed the minor groove [22]. This observation can be explained in part by the proximity to the phosphate backbone of the methyl group of 5mC [22]. Narrowing of the minor groove enhances the negative electrostatic potential and, thereby, attracts minor groove-binding basic side chains more efficiently [22, 25]. This mechanism could potentially be employed when A-tracts reside in vicinity of CpG dinucleotides, as previously reported for various methyl group-binding proteins that use arginine-carrying AT-hooks [48] to recognize A-tracts adjacent to a CpG-containing motif [11].

Application of *methyl*-DNAshape predictions: modeling of DNase I cleavage activity

The DNA shape-dependent mechanism by which DNase I cleaves naked genomic DNA [22] serves as appropriate test system for assessing the functional relevance of our predictions of methylation-induced shape changes. In particular, the hexamer-based model (3-bp up- or downstream of the phosphate cleavage site) explained most of the variance in cleavage rates (Additional file 9:

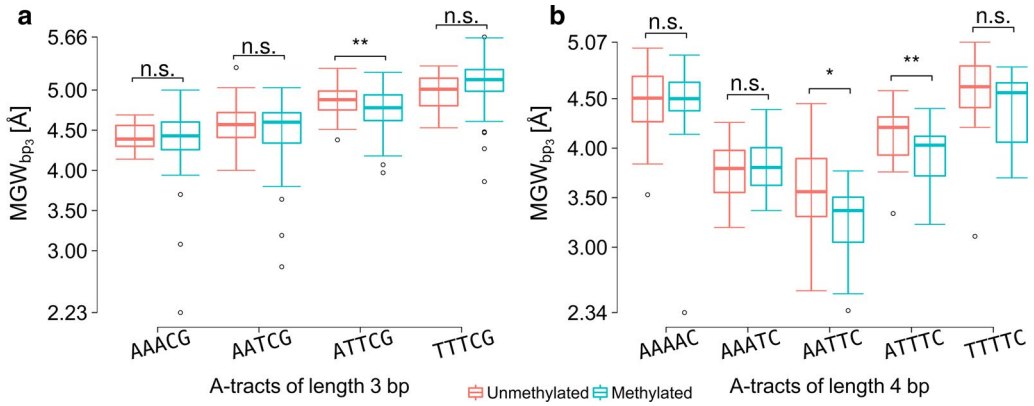


Fig. 4 Effect of CpG methylation on minor groove width (MGW) of adjacent A-tracts. **a** MGW values at the central nucleotide of 3-bp A-tracts, which are shown from AAACG to TTTCG with an exchange of one bp (A/T to T/A) from the 3′ end. Methylation did not decrease MGW at the central bp, except in the ATTCG sequence. Wilcoxon test *P* values were calculated for methylation narrowing the minor groove at the central nucleotide as the alternative hypothesis (*0.01 < *P* value ≤ 0.05; **0.001 < *P* value ≤ 0.01). Four A-tracts followed by a CpG step at the 3′ end include A-tracts preceded by a CpG step at the 5′ end because of symmetry in sequence and cytosine methylation. **b** MGW at the central nucleotide of 4-bp A-tracts follows a bell-shaped curve from AAAAC to TTTTC. One bp at a time was exchanged from A/T to T/A, starting at the 3′ end. Paired *t* test *P* values were calculated for methylation narrowing the minor groove at the central bp as the alternative hypothesis. Two pentamers, AATTC and ATTTC, showed significant *P* values, meaning that methylation narrowed the minor groove. MC simulations were performed on longer DNA fragments containing hexamer sequences with a CpG/mpg bp step at position 5, and MGW values were measured at the central position 3

Table S4; Additional file 10: Table S5). Enhanced cleavage by DNase I was observed for hexamers containing a CpG step at the + 1/+ 2 positions (referred to as $C_{+1}G_{+2}$ or positions 4 and 5 in a hexamer from the 5′ direction) immediately adjacent to the central cleavage site (Fig. 5a).

To assess how methylation-induced shape changes relate to the binding free energy ($\Delta\Delta G/RT$) of DNase I, we developed shape-based statistical models for unmethylated DNA (Fig. 5b). We used hexamers with an observed cleavage count of at least 25 to build our predictive models (Additional file 1). Next, we evaluated how well the resulting linear model predicted the effect of methylation on DNase I binding/cleavage ($\Delta\Delta\Delta G/RT = \Delta\Delta G/RT_{methylated} - \Delta\Delta G/RT_{unmethylated}$) in terms of the effect of methylation on shape ($\Delta shape = shape_{methylated} - shape_{unmethylated}$) (Additional file 1).

To evaluate the predictive power of each individual shape feature, we trained models based on each shape feature category and plotted the predicted $\Delta\Delta G$ shift against the maximum observed $\Delta\Delta G$ shift for a $C_{+1}G_{+2}$

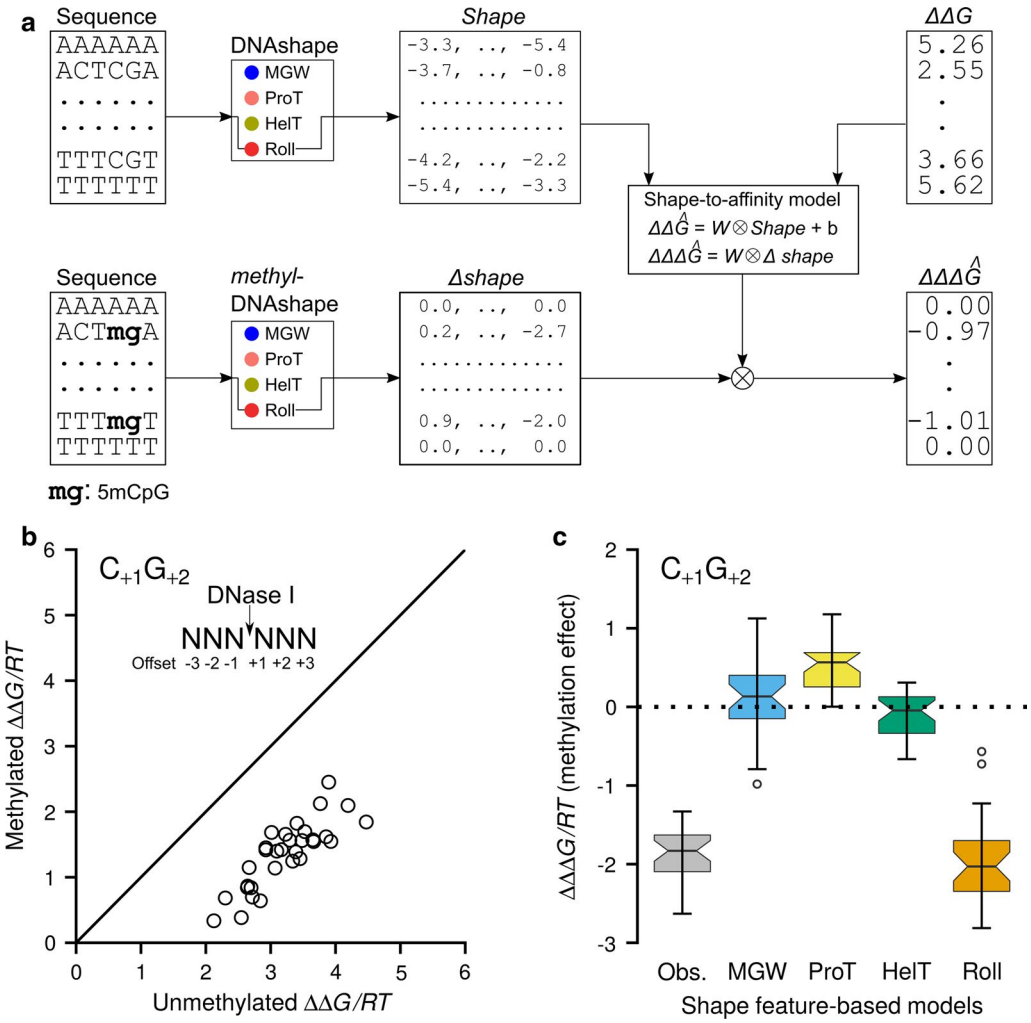


Fig. 5 Modeling of methylation-induced shifts in cleavage rates using methylation-induced shifts in shape feature profile. **a** Points on plot represent inferred binding free energy ($\Delta\Delta G/RT$) values of DNase I to unmethylated hexamers and corresponding methylated hexamers with absolute phosphate cleavage count ≥ 25. Methylation-induced effects are shown for sequences with $C_{+1}G_{+2}$ offset. Shift (downward) from diagonal indicates log-fold increase in cleavage activity of DNase I for methylated hexamers. **b** Shape-to-affinity modeling and use of *methyl*-DNAshape features. Shape-to-affinity model (L1- and L2-regularized linear regression model) built using unmethylated data. DNA shape features for unmethylated hexamers and their corresponding free energies ($\Delta\Delta G/RT$) were used as predictors and response variables, respectively. The model used the methylation effects on shape features (Δshape) calculated by *methyl*-DNAshape to predict $\Delta\Delta\Delta G$ (methylation effects on free energy, indicated by $\Delta\Delta\Delta\hat{G}$). Linearity of the model allowed direct use of Δshape as input variable. Roll values are shown for illustration purposes. **c** Predictive powers of different shape-based models. Observed $\Delta\Delta\Delta G/RT$ with median around − 2 is shown in gray colored box. Roll-based model accurately predicts the cleavage bias for $C_{+1}G_{+2}$ offset

offset (Fig. 5c). The Roll-based model better explained the shift than models based on other shape features. This observation may reflect the causal effect of the influence of methylation on DNA shape features (Fig. 3).

We observed an enhanced negative value (− 0.187) at the + 1/+ 2 offset in the weight vector $\boldsymbol{W}$ (Fig. 5b) of the Roll-based model. This finding suggested that the methylation-induced increase in Roll at this CpG offset caused a decrease in $\Delta\Delta G$ and, thus, an increase in binding affinity. For the $C_{+1}G_{+2}$ offset, the observed $\Delta\Delta G$ shift was well predicted by the change in Roll (Fig. 5c and Additional file 1). Compared to earlier work that was limited to MC simulations of a restricted set of methylated-DNA fragments [25], the *methyl*-DNAshape approach presented here enables systematic probing of the methylation effect for any CpG offset, number of sequences, or entire genomes.

CpG methylation effects on DNA binding of human Pbx-Hox complexes

In previous reports, SELEX-seq profiling followed by DNA shape analyses of binding by heterodimers of all eight *Drosophila melanogaster* Hox proteins in complex with their common co-factor Extradenticle (Exd) revealed an important role for MGW readout [26, 49]. More recently, an extension of the SELEX-seq method for methylated binding sites, EpiSELEX-seq, revealed that cytosine methylation modulates the affinity with which human orthologs (Pbx-Hox) of these heterodimers bind to CpG dinucleotide-containing sites [13]. The DNA sequences associated with the largest binding affinity for the Exd-Hox and Pbx-Hox complexes matched the 12-bp sequence pattern NTGAYNNAYNNN, where Y represents pyrimidine (C or T) and N any nucleotide (Fig. 6a).

As previously reported [13], direct comparison of the relative binding affinities for unmethylated versus methylated sequences (Fig. 6b, c) shows that cytosine methylation can either have a stabilizing or destabilizing effect on Pbx-Hox binding, depending on the position of the CpG dinucleotide within the binding site. For example, methylation of a CpG dinucleotide at offset 6/7 (NTGAYCGAYNNN; C_6G_7; green points/box in Fig. 6b, c) and offset 10/11 (NTGAYNNAYCGN; $C_{10}G_{11}$; blue points/box in Fig. 6b, c) suppresses binding, whereas methylation at offset 9/10 (NTGAYNNACGNN; C_9G_{10}; magenta points/box in Fig. 6b, c) enhances binding by an order of magnitude. We previously proposed a plausible mechanism for the latter stabilizing effect, which we postulated to involve direct contacts to the methyl group in the major groove [13]. However, an explanation of the suppressed binding at the CpG offsets 6/7 and 10/11 was lacking (Fig. 6a).

No protein–DNA contact was observed in the co-crystal structure (PDB ID: 1PUF) [50] at offset 6/7. However, the nucleotides at offset 6/7 form a spacer located between two AY dinucleotides (Fig. 6a), which were previously shown to exhibit strong shape preferences. Specifically, minor groove narrowing at AY positions adjacent to the central spacer was shown to be associated with enhanced binding when the nucleotide sequence was varied for unmethylated DNA [26, 49]. Therefore, we hypothesized that a methylation-induced change in DNA shape near the CpG dinucleotide could affect binding affinity. We used the pentamer-based shape tables that form the foundation of DNAshape [33] and *methyl*-DNAshape to investigate this effect systematically.

A pentamer window centered at the A_8 position includes a CpG dinucleotide at offset 9/10 within its 5 bp (NNGAYNNA**CG**NN). However, a CpG step at offsets 6/7 and 10/11 only includes one bp of the CpG dinucleotide (NNGAY**CG**AYNNN or NNGAYNNAY**CG**N) and indirectly constrains the nucleotide identity at a sixth position after the pentamer window. This distinction became important when we predicted MGW. In the case of the methylated-DNA table (*m*PQT), the presence of a (methylated) C at position 5 within the pentamer implies the presence of a G at the following position in the training set from which the pentamer tables were derived. This prediction is not the case for the unmethylated-DNA table (PQT). The pentamer tables do not capture a weak dependency of shape on the sixth position, which confounds our estimate of the methylation effect on shape. For this reason, we compiled an additional table consisting of unmethylated-DNA shape parameters for all hexamers ending with CpG and heptamers with CpG flanks (Additional file 1), which we used to estimate the effect of methylation on shape. Figure 6d shows that cytosine methylation in a sequence context consistent with the presence of a CpG step at offset 6/7 or 10/11 within the 12-bp Pbx-Hox binding site results in widening of the minor groove (see Additional file 1 for details on statistical tests performed). This observation, combined with the known inverse relationship between MGW and binding affinity for unmethylated DNA, provides a plausible explanation for the methylation-induced weakening of binding observed at these offsets (Fig. 6b). In contrast, no effect of methylation on MGW can be observed for the CpG offset 9/10, where direct contacts in the major groove already provided a mechanistic explanation [13].

Conclusions

Mechanisms of protein–DNA recognition remain incompletely understood. This lack of knowledge is particularly true for the readout of methylated DNA [15], despite its important role in gene regulation [22]. DNA sequence

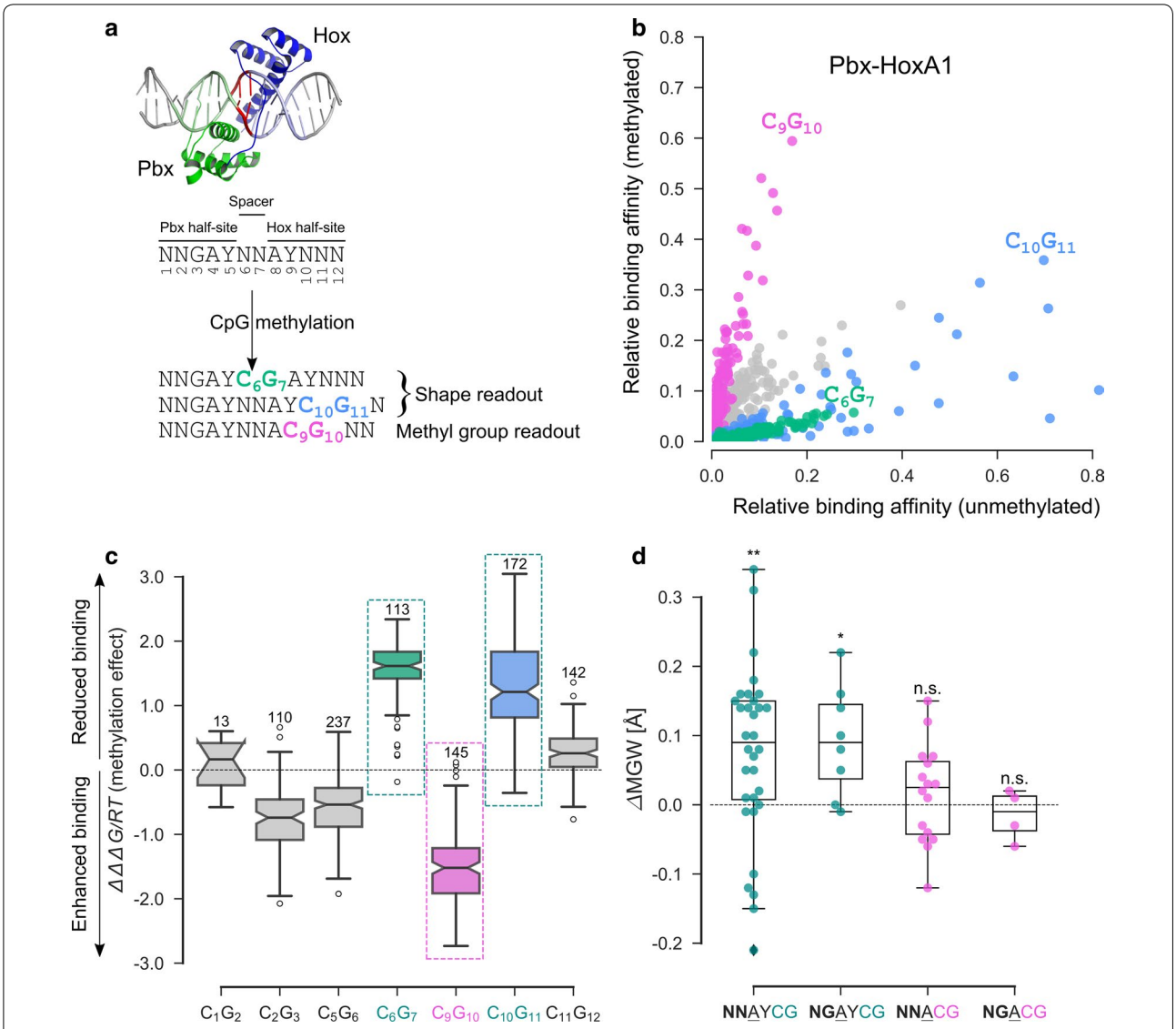


Fig. 6 CpG methylation induces a DNA shape change that explains its effect on Pbx-Hox binding. **a** Schematic representation of Pbx-Hox heterodimer bound to DNA (PDB ID 1PUF), and of the effect of CpG methylation on binding. Pbx (green) and Hox (blue) homeodomains bind up- and downstream of the central spacer region (indicated in red), respectively. CpG methylation at offsets 6/7 and 10/11 reduces binding, whereas methylation at offset 9/10 enhances binding. Methyl group readout was previously identified as underlying mechanism for the latter offset [13]. **b** Scatter-plot representation of relative binding affinities of methylated versus unmethylated sequences for Pbx-HoxA1 complex. Sequences carrying a single methylation event and their corresponding unmethylated part were considered. Green, magenta, and blue points correspond to methylation at offsets 6/7, 9/10, and 10/11, respectively. Sequences containing CpG dinucleotides at other offsets (relatively weakly affected by methylation) are colored gray. **c** Alternative representation of the data in **b**, showing the effect of methylation on binding free energy, denoted as $\Delta\Delta\Delta G/RT$. Positive (e.g., offsets 6/7 and 10/11) and negative (e.g., offset 9/10) shifts from the dashed line (indicating no methylation effect) reflect reduced and enhanced binding (on logarithmic scale) due to methylation. CpG dinucleotides at offsets 6/7 and 10/11 produce the same hexamer context for A_4 and A_8 (NNAYCG/NGAYCG) and, hence, were assigned a common color, dark-cyan. **d** Analysis of the methylation-induced change in MGW at positions A_4 and A_8 within the Pbx-Hox binding site (NNGAYNNAYNNN), for the different hexameric/pentameric contexts that the Pbx-Hox heterodimer may encounter within its binding sequence. Coloring corresponds to that of labels and rectangular patches in **c**. Statistically significant widening of minor groove (first two boxes) plausibly explains the observed reduced binding due to methylation at CpG offsets 6/7 and 10/11. No significant change in MGW upon methylation was observed for offset 9/10

and shape readout are key factors in achieving TF binding specificity. For base readout, presence of a bulky hydrophobic methyl group in the major groove may facilitate hydrophobic contacts with protein side chains [17]. For shape readout, local structural changes of the double helix induced by cytosine methylation may strengthen or weaken protein contacts to DNA [25]. Here, we describe an approach to probe and comprehend the shape readout mechanism of methylated DNA. As a high-throughput approach for predicting shape features of methylated DNA, our *methyl*-DNAshape method can be used to determine how the intrinsic shape of chemically modified DNA mediates recognition by TFs. Moreover, this method overcomes the limitation of the unavailability of experimental structures containing methylated cytosine.

One possible application of our method is to utilize high-throughput predictions of DNA shape features in quantitative models of protein–DNA binding. We found that the predicted change in shape features due to methylation partially explained the magnitude and context dependence of the experimentally measured effect of CpG methylation on DNase I cleavage [25]. Moreover, we were able to explain previously unexplained effects of DNA methylation on the binding specificity of human Pbx-Hox complexes. This study, therefore, represents a step forward toward a full mechanistic understanding of gene expression regulation.

Additional files

Additional file 1: Supplementary methods. Methodology for structure comparison, MC simulations, pentamer model, DNase I cleavage analysis, and statistical tests.

Additional file 2. Table S1. Types of DNA fragments and their counts considered in MC simulations.

Additional file 3: Table S2. Variables considered in MC simulations.

Additional file 4: Figure S1. Shape vector calculation.

Additional file 5: Table S3. Count breakdown of unique pentamer entries in *methyl*-DNAshape Pentamer Query Table (*m*PQT).

Additional file 6: Figure S2. Use of CpG context table in ΔMGW prediction.

Additional file 7. Sequence pool used in all-atom MC simulations.

Additional file 8: Figure S3. MGW profiles for selected DNA fragments or protein–DNA complexes.

Additional file 9: Table S4. Data preprocessing of DNase I cleavage data.

Additional file 10: Table S5. DNase I cleavage data in hexamer context.

Authors' contributions

SR, HJB, and RR conceived and designed the project. SR generated and analyzed MC simulation data, developed, and validated the *methyl*-DNAshape method, and performed statistical analyses. TPC updated the DNAshapeR/Bioconductor package to predict shape features of methylated DNA. SR, JFK, RSM, and HJB analyzed Pbx-Hox SELEX-seq and DNA shape data. SR, HJB, and RR wrote the manuscript with contributions from all authors. HJB and RR supervised the project. All authors read and approved the final manuscript.

Author details

[1] Computational Biology and Bioinformatics Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA. [2] Department of Biological Sciences, Columbia University, New York, NY 10027, USA. [3] Department of Systems Biology, Columbia University, New York, NY 10032, USA. [4] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA. [5] Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA. [6] Department of Neuroscience, Columbia University, New York, NY 10027, USA. [7] Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA. [8] Department of Physics & Astronomy, University of Southern California, Los Angeles, CA 90089, USA. [9] Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA.

Acknowledgements

The authors thank all members of the Mann, Bussenmaker, and Rohs laboratories for valuable input.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets analyzed in the current study are available in GEO under Accession Numbers GSM723024 (DNase I cleavage data) [25] and GSE98652 (Pbx-Hox SELEX-seq data) [13]. Additional data and code used for analyses are available at https://doi.org/10.5281/zenodo.1119141 and https://doi.org/10.5281/zenodo.1117976. Methylated-DNA query tables are available for download at http://rohslab.usc.edu/EPCH2018/.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

This work was supported by the National Institutes of Health (Grants R01HG003008 to HJB and RR; R01GM106056 to RR; R35GM118336 to RSM), an Andrew Viterbi Fellowship (to SR), a Manning Endowed Fellowship (to TPC), and a Howard Hughes Medical Institute International Student Research Fellowship (to JFK). RR is an Alfred P. Sloan Research Fellow.

Publisher's Note

Received: 29 July 2017 Accepted: 15 January 2018
Published online: 06 February 2018

References

1. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet. 1975;14:9–25.
2. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017;18:441–51.
3. Costello JF, Plass C. Methylation matters. J Med Genet. 2001;38:285–303.
4. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39:311–8.
5. Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev. 1993;3:226–31.
6. Brown CJ, Greally JM. A stain upon the silence: genes escaping X inactivation. Trends Genet. 2003;19:432–8.
7. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13:705–19.
8. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

9. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
10. Hayatsu H. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis: a personal account. Proc Jpn Acad Ser B Phys Biol Sci. 2008;84:321–30.
11. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356:eaaj2239.
12. Mann IK, Chatterjee R, Zhao J, He X, Weirauch MT, Hughes TR, et al. CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. 2013;23:988–97.
13. Kribelbauer JF, Laptenko O, Chen S, Martini GD, Freed-Pastor WA, Prives C, et al. Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. Cell Rep. 2017;19:2383–95.
14. Zuo Z, Roy B, Chang YK, Granas D, Stormo GD. Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. Sci Adv. 2017;3:eaao1799.
15. Buck-Koehntop BA, Stanfield RL, Ekiert DC, Martinez-Yamout MA, Dyson HJ, Wilson IA, et al. Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. Proc Natl Acad Sci. 2012;109:15229–34.
16. Spruijt CG, Vermeulen M. DNA methylation: old dog, new tricks? Nat Struct Mol Biol. 2014;21:949–54.
17. Liu Y, Zhang X, Blumenthal RM, Cheng X. A common mode of recognition for methylated CpG. Trends Biochem Sci. 2013;38:177–83.
18. Razin A, Riggs A. DNA methylation and gene function. Science. 1980;210:604–10.
19. Iguchi-Ariga SMM, Schaffner W. CpG methylation of the cAMP- responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev. 1989;3:612–9.
20. Boyes J, Bird A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell. 1991;64:1123–34.
21. Kemme CA, Marquez R, Luu RH, Iwahara J. Potential role of DNA methylation as a facilitator of target search processes for transcription factors through interplay with methyl-CpG-binding proteins. Nucl Acids Res. 2017;28:29–56.
22. Dantas Machado AC, Zhou T, Rao S, Goel P, Rastogi C, Lazarovici A, et al. Evolving insights on how cytosine methylation affects protein-DNA binding. Brief Funct Genomics. 2015;14:61–73.
23. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, et al. DNA methylation presents distinct binding sites for human transcription factors. Elife. 2013;2:e00726.
24. Rishi V, Bhattacharya P, Chatterjee R, Rozenberg J, Zhao J, Glass K, et al. CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. Proc Natl Acad Sci USA. 2010;107:20311–6.
25. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci USA. 2013;110:6376–81.
26. Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, et al. Deconvolving the recognition of DNA shape from sequence. Cell. 2015;161:307–18.
27. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein–DNA recognition. Nature. 2009;461:1248–53.
28. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein–DNA recognition. Annu Rev Biochem. 2010;79:233–69.
29. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.
30. Spivak AT, Stormo GD. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic Acids Res. 2012;40:162–8.
31. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016;44:D110–5.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000;28:235–42.
33. Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 2013;41:W56–62.
34. Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordân R, et al. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 2014;42:D148–55.
35. Chiu TP, Yang L, Zhou T, Main BJ, Parker SCJ, Nuzhdin SV, et al. GBshape: a genome browser database for DNA shape annotations. Nucleic Acids Res. 2015;43:D103–9.
36. Zhang X, Dantas Machado AC, Ding Y, Chen Y, Lu Y, Duan Y, et al. Conformations of p53 response elements in solution deduced using site-directed spin labeling and Monte Carlo sampling. Nucleic Acids Res. 2014;42:2789–97.
37. Lavery R, Zakrzewska K, Sklenar H. JUMNA (junction minimisation of nucleic acids). Comput Phys Commun. 1995;91:135–58.
38. Mak CH. Loops MC: an all-atom Monte Carlo simulation program for RNAs based on inverse kinematic loop closure. Mol Simul. 2011;37:537–56.
39. Rohs R, Sklenar H, Shakked Z. Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. Structure. 2005;13:1499–509.
40. Aduri R, Psciuk BT, Saro P, Taniga H, Schlegel HB, SantaLucia J. AMBER force field parameters for the naturally occurring modified nucleosides in RNA. J Chem Theory Comput. 2007;3:1464–75.
41. Lavery R, Sklenar H. Defining the structure of irregular nucleic acids: conventions and principles. J Biomol Struct Dyn. 1989;6:655–67.
42. Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016;32:1211–3.
43. Pérez A, Castellazzi CL, Battistini F, Collinet K, Flores O, Deniz O, et al. Impact of methylation on the physical properties of DNA. Biophys J. 2012;102:2140–8.
44. Haran TE, Mohanty U. The unique structure of A-tracts and intrinsic DNA bending. Q Rev Biophys. 2009;42:41–81.
45. Koo HS, Wu HM, Crothers DM. DNA bending at adenine. thymine tracts. Nature. 1986;320:501–6.
46. Nelson HC, Finch JT, Luisi BF, Klug A. The structure of an oligo(dA). oligo(dT) tract and its biological implications. Nature. 1987;330:221–6.
47. Hizver J, Rozenberg H, Frolow F, Rabinovich D, Shakked Z. DNA bending by an adenine–thymine tract and its role in gene regulation. Proc Natl Acad Sci USA. 2001;98:8490–5.
48. Aravind L, Landsman D. AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res. 1998;26:4413–21.
49. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, et al. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. Cell. 2011;147:1270–82.
50. LaRonde-LeBlanc NA, Wolberger C. Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. Genes Dev. 2003;17:2060–72.
51. Shannon P, & Richards M. MotifDb: An annotated collection of protein–DNA binding sequence motifs. R package 2017; version 1.20.0.