A direct readout mechanism regulates the sequence selectivity of human *de novo* DNA methyltransferases

Ayşe Berçin Barlas[1,2], Ezgi Karaca[1,2]

[1]Izmir Biomedicine and Genome Center, 35340, Izmir, Turkey

[2]Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, 35340, Izmir, Turkey

Correspondence: Ezgi Karaca, ezgi.karaca@ibg.edu.tr

ABSTRACT

Mammalian *de novo* DNA methylation is integral to gene regulation. This epigenetic modification is exerted by DNMT3A and DNMT3B enzymes predominantly on CpG islands. Over their enzymatic domains, DNMT3A/B shares over 90% of sequence similarity. Even so, DNMT3A prefers to methylate preferentially CGC, while DNMT3B prefers CGG sequences. To elucidate the mechanistic basis of this selective methylation profile, we performed extensive molecular dynamics simulations of DNMT3A/B enzymes bound to all possible CGX[C/G/T/A] motifs. As an outcome, we observed that the DNMT3 sequence selectivity at +2 position relies on the likelihood of a stable Arg/Lys-Guanine base-specific hydrogen bond formation. In the case of DNMT3A, Arg836 forms stable hydrogen bonds with a single guanine that is complementary to the second C in the CGC motif. In the case of DNMT3B, Arg836 is replaced by Lys777, where DNMT3B-Lys777 switches to hydrogen bonding with the two consecutive guanines in the CGG motif. The loss of branching in DNMT3B-Lys777 is further compensated by Lys777's cooperative action with Asn779. Our analysis also showed that the single nucleotide substitutions in CGX motifs lead to dramatic changes in the DNMT3A-DNA selective hydrogen bonding profile compared to the ones in DNMT3B. This finding is well correlated with the higher CpG specificity of DNMT3A compared to DNMT3B. All in all, we propose a simple explanation for a complex phenomenon, where the sequence selectivity of de novo DNMTs is directed by the guanine interaction capacity of these enzymes.

Keywords: *De novo* DNA methylation, DNMT3, sequence selectivity, protein-DNA interactions, molecular dynamics

INTRODUCTION

DNA methylation is an epigenetic mark introduced to regulate a diverse range of vital cellular processes in aging, embryonic development, gene expression, reprogramming, chromatin modification, and X chromosome inactivation (Dor & Cedar, 2018; Kaneda et al., 2004; Skvortsova et al., 2019). In humans, there are three catalytically active DNMT enzymes; DNMT1, DNMT3A, and DNMT3B. These enzymes predominantly methylate CpG dinucleotides (Schmitz et al., 2019; Vukic & Daxinger, 2019). DNMT3A and DNMT3B impose the methyl mark on genomic DNA from scratch (in a *de novo* manner) (Okano et al., 1999), while DNMT1 maintains an existing methylation pattern during DNA replication (Schneider et al., 2013).

Among these DNMT enzymes, DNMT3A/B-dependent methylation is orchestrated by highly similar paralogous methyltransferase (MTase) domains of DNMT3A/B, exhibiting a remarkable 91.2% sequence similarity over 283 residues (Supplementary Figure S1) (Shen et al., 2010; Dukatz et al., 2022). The DNMT3-MTase domains comprise catalytic and target recognition (TRD) loops, which play pivotal roles in locating and anchoring the methylation site upon interacting with the genomic DNA from minor and major grooves (Figure 1A). Once DNMT3-MTase is bound at the target site, the MTase flips the target cytosine, extracts a methyl group from the S-adenosyl-L-methionine (SAM) molecule and transfers it to the 5' carbon of the target cytosine (Lin et al., 2020; Schmitz et al., 2019; Vukic & Daxinger, 2019; Zhang et al., 2018) (Figure 1A). Despite the high sequence similarity and the conserved reaction pattern shared between DNMT3A/B-MTases, in-depth enzymology and gene knockout experiments have revealed striking distinct sequence preferences for these enzymes: DNMT3A exhibits a preference for methylating CpG sequences followed by a pyrimidine (C/T) at the +2 position (i.e., CGC/T), while DNMT3B favors a purine (G/A) nucleotide at the same +2 site (i.e., CGG/A). These experiments also noted CpGpC (CGC) sequences as 3A-MTase, and CpGpG (CGG) sequences as 3B-MTase cognate methylation sequences (Zhang et al., 2018; Lin et al., 2020; Gao et al., 2020; Mao et al., 2020; Mallona et al., 2021). Hereafter, 3A-MTase will be referred to as DNMT3A and 3B-MTase as DNMT3B for simplicity.

In the past decade, a handful of structural studies concentrated on resolving the atomic scale differences of DNMT3A/B to understand the molecular grounds behind their sequence selectivity patterns. To this end, the atomistic resolution structures of human DNMT3A/B bound to their cognate sequences (CGC and CGG) identified Arg836-to-Lys777 TRD loop substitution (DNMT3A-to-DNMT3B) as the primary determinant for +2 flanking sequence preference (Z.-M. Zhang et al., 2018; Emperle et al., 2019; Lin et al., 2020; Gao et al., 2020; Mallona et al., 2021) (Supplementary Figure S1). Specifically, these structures showed that when DNMT3A is bound to CGC, its Arg836 acts as a specificity-determining position (SDP) by forming a base-specific hydrogen bond with guanine's (G+1) O6 atom (Figure 1A-B) (Z.-M. Zhang et al., 2018). In the case of the cognate DNMT3B-CGG complex, the SDP arginine is substituted by a lysine, which engages in van der Waals interactions with G+2's N7 (Figure 1B) (Lin et al., 2020). In the same work, it was proposed that Arg-to-Lys substitution in DNMT3B is compensated with the involvement of a conserved amino acid, Asn779, where it forms a hydrogen bond with G+1's O6 (Gao et al., 2020).

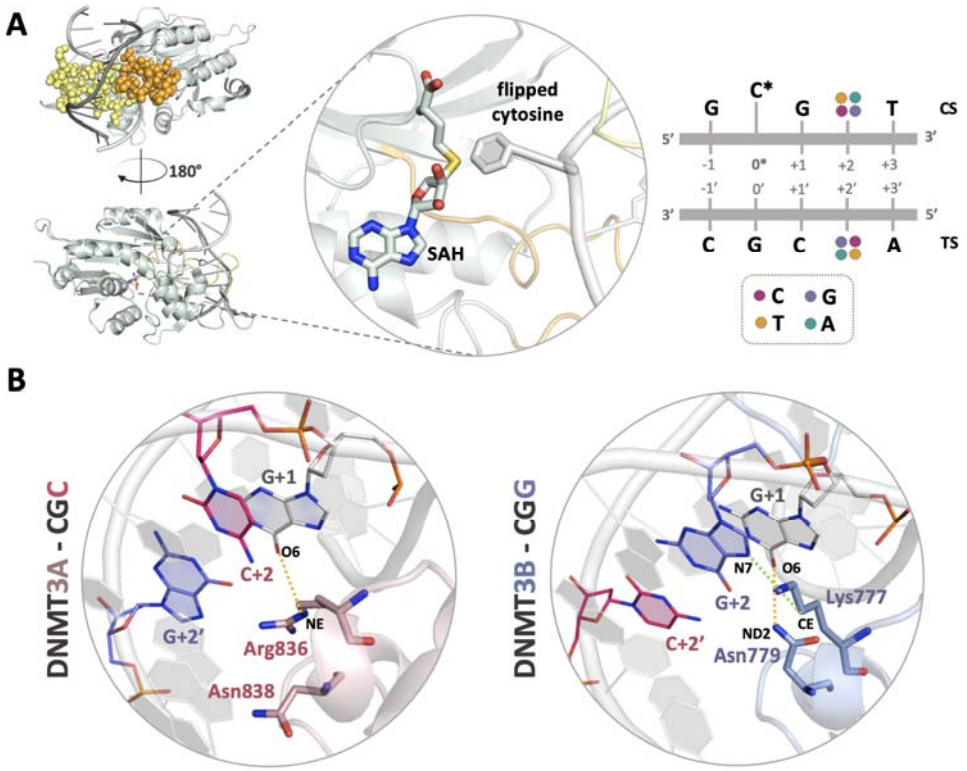

Figure 1: Structural features of DNMT3A and DNMT3B MTase domains. (A) Left: 3D structure of DNMT3A/B-MTase domain from the front (top) and back (bottom) sides. Orange region represents the catalytic loop and yellow region represents the target recognition domain (TRD) loop. The inset depicts the SAM methyl donor and the flipped target cytosine (the cytosine to be methylated) (PDB ID: 6KDA (Lin et al., 2020)). Right: Target DNA site coordinate system. TS and CS stand for target and complementary strands, respectively. 0* denotes the center of the sequence, which is the cytosine to be methylated. Nucleotide numbers with apostrophe (') state the complementary strand nucleotides. The nucleotide mutations explored in this study fall at the +2 and 2' positions. **(B)** Crystal structures of cognate human DNMT3A and DNMT3B complexes. Right: DNMT3A is bound to CGC (PDB ID: 6F57). In this crystal structure, Arg836 forms a base specific hydrogen bond with G+1 of CGC (yellow) (Zhang et al., 2018). Right: DNMT3B is bound to CGG (PDB ID: 6KDA). In this crystal structure, Lys777 forms van der Waals interactions with G+2 (green) (Lin et al., 2020) and Asn779 forms a hydrogen bond with G+1 of CGG (yellow) (Gao et al., 2020).

Apart from their cognate structures, the complexes of DNMT3A and DNMT3B with CGT and CGA sequences were also resolved (Anteneh et al., 2020; Zhang et al., 2018; Gao et al., 2020; Lin et al., 2020). In the DNMT3A–CGT complex, Arg836 forms a hydrogen bond with the O6 of the CpG guanine and a water-mediated contact with the N7 atom (Zhang et al., 2018). Conversely, in DNMT3B, Lys777 points towards T+2, and Asn779 donates a hydrogen bond to the O6 atom of G+1 (Gao et al., 2020). In the structure of DNMT3A bound to CGA, Arg836 forms a hydrogen bond with the N6 atom of the A+2 position, while Asn838 contacts the O6 atom of G+1 (Anteneh et al., 2020). Similarly, in the DNMT3B–CGA complex, Lys777 interacts with A+2, engaging in van der Waals contacts (Gao et al., 2020). The sequence selectivity of cognate DNMT3A-CGC and noncognate DNMT3A-CGG complexes were further explored with molecular dynamics (MD). The analysis of dynamic interaction profiles of these DNMT3A systems strikingly showed that, in the case of the cognate CGC complex, Arg836's G+1 interaction is instantly lost at the beginning of the simulation. Instead, Arg836 forms a stable base-specific interaction with the complementary G+2' (i.e., GCG) nucleotide (Mallona et al., 2021), explaining the +2 selectivity of this enzyme. The authors could also validate this experimentally.

The work of Mallona *et al.* unraveled that analyzing crystal structures was pivotal to unraveling the critical amino acids, i.e. SDPs responsible for selecting the cognate sequences. On the other hand, incorporating the dynamics perspective was necessary to elucidate the selective amino acid-nucleotide interactions. Building on this insight, we conducted the most comprehensive MD study performed to date on *de novo* DNA methyltransferases: we ran MD simulations of DNMT3A and DNMT3B bound to all conceivable nucleotide substitutions at the +2 position (i.e., CGC/G/T/A). We then devised and employed an Interaction Dynamics Framework (IDF) for a systematic exploration of base-specific hydrogen bonding interactions

across all complexes. As an outcome, we observed that the DNMT3 sequence selectivity +2 position relies on the likelihood of a stable Arg/Lys-Guanine base-specific hydrogen bond formation. We also showed that this direct readout mechanism is dictated by the DNA groove geometry. This straightforward yet compelling perspective proposes a simple explanation for a complex phenomenon, which can be directly applied to understand the partner selectivity of other paralogous DNA-binding proteins.

RESULTS

An Interaction Dynamics Framework to deliver selective direct readout signals

To understand the sequence selectivity of DNMT3 enzymes, we took the available DNMT3A/B-DNA structures as a basis and built eight DNMT3A/B-DNA complex models, i.e., DNMT3A-CGC/T/A/G and DNMT3B-CGG/T/A/C, where the DNA sequences differed only at the +2 position (Methods). This setup allowed us to concentrate on the interaction differences emanating from the +2 flanking nucleotides and differential DNMT3A/B amino acids (Supplementary Figure S1A). We simulated the DNMT3 complexes (DNMT3A/B-CGX) for 2 µs long (4x0.5 µs replicas) to cover a 16 µs long DNMT3-CGX dynamics space (Methods). We then devised an interaction dynamics framework (IDF), which performs a differential analysis of selective protein-DNA interactions upon primarily focusing on base-specific protein-DNA hydrogen bonding (direct readout) profiles, since these are the ones holding the specific binding signals (Figure 2A, Methods) (Deepa et al., 2011; Luscombe et al., 2001; Rohs et al., 2010).

Our IDF tool analyzes DNMT3-DNA direct readout interaction profiles in a stepwise manner with an increasing resolution. For this, it initially calculates all base-specific hydrogen bonding profiles of all complexes exerted throughout their simulation times. Afterward, it pools the data coming from replica simulations of each complex to cover a wider sampling space. Within the merged interaction lists, it systematically excludes SAM-DNA and DNMT3-methylated-cytosine interactions, since these stay strictly conserved across different complex sets. IDF finally discards all the interactions observed for less than 10% of the simulation time, to be able to eliminate the noisy ones.

Over the filtered interaction sets, IDF first compares the global base-specific hydrogen bond formation potentials across all complexes (Figure 2B). Then, it concentrates on the two nucleotides around the selective +2 position, the +1 to +2 flanking region, and calculates the interaction propensities at the residue/nucleotide level (Figure 2C). IDF finally zooms into the same flanking region interactions, but this time it specifically analyzes the hydrogen bonding patterns at the atomistic level (Figure 2D). While interpreting the atom-atom interactions, it considers the hydrogen acceptance and donation capabilities of

selected amino acids with the flanking bases that are dictated by the DNA groove geometry (Figure 2E) (Luscombe et al., 2001). Our IDF's input/output files and analysis scripts are deposited at https://github.com/CSB-KaracaLab/DNMT3AB_specificity. Below, we investigate the IDF results of DNMT3A and DNMT3B complexes separately.

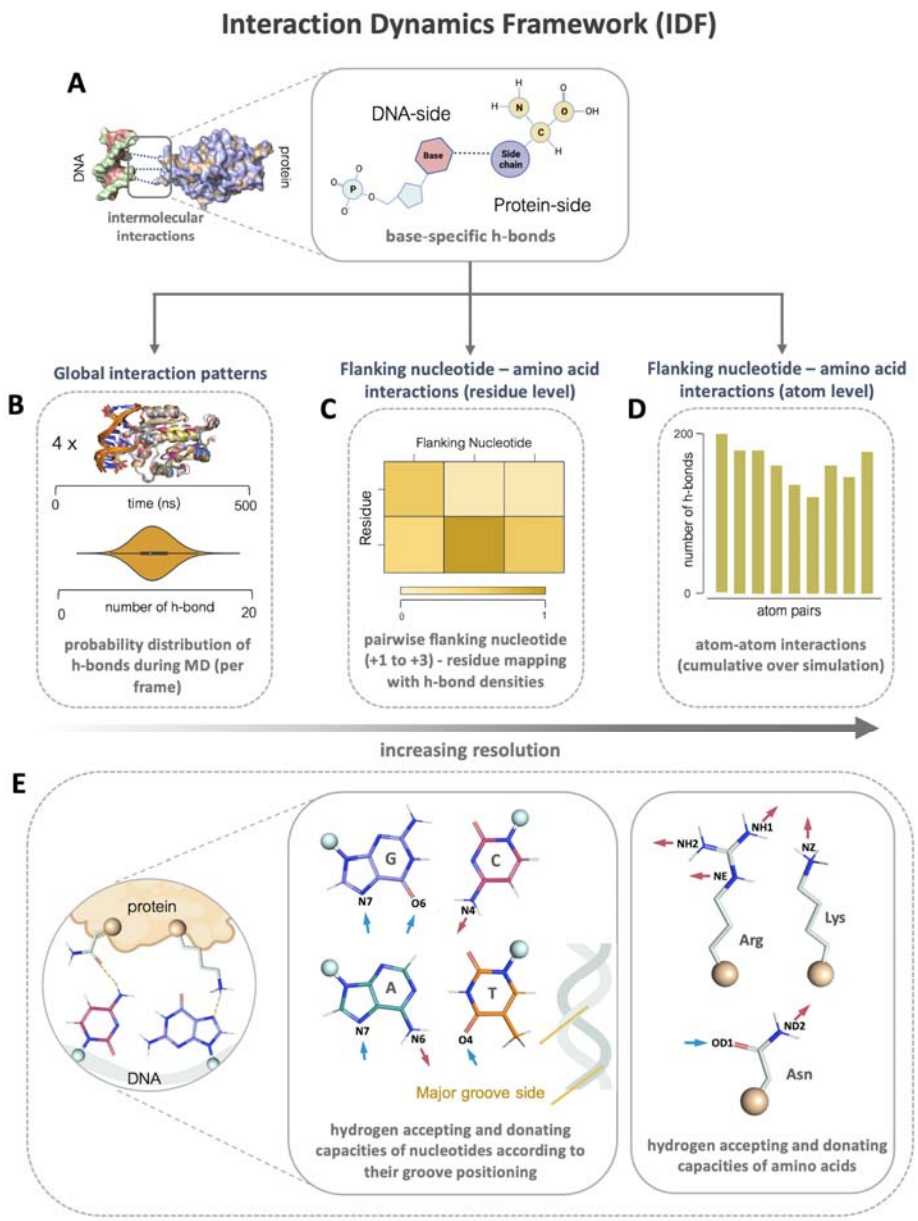

Figure 2: An interaction dynamics framework (IDF) to identify DNMT3 selectivity +2 flaking position. (A) Evaluation of intermolecular interactions between protein and DNA, in terms of base-specific hydrogen bonds. **(B)** Global interaction analysis of base-specific hydrogen bonds throughout 2 µs MD simulations for each DNMT3A/B system. The violin plots show the probability distributions of the number of h-bonds formed in each conformation generated during MD. **(C)** Residue level flanking nucleotide – amino acid interactions. Heatmap shows the mean hydrogen bonding capacity between observed amino acid-nucleotide pairs. Here we only focus on the +1 to +3 DNA interactions. **(D)** Atom level flanking nucleotide – amino acid interactions

plotted as frequency bar graphs. **(E)** Hydrogen accepting and donating capacities of nucleotides (from major groove) and important amino acids. Blue arrows represent hydrogen-accepting, where red arrows show hydrogen-donating positions. Blue spheres represent the attachment point of the bases to the backbone; gold spheres represent the attachment point of side chains to the backbone. Through the major groove: Guanine accepts two hydrogens at N7 and O6, cytosine donates a hydrogen from N4, adenine accepts and donates hydrogens at N7 and from N6, thymine accepts a hydrogen at O4. Arg and Lys donate hydrogen from their NH1, NH2, NE, and NZ atoms, while Asn both accepts and donates hydrogens at its OD1 and from ND2 atoms, respectively.

DNMT3A's target specificity is modulated by the hydrogen donation capacity of Arg836

We initially focused on the IDF analysis of DNMT3A-CGC/T/A/G systems. For this, as described above (Figure 2B), we calculated the base-specific hydrogen bonding probabilities of DNMT3A complexes. As an outcome, we could not observe any significant differences between the base-specific hydrogen bonding capabilities of DNMT3A-CGX complexes (the median h-bonding capacities varied between 3-4 bonds) (Supplementary Figure S2A). This led us to delve deeper into specific interactions formed around the selective +2 region (+1 to +3, as pictured in Figure 2C). Here, we counted all base-specific hydrogen bonds formed between any amino acid of DNMT3A with the flanking bases (+1 to +3) of DNA, followed by their normalization over the number of conformations generated (Figure 3A, Methods).

As an outcome, in the cognate DNMT3A-CGC complex, we observed a consistent hydrogen bonding between Arg836 and complementary G+2', with 1.9 intensity (Figure 3A). When we zoomed into the atom-atom frequency distribution of this specific interaction, we saw that NE and NH2 atoms of Arg836 form bidentate and stable hydrogen bonds with N7 and O6 atoms of G+2' (Figure 3A), explaining the normalized intensity value over 1.0. This strong interaction is observed during 70% of conformations. We observed another recessive interaction mode of this amino acid-nucleotide pair, which is made through a single hydrogen bonding established between NH1/2 groups of Arg836 and O6 of G+2'. All in all, in the cognate complex, we observed an extremely specific and strong interaction between arginine's flexible and bulky side chain with the most attractive nucleotide around that is the +2' guanine (Supplementary Figure S3). This cognate interaction mode is in agreement with the observations outlined in a previous MD study of the same cognate system (Mallona et al., 2021).

Setting the cognate DNMT3A interaction mode as the specific G+2' recognition, we moved to the DNMT3A-CGT complex. The mutation of C+2 to T strikingly leads to the total disappearance of the complementary strand contacts. Instead, a moderate-intensity hydrogen bonding is established with the target stand's G+1 (Figure 3B). For this sequence, the intensity drop compared to the cognate sequence is almost threefold (0.69 vs. 1.9). The atom-atom interactions of the CGT motif showed that Arg836 makes dampened and distributed h-bonds with its NH1, NH2, and NE atoms and the N7 and O6 atoms of G+1, reflecting a less stable and directed hydrogen bonding pattern compared to the cognate sequence. One such interaction mode is presented in Figure 3B, where Arg836's NH1 and NH2 atoms form bidentate hydrogen bonds with the N7 and O6 atoms of G+1. So, C-to-T mutation at the +2 site results in the loss of selectivity at the +2 position, since Arg836 cannot donate hydrogens to A+2', which leaves the G+1 as the most attractive spot for Arg836, though with a three-fold lower probability of hydrogen bond formation compared to the cognate Arg836-G+2' interaction mode.

Interestingly when DNMT3A is bound to the CGA motif, Arg836's interaction with the T+2' position reemerges, though this time with a much lower intensity compared to the cognate one (0.84 vs. 1.9) (Figure 3C). This happens since T+2' has the capacity to accept one hydrogen, which apparently is attractive enough for Arg836 to form a hydrogen bond. The examination of the pairwise atomistic interactions reveals a single binding mode between the T+2' and Arg836 pair, where O4 of T+2' and NH1 of Arg836 establishes a single and stable bond in almost 70% of the pooled conformations (Figure 3C). As a result, at the atomistic scale, the +2 selectivity in the case of CGA sequence is kept.

In all three previous cases, we only have Arg836 coming up as the specificity-determining amino acid that selectively interacts with a single nucleotide. We surprisingly observed that this pattern is lost when DNMT3A is in complex with its noncognate CGG sequence. When G is present at +2 position, the conserved Asn838 comes into play as the other amino acid recognizing the +1-to-+3 region. To interact with the CGG sequence, Arg836 and Asn838 exert two different tasks where Arg836 interacts with the G+1 and G+2 in the target strand and Asn838 with the C+2' and T+3' on the complementary strand. The interaction intensity values of Arg836, in this case, are the lowest ones observed compared to other DNMT3A complexes (0.28 and 0.49, Figure 3D). The atom-atom pairwise analysis

reveals that the Asn838 forms a higher frequency of hydrogen bonds (approx. in 40% of the conformers) compared to the bonds made by Arg836. At the atomic level, the highest frequency interaction corresponds to the one between the OD1 atom of Asn838 and the N4 atom of C+2'. In the context of the CGG sequence, Arg836 can still form bidentate hydrogen bonds with the G+2, though this time with an extremely low probability, since such bonds were observed in 15% of the conformations.

To assist these findings, we finally calculated the mean and standard deviation values of the DNA shape parameters sampled for each DNMT3A complex (major groove width and base pair step parameters, Supplementary Figure S6A). This analysis showed that the major groove width signature of the cognate CGC is lower (varied between 12-25 Å) compared to the other sequences. It also highlighted that the most variation in the base pair step parameters was observed for CGG and CGT sequences, aligning with the lost +2 selectivity observation, as explained above.

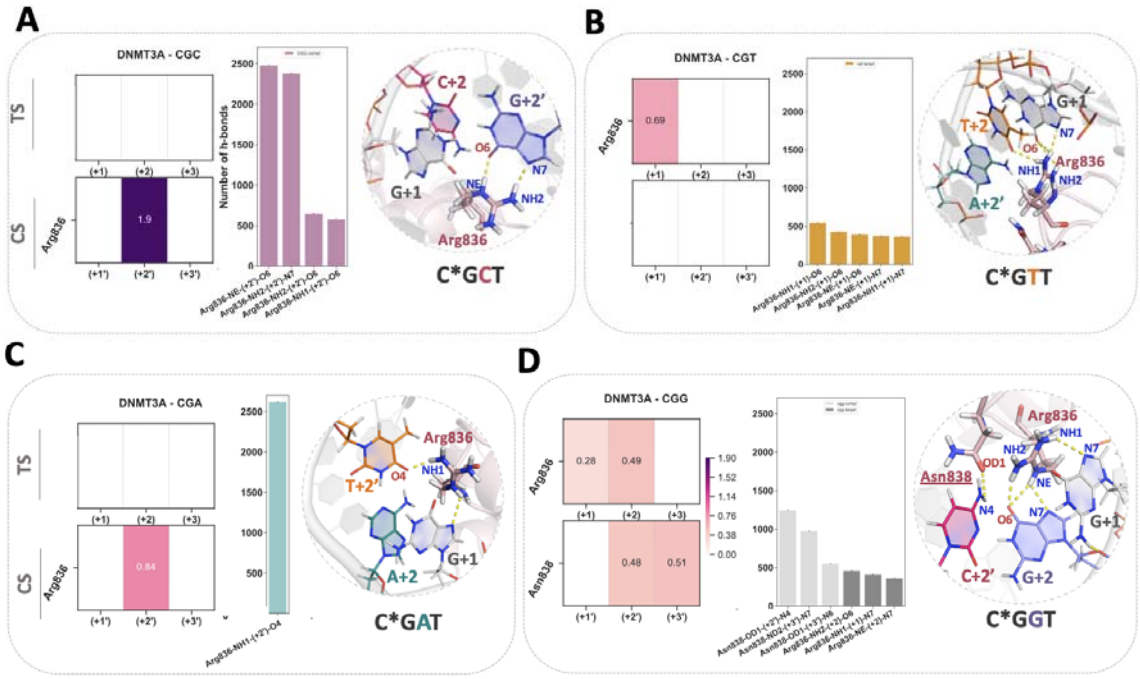

Figure 3: Interaction analysis of DNMT3A cognate (CGC), cognate-like (CGT) and noncognate systems (CGA/G). Flanking nucleotide interactions of DNMT3A complex with **(A)** CGC, **(B)** CGT, **(C)** CGA, **(D)** CGG. Heatmaps of positional base-specific interaction intensity of amino acids and flanking nucleotide positions of All interaction frequencies were normalized with respect to the number of frames (N=3204). The calculated values are mapped on the heatmap cells. In the heatmaps, the first row represents the target strand (TS) interactions, while the second row shows the complementary strand (CS) ones. Atomic details of base-specific hydrogen bonds formed between DNMT3A and CGC, CGT, CGA and CGG simulations given by barplots, accompanied by structural snapshots of the base-specific interactions formed by DNMT3A- Arg836 with +1, +2 and +2' flaking nucleotide positions. Each bar shows the interaction number during 2 µs MD simulations. Light colored bars represent the complementary strand interactions, whereas dark colored bars represent target

strand interactions. The x-axis of the plot shows the interacting protein residue, interacting side chain atom, nucleotide position and nucleotide base atom.

The loss of hydrogen donation capacity due to the Arg-to-Lys substitution is compensated by a conserved Asn

We applied our IDF protocol also on the DNMT3B complexes. The initial base-specific hydrogen bonding probabilities showed that the CGG sequence shares similar distribution characteristics with CGA and CGC systems in terms of median values, peak points, and probability density (8 hydrogen bonds per frame, Supplementary Figure S2B). In the case of the CGT complex, the distribution showed a higher probability, with a higher median value compared to other systems. Still, these probability distributions of the DNMT3B complexes did not report the essential factors for its sequence selectivity.

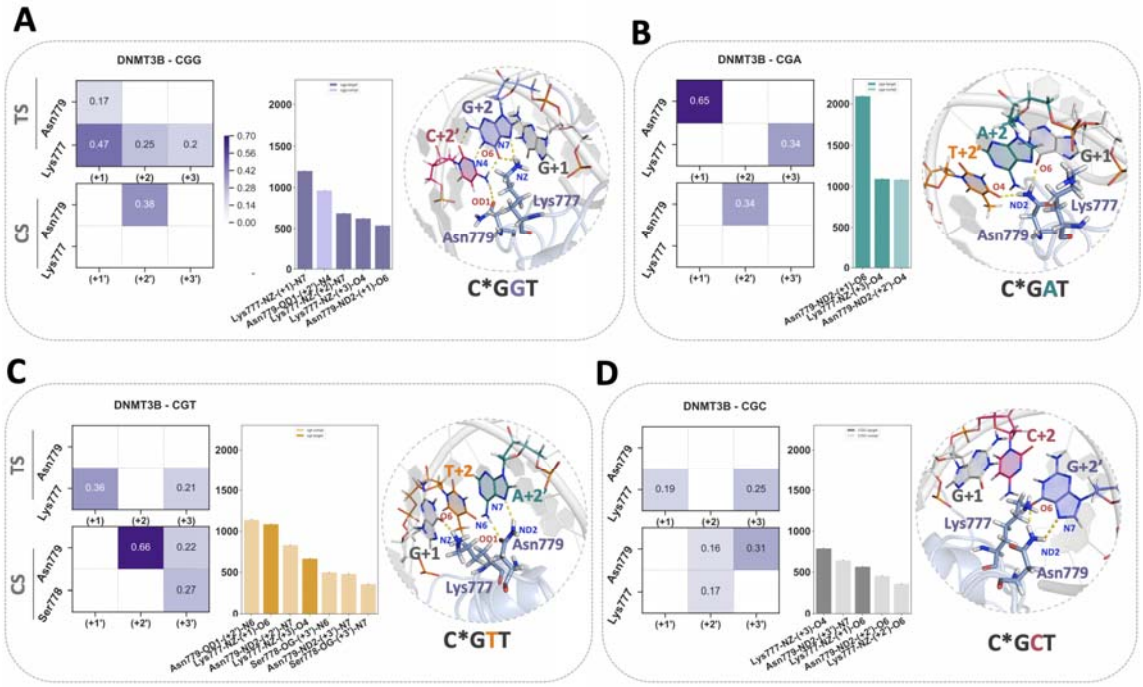

Figure 4: Interaction analysis of DNMT3B cognate (CGG), cognate-like (CGA) and noncognate systems (CGT/C). Flanking nucleotide interactions of DNMT3B complex with **(A)** CGG, **(B)** CGA, **(C)** CGT, **(D)** CGC. Heatmaps of positional base-specific interaction intensity of amino acids and flanking nucleotide positions of All interaction frequencies were normalized with respect to the number of frames (N=3204). The calculated values are mapped on the heatmap cells. In the heatmaps, the first row represents the target strand (TS) interactions, while the second row shows the complementary strand (CS) ones. Atomic details of base-specific hydrogen bonds formed between DNMT3B and CGG, CGA, CGT and CGC given by barplots, accompanied by structural snapshots of the base-specific interactions formed by DNMT3B-Lys777 and Asn779 with +1, +2 and +2′ flaking nucleotide positions. Each bar shows the interaction number during 2 µs MD simulations. Light colored bars represent the complementary strand interactions, whereas dark colored bars represent target

strand interactions. The x-axis of the plot shows the interacting protein residue, interacting side chain atom, nucleotide position and nucleotide base atom.

When we moved to the residue-specific analyses, we observed that, in the cognate CGG motif, Lys777 forms base-specific hydrogen bonds with the +1, +2, and +3 positions on the target strand with moderate to low intensities (ranging from 0.43 to 0.17). Notably, we did not observe any complementary strand interactions of Lys777 (Figure 4A). In this DNMT3B cognate complex, the selective +2' complementary strand position is recognized by the conserved Asn779 (corresponding to Asn838 of DNMT3B). Asn779 formed specific hydrogen bonds +1 and +2' positions with a moderate intensity (0.38). The CGG interaction network spanned by Lys777 and Asn779 is strikingly very similar to the one we observed for the DNMT3A-CGG complex, complementary C+2' cannot accept hydrogen through the major groove (Figures 2E and 4D). C+2' N4 hydrogen donor is recognized through Asn779's hydrogen acceptor OD1, which comes up as the second dominant interaction pair observed for this complex (observed in ~30% of conformations, Figure 3A). The first one is formed between Lys777's NZ and G+1's N7 for almost 40% of the conformations. G+1 was further stabilized with Asn779's hydrogen donor ND2 and G+1's free hydrogen acceptor O6. Importantly, Lys777 formed base-specific hydrogen bonds with G+1 and G+2 nucleotides through its hydrogen donor NZ (Figure 4A). All in all, in the DNMT3B cognate sequence, we observed that a cooperative recognition between Lys777-G+2 and Asn779-C+2' describes the specificity of this complex.

Upon mutating the +2 position to A, Lys777 shifts to form specific hydrogen bonds only with T+3 position, with a low intensity (0.34, Figure 4B). In this complex, G+1 recognition is handed to Asn779 (0.65 intensity), while Asn779 still keeps its role for the recognition of +2' with the same intensity as in the CGG complex. In this case, Asn779 acts as a hydrogen donor to the +2' position during 30% of conformers. As G+1 is not occupied by Lys777 for the CGA complex, G+1 becomes an attractive hydrogen donation target for Asn779, as observed in 65% of the conformers. O4 of T+1 could be reached by Lys777 in 30% of the conformers, making it the only interaction of Lys777 with the target strand (Figure 4B). The interaction between Asn779 and T+2' was also highlighted by Gao et al., 2020.

When we moved to the CGT motif, Asn779 started to form a specific hydrogen bond with A+2' with the highest intensity (0.66), compared to the other complexes. Asn779 also makes low-intensity bonds (0.22) with A+3'. Interestingly, the atom-atom interaction analysis revealed that Asn779 and adenine at +2' and 3' positions use their dual hydrogen donation and acceptance capacity while making hydrogen bonds (Figure 2E, Figure 4C). Strikingly, we start seeing a new amino acid coming into play, i.e., Ser778, which forms h-bonds through its OG atom, both accepting hydrogen from the N6 atom of A+3' and donating hydrogen from the N7 atom of A+3'. At the atomistic scale, Lys777 forms hydrogen bonds with the G+1-O6 atom (Figure 4C). When we compared the replica simulations, Lys777 reads CGT in two different conformations: it reads +4' guanine and +3 thymine, or it reads +1 guanine (Supplementary Figure S4). The structural snapshots of the MD simulation showed that Asn779 forms bidentate hydrogen bonds with A+2' with both donating and accepting hydrogen, so the interaction intensity increases (Figure 4C).

In the non-cognate case of DNMT3B-CGC, Lys777 interactions are distributed to +1 and +3 like the other motifs (varying between 0.19-0.25 intensities). It also forms h-bonds with +2' position, possibly due to the attractive field posed by G+2'. This interaction has very low intensity (0.17). Asn779 continues to form hydrogen bonds with +2', like the other motifs, however, the intensity of this interaction decreased 2.4-fold compared with the cognate motif and 4.1-fold compared with the CGT. In this motif, the interactions of Asn779 with the target strand are entirely lost and shifted to G+2'/A+3' (Figure 4D). In atomic detail, we observed that Lys777 forms hydrogen bonds with the O6 atom of G+1 and O4 of T+3. Furthermore, Lys777 also goes beyond the sequence space we concentrated on and turns toward G+4' and forms h-bonds for 75% of the total simulation time (Supplementary Figure S4). Notably, one of the interesting interactions of Lys777 is a hydrogen bond with the O4 atom of T+3. All in all, while Asn779 interprets +2' guanine and +3' adenine, Lys777 reads +4' guanine. In addition, for the remaining 25% of the total simulation time, Lys777 can interpret +2' guanine and +1 guanine simultaneously (Supplementary Figure S4).

Following the h-bond analysis at atomic resolution, we calculated ionic bonds and hydrophobic interactions of the DNMT3B systems as well. In DNMT3B, the number of interactions per frame in ionic and hydrophobic interactions exhibit a similar distribution profile in both the cognate and non-cognate motifs (Supplementary Figure S5B). As we

performed for DNMT3A, we examined the major groove width and base-pair step parameters of DNMT3B bound DNAs to look at the relationship of interaction patterns with DNA shape. Major groove width of the CGG motif, which represents the cognate sequence, was observed to be larger than other motifs (17-19 Å) (Supplementary Figure S6A). The base pair step parameters at the DNA ends showed similar error ranges for the CGG and CGA motifs, as well as for the CGC and CGT motifs (Supplementary Figure S6B).

DISCUSSION

DNMT3 sequence selectivity at +2 position relies on the likelihood of a stable Arg/Lys-Guanine base-specific hydrogen bond formation

In mammals, de novo DNA methylation is an integral epigenetic modification, vital for regulating essential cellular processes. The primary cell-specific methylation patterns are imposed by the two paralogous de novo methyltransferase enzymes, DNMT3A and DNMT3B. Despite their high sequence and structural similarity at their enzymatic modules, these enzymes preferentially methylate close but distinct sequences. This distinction poses an intriguing puzzle, which requires further elucidation. To solve this puzzle, cell-scale methylation and structural studies were performed. These studies pointed out one distinct (Arg836 of DNMT3A and Lys777 in DNMT3B) and one conserved (Asn838 of DNMT3A and Asn779 of DNMT3B) amino acid position as major players in sequence selectivity. In this work, we explored the dynamic mechanisms at the atomistic scale behind the selective roles of these amino acids. In this context, we concentrated on the +2 (CGX) position selectivity of these enzymes and so constructed complexes of DNMT3A-DNA and DNMT3B-DNA complexes, which differ on the DNA side at the +2 position. We then run four 500 ns simulations on each complex, covering a wide range of conformational space. Then, we devised an Interaction Dynamics Framework (IDF), focusing on the differential analysis of base-specific hydrogen bonding patterns across DNMT-DNA complexes, to be able to encode pairwise partner-specific signals.

Initially concentrating on the cognate complexes, we observed that in the DNMT3A-CGC complex, DNMT3A-Arg836's long and branched side chain consistently and strongly bound to the complementary strand G+2', while DNMT3B-Lys777 forms less intense but considerable hydrogen bonding pattern with the target G+2. Moreover, DNMT3A-CGC sequence interaction is extremely specific and concentrated solely on the +2' position, whereas in DNMT3B-CGG, as well as in the other complexes, such a specific and concentrated interaction pattern was never observed. Accordingly, we found that for its cognate sequence, DNMT3B-Lys777's recognition is further assisted with conserved Asn779's C+2' reading. Surprisingly, we observe the same cooperation with Arg836 for DNMT3A-CGG complex. Another striking observation was that DNMT3A's interaction pattern

changes significantly when we impose a mutation on its DNA sequence, whereas DNMT3B was less influenced by such nucleotide changes.

These findings can all be explained by the simple, yet elegant early observations made by Luscombe et al.'s 2001 study (Luscombe et al., 2001). In this work, 129 protein-DNA structures were analyzed and found that among the base-specific hydrogen bonds made, there is a strong preference of arginine and lysine residues for a major groove guanine. DNMT3B-Lys777's interactions with G+1 positions in CGG, CGT, and CGC sequences, along with its interactions with G+2' in the CGC sequence follow this description. The same work also underscored the significance of bidentate interactions in enhancing specificity in DNA sequence recognition, especially by the Arg–G interaction happening through the major groove. In such an interaction, arginine engages with both O6 and N7 atoms, thereby stably and selectively recognizing guanine. Our analyses, conducted at an atomic resolution for the DNMT3A-CGC complex, reproduce this interaction pattern. Furthermore, Luscombe et al. termed an "end-on approach", where arginine interacts with one or both of its distal nitrogen atoms (dominant mode), and a "side-on approach" where it uses NE or NH1 atoms (recessive mode), a classification that aligns with the two interactions modes we noted in the DNMT3A-CGC complex. Also, in the case of DNMT3B, Lys777 places its NZ atom between the guanine acceptor atoms to perform an interaction that is similar to the end-on conformation. Additionally, the observed Arg-G interactions in the CGT and CGG complexes resonate with the reported preferences. So, the distinctive attributes of arginine, such as its long side chain, capacity to engage in different conformations, and ability to generate robust hydrogen bond geometries, play a pivotal role in its selective and specific interactions with CGC. In contrast, lysine, while also possessing a long side chain, can only interact in a singular configuration, which largely limits its interaction versatility. So, lysine's ability to employ only a single point for hydrogen donation makes it less likely to mimic ideal geometries compared to arginine. This fundamental distinction elucidates why, in our analyses of DNMT3A and DNMT3B, Arg836 singularly directs the interaction network in DNMT3A, while Lys777 has to coordinate with Asn to compensate for the loss in branching in DNMT3B.

How atomistic scale dynamics-based observations can be interpreted within the perspective of earlier omics findings on DNMT3 sequence selectivity

To contextualize our observations, we compared our MD-based findings with the available studies reporting methylation profiles at the general (macro) scale. The first class of such studies relies on deep enzymology, such as the ones carried out by Gao et al. in 2020, focusing on the methylation activity of DNMT3A/B's MTase domain using DNA substrates of limited length (e.g., 10-15 nucleotides). These experiments revealed notable variations in methylation rates at CpG sites, with DNMT3A preferring a cytosine after CG dinucleotide and DNMT3B favoring guanine or adenine (Jeltsch et al., 2021). The second type of such macro-scale studies was carried out by Mallona et al. using whole-genome bisulfite datasets, which demonstrated genome-wide sequence selectivity at both CpG and non-CpG sites, emphasizing purine preference for DNMT3A and pyrimidine preference for DNMT3B.

When comparing our MD simulations and interaction network analyses with these experimental results, we see that, like in our observation, higher enzyme activity was noted in the CGA sequence than in CGT by Gao et al.. Furthermore, studies by Gao (2020) and Mallona (2021) demonstrated that DNMT3A exhibits higher CpG specificity compared to DNMT3B. In our study, analyzing the base-specific hydrogen bond networks of DNMT3A and DNMT3B, we observed that DNMT3A displayed higher scores in specific hydrogen bonding intensity, consistent with experimental data. Moreover, we noted that single nucleotide substitutions in other CG motifs significantly altered the DNMT3A-DNA interaction profile, whereas DNMT3B appeared less affected, due to DNMT3A's higher specificity to CpG motifs.

In addition to different types of experimental studies, in 2018, Norvil et al performed a biochemical study on enzyme kinetics of DNMT3A and DNMT3B. The cooperativity of DNMTs is characterized by the collective binding of multiple DNMTs to DNA, enabling the methylation of multiple sites on the same DNA molecule. Conversely, processivity is denoted by their capacity to move along the DNA and methylate multiple sites before dissociating (Norvil et al., 2018). This enzymatic study illuminated the cooperative nature of DNMT3A and the non-cooperative behavior of DNMT3B. Lesser pairwise ionic bond interaction in DNMT3B can be attributed to its cooperative nature and a more freely sliding pattern on DNA. As a result, in DNMT3A motifs, ionic bond distributions are higher than DNMT3B

motifs. Concentrating primarily on the cognate sequence interactions, namely DNMT3A-CGC and DNMT3B-CGC, we observed that the most frequent DNMT3A-DNA interactions are 3-fold higher compared to the most frequent (Figure 3 and 4). In addition to probability distributions, we evaluate the intermolecular ionic bond profiles for both enzymes. As an outcome, we observed more interacting amino acid-nucleotide pairs in DNMT3A (Supplementary Figure S7A). So, DNMT3A anchors to the DNA at more positions than DNMT3B (Supplementary Figure S7B). We observed that this was due to amino acid differences on the TRD loop. Especially, Arg831, for which at the same position in DNMT3B Gln772 is present, which causes more ionic bonds in DNMT3A, explaining the enzymatic observation at the atomistic scale.

Future Perspective

Our study sheds light on the sequence selectivity of DNMT3A/B, emphasizing the need for a more comprehensive exploration of the impact of the broad flanking region and the sequence range from -3 to +3 on flanking nucleotide selectivity, as highlighted in recent studies (Dukatz et al., 2022; Gao et al., 2020; Jeltsch et al., 2021). The intricate allosteric effect introduced by DNMT3L further complicates the influencing factors. Although we focused on the minimum enzymatic module in our MD simulations, specifically the MTase domain of DNMT3A/B, a deeper understanding could be gained by investigating the effects of changes in other flanking nucleotides and exploring non-CpG DNA sequences. Understanding how the interaction network evolves for selectivity in non-CpG DNA sequences will be pivotal for a comprehensive understanding of these systems. In this context, our study pioneers the elucidation of interaction networks in all flanking regions of the DNA methylation mechanism, contributing to the development of therapeutics targeting these mechanisms and the design of enzymes with analogous functions.

In conclusion, we present the recognition mechanisms of DNMT3A and DNMT3B through computational approaches. Ultimately, our study not only play a crucial role in advancing the development of therapeutics for the treatment of diseases that may arise due to disruptions in the de novo DNA methylation pattern, but also provides a computational framework to understand protein-DNA recognition mechanisms.

METHODS

Pairwise sequence alignment

Pairwise sequence alignment was performed on the EMBOSS Needle web-server(Madeira et al., 2019). Default settings and BLOSUM62 matrix were used for alignment. Then, the sequence differences were determined and alignment results were visualized with using Espript 3.0 web-server (Robert & Gouet, 2014).

System preparation for molecular dynamics

The molecular dynamics simulations were conducted utilizing crystal structures obtained from the Protein Data Bank (PDB). The structures included the DNMT3A-DNMT3L complex with single CpG-containing DNA (PDB ID: 6F57) and the human DNMT3B-DNMT3L complex with DNA containing a CpGpT site (PDB ID: 6KDB) (Lin et al., 2020a; Z. M. Zhang et al., 2018). From these initial structures, individual monomers of DNMT3A and DNMT3B, 10 bp long CGC-containing DNA (5'-CATG*CGCTCT-3') and S-Adenosyl-L-Homocysteine (SAH) were extracted. To create the pre-reaction forms of these complexes, SAH molecule was replaced by the S-Adenosylmethionine (SAM) molecule in the crystal structure of mouse DNA methyltransferase 1 with AdoMet (PDB ID: 3AV6)(Takeshita et al., 2011). Additionally, zebularine in the crystal structure of DNA was converted to cytosine. To ensure equal length, two and three amino acids were removed from the N-terminal ends of DNMT3A and DNMT3B structures, respectively. The cytosine at +2 in DNMT3A-CGC modified to the other three nucleotide possibilities, i.e., thymine, adenine, and guanine using 3DNA web server was used. (Li et al., 2019). The modeled DNA structures were further isolated to generate the corresponding DNMT3B complexes. All created 8 complexes were refined with the advanced refinement protocol using the HADDOCK2.2 web-server Guru Interface, and 200 structures were generated for each system (van Zundert et al., 2016). The structures with the best scores were selected for molecular dynamics simulations. For free simulations, DNA molecules were removed from the refined structures.

Molecular dynamics simulations and quality control analysis

GROMACS 2020.4 software was used for molecular dynamics simulations (van der Spoel et al., 2005). AMBER14SB-parmbsc1 force-field was used for both protein-DNA and free systems and TIP3P water was chosen as the water model (Jorgensen et al., 1983). The SAM molecule is not defined in the force-field, so the topology file (.itp), GROMACS formatted structure file (.gro), and positional restrain file (posre.itp) of the SAM molecule were created using the Acpype web-server (Sousa Da Silva & Vranken, 2012). In addition, the 5' and 3' ends of the DNA are defined in the Gromacs formatted structure file of the complex.

The simulation box was chosen as a dodecahedron and the protein-DNA complex was located 1.4 nm from the edges of the box. Afterward, positional restraints were increased from 1000 to 2500 and vacuum minimization was applied. After vacuum minimization, the simulation box was filled with water and energy minimization was performed. The ion concentration of the system was equilibrated to 0.15 M with K+ and CL- ions and the systems were geometrically relaxed with energy minimization again. Before starting the MD run, the system was heated to 310 K under constant volume to equilibrate the solvent and ions around the complex. System was chosen as the energy group. NVT equilibration is completed in 6 steps lasting 2 ps each, with each step after the first step positional restraints were decreased from 2500 to 500, 500 to 400, 400 to 300, 300 to 200, and 200 to 100. After the temperature of the system was stabilized, the pressure was stabilized under a constant particle number. NPT equilibration was performed in 3 steps of 20 ps and positional restraints were reduced from 100 to 50 and 50 to 10 in the first two steps. In the last step, the positional restraints were completely removed and the system was made ready for production run. For each DNA-bound system, 500 ns length 4 repetitive simulations were performed with different initial velocities (total of 2 µs for each system). Free systems were simulated for 500 ns for both systems. After the periodic boundary conditions of each simulation were corrected, an ensemble was created by taking 1 frame every 0.5 ns for each simulation.

After completing the molecular dynamics simulations, thorough quality control analyses were conducted using GROMACS 2020.4 software. These analyses included Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), β-factor, and cross-

correlation analysis. To enhance the interpretability of the data, the covariance matrices were transformed into cross-correlation matrices. This conversion was achieved using the Numpy and Pandas Python libraries. Detailed implementation steps are available on our GitHub repository (https://github.com/aysebercin/gmx_corr).

Before the analysis part, the 100 ns of all simulations was accepted as equilibration time and discarded.

Calculation and analysis of the intermolecular contacts

Intra and intermolecular contacts were calculated using the Interfacea Python package (Rodrigues et al., 2019). This package facilitated the calculation of non-covalent interactions, including hydrogen bonds, salt bridges, and hydrophobic interactions. Interfacea outputs for each frame are processed using bash and the Pandas library of Python. To refine the dataset, interactions between the protein and methylated cytosine were systematically removed from the interaction dataframe. For base-specific hydrogen bonds, interactions of backbone atoms of amino acids and nucleotides were also filtered out from the interaction dataframe. Statistical visualization of the interaction data was performed using the Seaborn Python library (Waskom, 2021). For statistical visualization of the interaction data, Seaborn Python library were used (Waskom, 2021). Additionally, for residue-level and atom-level analysis, interactions that exceeded 10% of the simulation time were considered significant. Subsequently, for residue level analysis, the significant hydrogen bonding interactions were normalized by dividing the total hydrogen bonding count by the total frame number.

Interaction Dynamics Framework

To elucidate the sequence selectivity of DNMT3 enzymes, we developed an Interaction Dynamics Framework (IDF) focusing on base-specific protein-DNA hydrogen bonding profiles, key to understanding direct readout interactions. The IDF approach comprises three progressively detailed steps, each explained below:

(i) Global interaction patterns: For each MD simulation involving DNMT3A/B-CGC/G/T/A complexes, intermolecular interactions were systematically computed. Interaction dataframes from replica simulations were aggregated for each complex within all steps of IDF. Interactions involving SAM-DNA and DNMT3-methylated cytosine were

removed from the dataframes. Additionally, base-specific interactions were identified by excluding backbone atoms of amino acids and sugar-phosphate backbone atoms of nucleotides. The remaining hydrogen bonds were quantified for each complex, and probability distributions of these interactions throughout the simulations were visualized using violin plots.

(ii) Residue level flanking nucleotide – amino acid interactions: To identify key interactions in the flanking region, interactions within positions +1 to +3 of the DNA were isolated. The number of base-specific hydrogen bonds for each interacting pair was calculated. Interactions occurring in less than 10% of the simulation time (4x400 ns, 3204 frames total, and less than 320 hydrogen bonds) excluded. Interaction densities for each complex were calculated by dividing the total number of hydrogen bonds by the total number of frames (3204) and visualized using heatmaps.

(iii) Atom level flanking nucleotide – amino acid interactions: To delve into the base-specific hydrogen bond formation identified in Step (ii), the interacted atoms of each pair were identified. The total number of hydrogen bonds formed in each atom pair was calculated, and cumulative numbers were visualized using barplots. This detailed atom-level analysis provided insights into the specificity and dynamics of base-specific hydrogen bond interactions within the flanking nucleotide – amino acid interface.

DNA geometry analysis

To derive DNA geometrical features such as major groove width and base-pair step parameters, the x3DNA suite was employed. The resulting x3DNA outputs were meticulously processed using a Python script, previously made available on our GitHub repository (https://github.com/CSB-KaracaLab/NucDNADynamics).

For each simulation set, we calculated essential metrics, including the minimum, maximum, and mean values for major groove width and base-pair step parameters. This comprehensive analysis provides a detailed understanding of the dynamic structural changes in DNA throughout the course of the simulations.

ACKNOWLEDGEMENTS

The preliminary data of this work is produced by Deniz Dogan on SURFsara, the Dutch HPC center, in the lab of Prof. Siewert Jan Marrink under the umbrella of HPC-Europa3 visitors program. We also thank to IBG's HPC resources for our analyses and simulations.

FUNDING

E.K. acknowledges EMBO Installation Grant (Project Number: 4421), Alexander von Humboldt Foundation Return Fellowship, and Young Investigator Award granted by the Turkish Science Academy. A.B.B. acknowledges TÜBİTAK 2232/A International Fellowship for Outstanding Researchers Program (Project Number: 118C354) and TÜBİTAK 2211/A National PhD Scholarship Program.

References

Das, P. M., & Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *22*(22), 4632–4642. https://doi.org/10.1200/JCO.2004.07.151

Dong, Y., Zhao, H., Li, H., Li, X., & Yang, S. (2014). DNA methylation as an early diagnostic marker of cancer (Review). *Biomedical Reports*, *2*(3), 326–330. https://doi.org/10.3892/br.2014.237

Dor, Y., & Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *The Lancet*, *392*(10149), 777–786. https://doi.org/10.1016/S0140-6736(18)31268-6

Emperle, M., Adam, S., Kunert, S., Dukatz, M., Baude, A., Plass, C., Rathert, P., Bashtrykov, P., & Jeltsch, A. (2019). Mutations of R882 change flanking sequence preferences of the DNA methyltransferase DNMT3A and cellular methylation patterns. *Nucleic Acids Research*, *47*(21), 11355–11367. https://doi.org/10.1093/nar/gkz911

Gao, L., Emperle, M., Guo, Y., Grimm, S. A., Ren, W., Adam, S., Uryu, H., Zhang, Z.-M., Chen, D., Yin, J., Dukatz, M., Anteneh, H., Jurkowska, R. Z., Lu, J., Wang, Y., Bashtrykov, P., Wade, P. A., Wang, G. G., Jeltsch, A., & Song, J. (2020). Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nature Communications*, *11*(1), 3355. https://doi.org/10.1038/s41467-020-17109-4

Gowher, H., & Jeltsch, A. (2018). Mammalian DNA methyltransferases: New discoveries and open questions. *Biochemical Society Transactions*, *46*(5), 1191–1202. https://doi.org/10.1042/BST20170574

He, Y., & Ecker, J. R. (2015). Non-CG Methylation in the Human Genome. *Annual Review of Genomics and Human Genetics*, *16*, 55–77. https://doi.org/10.1146/annurev-genom-090413-025437

Jeltsch, A., Adam, S., Dukatz, M., Emperle, M., & Bashtrykov, P. (2021). Deep Enzymology Studies on DNA Methyltransferases Reveal Novel Connections between Flanking Sequences and Enzyme Activity. *Journal of Molecular Biology*, *433*(19), 167186. https://doi.org/10.1016/j.jmb.2021.167186

Jin, Z., & Liu, Y. (2018). DNA methylation in human diseases. *Genes & Diseases*, *5*(1), 1–8. https://doi.org/10.1016/j.gendis.2018.01.002

Klutstein, M., Nejman, D., Greenfield, R., & Cedar, H. (2016). DNA Methylation in Cancer and Aging. *Cancer Research*, *76*(12), 3446–3450. https://doi.org/10.1158/0008-5472.CAN-15-3278

Li, K. K., Li, F., Li, Q. S., Yang, K., & Jin, B. (2013). DNA methylation as a target of epigenetic therapeutics in cancer. *Anti-Cancer Agents in Medicinal Chemistry*, *13*(2), 242–247. https://doi.org/10.2174/1871520611313020009

Lin, C.-C., Chen, Y.-P., Yang, W.-Z., Shen, J. C. K., & Yuan, H. S. (2020). Structural insights into CpG-specific DNA methylation by human DNA methyltransferase 3B. *Nucleic Acids Research*, *48*(7), 3949–3961. https://doi.org/10.1093/nar/gkaa111

Mallona, I., Ilie, I. M., Karemaker, I. D., Butz, S., Manzo, M., Caflisch, A., & Baubec, T. (2021). Flanking sequence preference modulates *de novo* DNA methylation in the mouse genome. *Nucleic Acids Research*, *49*(1), 145–157. https://doi.org/10.1093/nar/gkaa1168

Mao, S.-Q., Cuesta, S. M., Tannahill, D., & Balasubramanian, S. (2020). Genome-wide DNA Methylation Signatures Are Determined by DNMT3A/B Sequence Preferences. *Biochemistry*, *59*(27), 2541–2550. https://doi.org/10.1021/acs.biochem.0c00339

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257. https://doi.org/10.1016/s0092-8674(00)81656-6

Schmitz, R. J., Lewis, Z. A., & Goll, M. G. (2019). DNA Methylation: Shared and Divergent Features across Eukaryotes. *Trends in Genetics*, *35*(11), 818–827. https://doi.org/10.1016/j.tig.2019.07.007

Schneider, K., Fuchs, C., Dobay, A., Rottach, A., Qin, W., Wolf, P., Álvarez-Castro, J. M., Nalaskowski, M. M., Kremmer, E., Schmid, V., Leonhardt, H., & Schermelleh, L. (2013). Dissection of cell cycle-dependent dynamics of Dnmt1 by FRAP and diffusion-coupled modeling. *Nucleic Acids Research*, *41*(9), 4860–4876. https://doi.org/10.1093/nar/gkt191

Skvortsova, K., Stirzaker, C., & Taberlay, P. (2019). The DNA methylation landscape in cancer. *Essays in Biochemistry*, *63*(6), 797–811. https://doi.org/10.1042/EBC20190037

Vukic, M., & Daxinger, L. (2019). DNA methylation in disease: Immunodeficiency, Centromeric instability, Facial anomalies syndrome. *Essays in Biochemistry*, *63*(6), 773–783. https://doi.org/10.1042/EBC20190035

Zeng, Y., & Chen, T. (2019). DNA Methylation Reprogramming during Mammalian Development. *Genes*, *10*(4), 257. https://doi.org/10.3390/genes10040257

Zhang, W., & Xu, J. (2017). DNA methyltransferases and their roles in tumorigenesis. *Biomarker Research*, *5*(1), 1. https://doi.org/10.1186/s40364-017-0081-z

Zhang, Z.-M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., Liu, S., Ji, D., Rothbart, S. B., Wang, Y., Wang, G. G., & Song, J. (2018). Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature*, *554*(7692), 387–391. https://doi.org/10.1038/nature25477