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

AUTHORS

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

[1] Basic and Translational Research Program, Izmir Biomedicine and Genome Center, Balçova, Izmir, 35340, Turkey
[2] Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, Balçova, Izmir, 35340, Turkey

* To whom correspondence should be addressed. Tel: +905334566096; Fax: +9002322776353; Email: ezgi.karaca@ibg.edu.tr

ABSTRACT

De novo DNA methylation is crucial for mammalian gene regulation, and it is mediated by DNMT3A and DNMT3B enzymes. Despite their catalytic domains sharing over 90% sequence similarity, they preferentially methylate different flanking nucleotides adjacent to a classical CG motif. To uncover the basis of this selective methylation pattern, we conducted extensive molecular dynamics simulations of DNMT3A/B bound to eight possible CGX[C/G/T/A] motifs. Our comparative analysis of dynamic enzyme-DNA interactions revealed that single nucleotide substitutions in CGX motifs significantly alter DNMT3A-DNA hydrogen bonding profiles, resulting in a more specific readout for DNMT3A compared to DNMT3B. We found that +2 CGX nucleotide selectivity is driven by Arg/Lys-Guanine base-specific hydrogen bonds formed between DNMT3A/B and DNA. More specifically, Arg836 in DNMT3A reads the guanine complementary to C+2 in its cognate CGC motif, while Lys777 in DNMT3B recognizes two consecutive guanines in its cognate CGG. In DNMT3B, the impact of Arg-to-Lys mutation is mitigated through the cooperative nucleotide recognition directed by Asn779. All in all, our study provides fundamental insights into the molecular mechanisms underlying DNA sequence selectivity in *de novo* DNA methylation at the atomistic scale, which will have direct implications in understanding other DNA sequence recognition mechanisms.

Number of words: 196

INTRODUCTION

DNA methylation is an epigenetic mark introduced to regulate a diverse range of vital cellular processes, including aging, embryonic development, gene expression, reprogramming, chromatin modification, and X chromosome inactivation (1–4). In humans, there are three catalytically active DNMT enzymes: DNMT1, DNMT3A, and DNMT3B. These enzymes predominantly methylate CpG dinucleotides (5, 6). DNMT3A and DNMT3B perform *de novo* DNA methylation, imposing the methyl marks on genomic DNA from scratch (7), while DNMT1 maintains existing methylation patterns during DNA replication (8).

Among these DNMT enzymes, DNMT3A/B-dependent methylation is orchestrated by highly similar paralogous methyltransferase (MTase) domains of DNMT3A/B, exhibiting a 91.2% sequence similarity over 283 residues (Figure 1) (9, 10). The DNMT3-MTase domains contain catalytic and target recognition (TRD) loops, playing pivotal roles in locating and anchoring the methylation site through interacting with genomic DNA from both minor and major grooves (Figure 2A). The methylation reaction involves DNMT3-MTase flipping the target cytosine, extracting a methyl group from the S-adenosyl-L-methionine (SAM) molecule and transferring it to the 5' carbon of the target cytosine (6, 10–13) (Figure 2A). Despite their high sequence similarity and the conserved reaction patterns, in-depth enzymology and gene knockout experiments revealed 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), whereas 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 (10, 13–16). Hereafter, 3A-MTase will be referred to as DNMT3A and 3B-MTase as DNMT3B for simplicity.

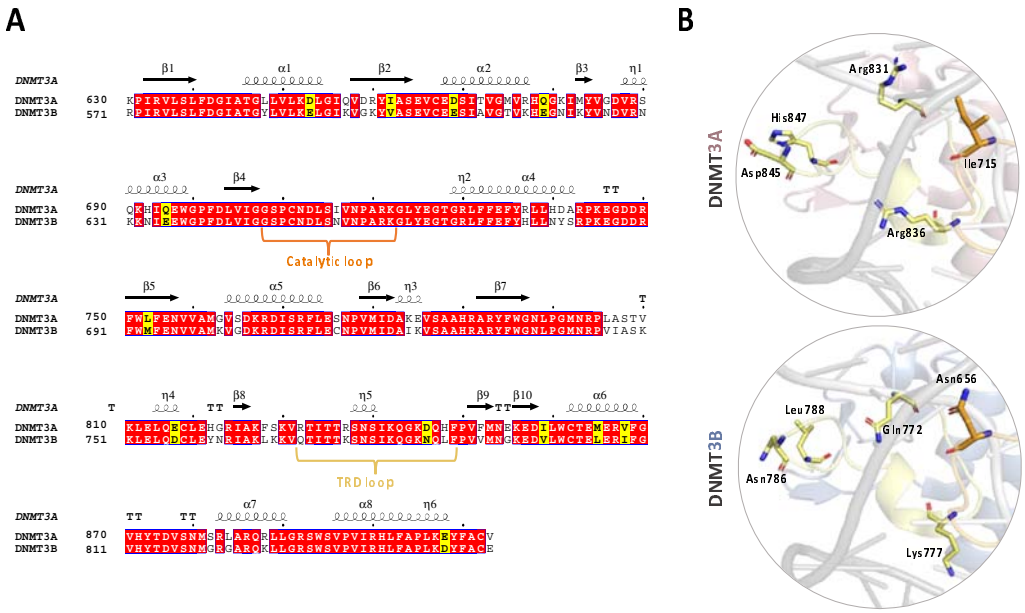

Figure 1. Differential amino acid positions in DNMT3A and DNMT3B MTase domains. (A) A pairwise sequence alignment (with EBLOSUM62 matrix) was performed for the MTase domains of DNMT3A and DNMT3B over their 283 amino acids. The conserved residues are highlighted in red, while the non-conserved residues with similar physicochemical properties are marked in yellow. Black residues with a white background indicate non-conserved amino acids. The sequence identity and similarity values are 80.2% and 91.2%, respectively. **(B)** The non-conserved amino acids of DNMT3A (upper) and DNMT3B (down) are located on the TRD-loop (Arg831-to-Gln772, Arg836-to-Lys777, Asp845-to-Asn786, His847-to-Leu788) and on the catalytic loop (Ile715-to-Asn656). Yellow residues represent the TRD loop, while orange residues represent the catalytic loop.

Over the past decade, several structural studies have focused on resolving the molecular grounds behind the DNMT3A/B sequence selectivity. These studies identified a TRD loop substitution, Arg836 in DNMT3A to Lys777 in DNMT3B, as the primary determinant for +2 flanking sequence preference (10, 13, 14, 17) (Figure 1). Specifically, available structures showed that when DNMT3A is bound to its cognate CGC, Arg836 acts as a specificity-determining position (SDP) through forming a base-specific hydrogen bond with O6 atom of guanine (G+1) (Figure 2A-B) (13). In the case of cognate DNMT3B-CGG complex, the SDP Arg836 is substituted by a lysine, which engages in van der Waals interactions with N7 atom of guanine G+2′ (Figure 2B) (10). It was proposed that the Arg-to-Lys substitution in

DNMT3B is compensated by the involvement of a conserved amino acid, Asn779, which forms a hydrogen bond with O6 atom of G+1 (14).

Apart from their cognate complexes, in the resolved DNMT3A–CGT complex, SDP Arg836 forms a hydrogen bond with the O6 atom of the CpG guanine (G+1). G+1 also forms a water-mediated hydrogen bond with the NE atom of Arg836 via its N7 atom (13). 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 (18). On the other hand, in the case of DNMT3B-CGT complex, Lys777 points towards T+2, while Asn779 donates a hydrogen to the O6 atom of G+1 (14). In the DNMT3B–CGA complex, similarly to DNMT3A, Lys777 interacts with A+2, through van der Waals contacts (14).

In 2021, Mallona et al. further investigated the sequence selectivity of cognate DNMT3A-CGC and noncognate DNMT3A-CGG complexes using molecular dynamics (MD) simulations. This sequence selectivity was first validated by whole genome bisulfite sequencing experiments, followed by analysis of dynamic interaction profiles of these DNMT3A systems. This analysis 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 (16), explaining the +2 selectivity of this enzyme. The work of Mallona *et al.* and our previous work (19, 20) demonstrated that analyzing crystal structures was pivotal to unraveling the critical amino acids, i.e., SDPs responsible for selecting the cognate sequences (21, 22). However, incorporating a dynamics perspective was necessary to elucidate the selective amino acid-nucleotide interactions. Building on this insight, we conducted the most comprehensive MD study to date on *de novo* DNA methyltransferases: we ran MD simulations of DNMT3A and DNMT3B bound to all possible nucleotide substitutions at the +2 position (i.e., CGC/G/T/A). We then devised and employed an Interaction Dynamics Framework (IDF) for systematically exploring of base-specific hydrogen bonding interactions across all complexes. As an outcome, we observed that DNMT3 sequence selectivity +2 position relies on the likelihood of a stable Arg/Lys-Guanine base-specific hydrogen bond formation. This observation aligns with prior research and extends the findings of MD studies to a broader range of structures, suggesting a generalizable trend. Furthermore, our study provides a workflow to shed light

on the mechanistic basis of the sequence selectivity of DNMT3 enzymes. We also showed that this direct readout mechanism is dictated by the DNA groove geometry. This straightforward yet compelling perspective offers a simple explanation for a complex phenomenon, which can be directly applied to understand the partner selectivity of other paralogous DNA-binding proteins.

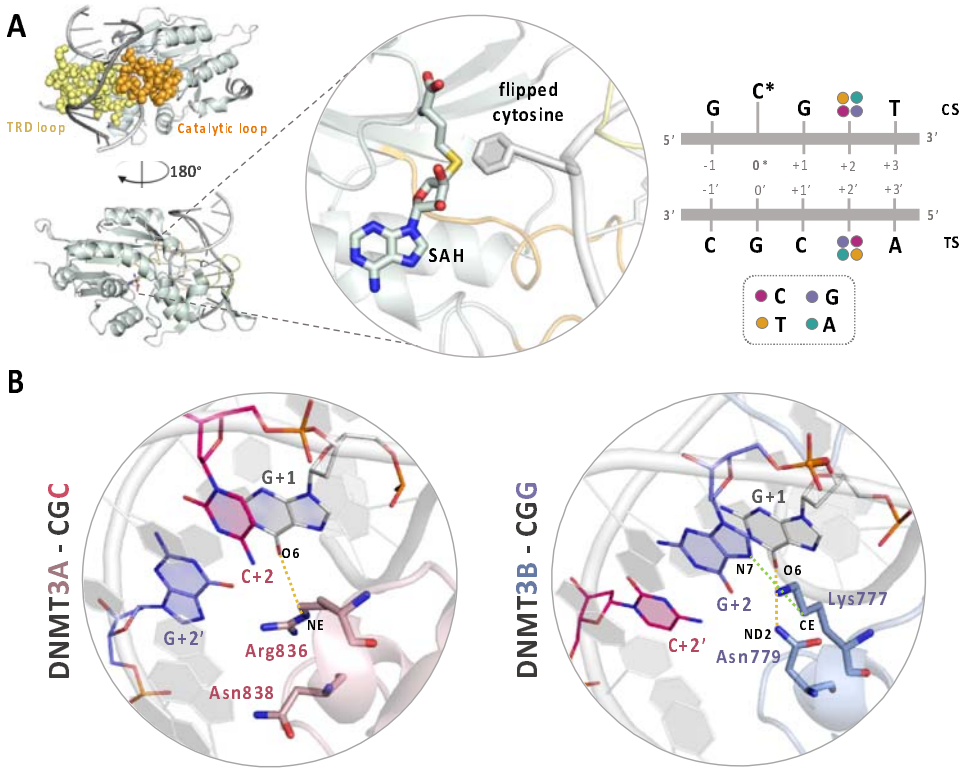

Figure 2. Structural features of DNMT3A and DNMT3B MTase domains. (A) Left: 3D structure of DNMT3A/B-MTase domain from the front (top) and back (bottom) sides. Orange region represents the catalytic loop and yellow region represents the target recognition domain (TRD) loop. The inset depicts the SAM methyl donor and the flipped target cytosine (the cytosine to be methylated). 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 (13)). In this crystal structure, Arg836 forms a base specific hydrogen bond with G+1 of CGC (yellow) (13). Right:

DNMT3B is bound to CGG (PDB ID: 6KDA (10)). In this crystal structure, Lys777 forms van der Waals interactions with G+2 (green) (10) and Asn779 forms a hydrogen bond with G+1 of CGG (yellow) (14).

MATERIAL AND METHODS

Pairwise sequence alignment

Pairwise sequence alignment of MTase domains of DNMT3A and DNMT3B was performed on the EMBOSS Needle web-server (23). Default settings and the EBLOSUM62 matrix were used for alignment. Alignment results were visualized with using Espript 3.0 web-server (24).

Molecular dynamics simulations

To assess the sequence selectivity of DNMT3A and DNMT3B enzymes, all-atom molecular dynamics simulations in explicit solvent were performed to capture atomistic interactions between protein and DNA. Atomic coordinates of DNMT3A-DNMT3L complex with single CpG-containing DNA (PDB ID: 6F57 (13)) and the human DNMT3B-DNMT3L complex with DNA containing a CpGpG site (PDB ID: 6KDA (10)) were used as starting structures for MD simulations. 10 bp long CGC-containing DNA (5'-CATG*CG**C**TCT-3') of DNMT3A-CGC complex were used as template DNA. The cytosine at +2 in DNMT3A-CGC modified to the other three nucleotide possibilities, i.e., thymine, adenine, and guanine using 3DNA web server (25). The modeled DNA structures were further isolated to generate the corresponding DNMT3B complexes. To create the pre-reaction forms of these complexes, the SAH molecule was replaced by the S-Adenosylmethionine (SAM) molecule in the crystal structure of mouse DNA methyltransferase 1 with AdoMet (PDB ID: 3AV6 (26)). Additionally, zebularine in the crystal structure of DNA was converted to cytosine. All created DNMT3A-CGC/T/A/G and DNMT3B-CGC/T/A/G complexes were refined with the advanced refinement protocol using the HADDOCK2.2 web-server Guru Interface, and 200 structures were generated for each system (27). The structures with the best scores were selected for molecular dynamics simulations. For DNA-unbound simulations, DNA molecules were removed from the refined structures.

MD simulations were performed using GROMACS 2020.4 (28) under the effect of AMBER14SB-parmbsc1 (29) force field. Parameters of SAM cofactor were generated by the

Acpype web-server (30). The simulation box was set up as a rhombic dodecahedron with a minimum distance between the solute and the edge of the box of 1.4 nm. Before running the MD simulations, the force constant of the positional restraints were increased from 1000 to 2500 and protein-DNA complexes were minimized by the steepest descent algorithm in the vacuum. Then, they were solvated with TIP3P water (31) and 0.15 M KCl and the system was relaxed with energy minimization again.

The systems were heated to 310 K using the velocity-rescaling thermostat (32). NVT equilibration was completed in 6 steps, each lasting 20 ps, with each step after the first step force constants of the positional restraints were decreased from 2500 to 500, 500 to 400, 400 to 300, 300 to 200, and 200 to 100. Then sequential 20 ps MD simulations with gradual release of position restraints on heavy atoms were performed under constant pressure at 1 bar using Berendsen barostat (33). Before starting the MD run, the positional restraints were completely removed. For each DNA-bound system, 500 ns length 4 independent simulations were performed with different initial velocities (total of 2 μs for each system). DNA-unbound systems were simulated for 500 ns. The PME approach (34) was employed to address long-range electrostatic interactions, utilizing a nonbonded cutoff of 1.2 nm. The LINCS method (35) was used to restrict hydrogen bond lengths. After accounting for periodic boundary conditions, an ensemble was created by taking 1 frame every 0.5 ns for each simulation. Before the analysis part, equilibration time was set as 100 ns and discarded from the simulations. As quality control measures we calculated RMSD and RMSF values of each complex with GROMACS 2020.4 software (28) (Supplementary Figures S1-3).

Calculation and analysis of the non-covalent interactions

Intra and intermolecular interactions were calculated using the Interfacea Python package (38). 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 (37) 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. For statistical visualization of the interaction data, Seaborn (39) Python library were used.

Additionally, for residue-level and atom-level analysis, interactions that exceeded 10% of the simulation time were considered significant.

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 independent trajectories of the same system were merged 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 the number 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 (40) was employed. The resulting x3DNA outputs were 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.

RESULTS

An Interaction Dynamics Framework to understand the selective protein-DNA direct readout signals

To understand the sequence selectivity of DNMT3 enzymes, we used the available DNMT3A/B-DNA structures as a basis to build eight DNMT3A/B-DNA complex models: DNMT3A-CGC/T/A/G and DNMT3B-CGG/T/A/C. In this model set, the DNA sequences differed only at the +2 position (Methods). This setup allowed us to concentrate on the interaction differences caused by the +2 flanking nucleotides and neighboring DNMT3A/B amino acids (Figure 2). We ran MD simulations of each DNMT3 complex (DNMT3A/B-CGX) for 2 μs long (4x0.5 μs independent trajectories), leading to a cumulative 16 μs-long DNMT3-CGX dynamics space (Methods). We then devised an Interaction Dynamics Framework (IDF). The IDF concentrates on the selective protein-DNA interactions upon analyzing the differential base-specific hydrogen bonding profiles. These profiles were earlier shown to correspond to the direct readout, i.e., selectivity signals (Figure 3A, Methods) (41–44).

Our IDF starts by calculating all base-specific hydrogen bonding profiles of DNMT3 complexes. For a given complex, IDF pools the interaction lists coming from four different trajectories to cover a wider sampling space. At this point, it ignores SAM-DNA and DNMT3-methylated-cytosine interactions since these remain unaltered across different complexes (Supplementary Figures S4 and S5). IDF also discards interactions observed for less than 10% of the simulation time, considering them to be unimportant to the downstream analysis. Over the filtered interaction sets, IDF initially compares the global base-specific hydrogen bond formation potentials across all complexes (Figure 3B). It then focuses on the +1 to +3 flanking region to calculate the interaction propensities at the residue/nucleotide level (Figure 3C). The propensities are reported as an interaction intensity value, which is the ratio of hydrogen bonds formed between nucleotides at positions +1 to +3 and DNMT3 amino acids to the frame count (see Methods). IDF finally analyzes the calculated hydrogen bonding patterns at the atomistic level (Figure 3D). When interpreting atom-atom interactions, it considers the hydrogen acceptance and donation capabilities of selected amino acids with the flanking bases, as dictated by the DNA groove geometry (Figure 3E) (41). Our IDF 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.

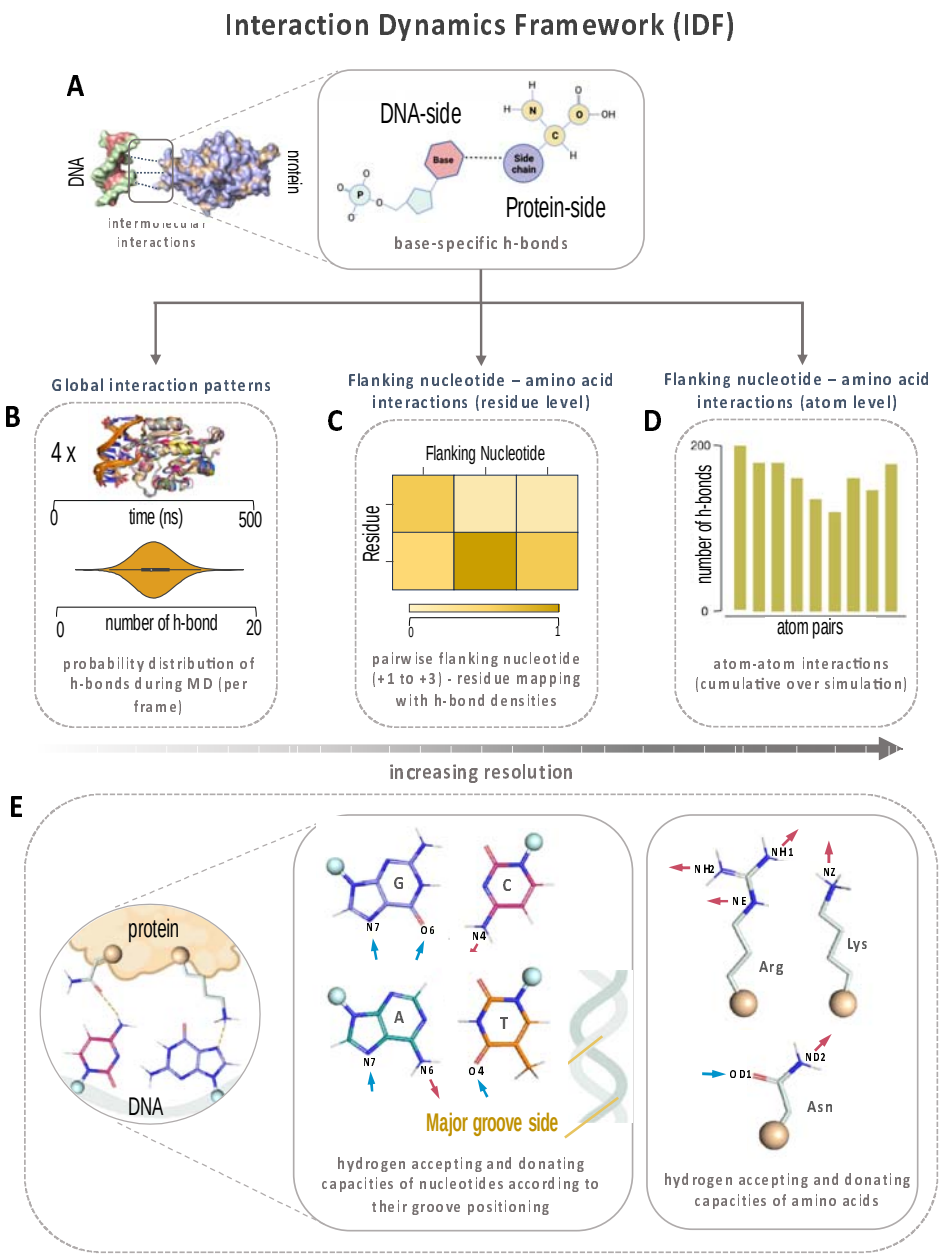

Figure 3. An interaction dynamics framework (IDF) to identify DNMT3 selectivity +2 flanking position. (A) Evaluation of intermolecular interactions between protein and DNA, in terms of base-specific hydrogen bonds. **(B)** Global interaction analysis of base-specific hydrogen bonds throughout 2 μs MD simulations for each DNMT3A/B system. The violin plots show the probability distributions of the number of hydrogen 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 target specificity is modulated by the hydrogen donation capacity of Arg836 and the groove positioning of Guanine

We calculated the base-specific hydrogen bonding probabilities of all DNMT3A (DNMT3A-CGC/T/A/G) complexes (Figure 3B). The median hydrogen bonding capacity of DNMT3A-CGX complexes varied between three-four hydrogen bonds (Supplementary Figure S6A). When we delved deeper into specific interactions formed around the +2 region (+1 to +3, Figure 3C). In the cognate DNMT3A-CGC complex, we observed a stable hydrogen bonding between Arg836 and complementary G+2′, with an interaction intensity of 1.9 (observed for over 70% of conformations, Figure 4A, Table 1). The atom-atom frequency distribution of the Arg836-and-G+2′ interaction showed that the NE and NH2 atoms of Arg836 form bidentate (dual hydrogen bonding with a base or base pair in a single conformation) stable hydrogen bonds with the N7 and O6 atoms of G+2′ (Figure 4A, Supplementary Movie 1), explaining the observed intensity value over 1.0. In the other 20.2% of conformations, we saw a single hydrogen bonding pattern established between NH1/2 groups of Arg836 and O6 of G+2′ (Supplementary Figure S4, Table 1).

Setting the cognate DNMT3A interaction mode as specific G+2′ recognition, we moved to the DNMT3A-CGT complex. The mutation of C+2 to T+2 leads to the disappearance of the complementary strand contacts, establishing moderate-intensity hydrogen bonding with the target strand's G+1 (Figure 4B). For this sequence, the intensity drop compared to the cognate sequence is almost three-fold (0.69 vs. 1.9). The atom-atom interactions of the CGT

motif showed that Arg836 forms dampened and distributed hydrogen bonds with its NH1, NH2, and NE atoms, as well as the N7 and O6 atoms of G+1. This reflects a less stable and less directed hydrogen bonding pattern compared to the cognate sequence. One such interaction mode is shown in Figure 4B, where Arg836 NH1 and NH2 atoms form bidentate hydrogen bonds with 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, as Arg836 cannot donate hydrogens to A+2′, leaving 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, the interaction of Arg836 with the T+2′ position reemerges, this time with a much lower intensity compared to the cognate sequence (0.84 vs. 1.9) (Figure 4C). This happens as T+2′ has the capacity to accept one hydrogen, which apparently is attractive enough for Arg836 to form a hydrogen bond. Examination of the pairwise atomic interactions reveals a single binding mode between T+2′ and Arg836, where O4 of T+2′ and NH1 of Arg836 establish a single and stable bond in almost 82% of the pooled conformations (Figure 4C). As a result, at the atomic scale, the +2 selectivity in the case of CGA sequence is maintained.

In all three previous cases, we only have Arg836 coming up as the SDP 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 the +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 G+1 and G+2 on the target strand, and Asn838 with C+2′ and T+3′ on the complementary strand. The interaction intensity values of Arg836 are the lowest ones observed compared to other DNMT3A complexes (0.28 and 0.49, Figure 4D). The atom-atom pairwise analysis reveals that Asn838 forms more frequent hydrogen bonds (approx. in 40% of conformers) compared to the bonds made by Arg836. At the atomic level, the highest frequency interaction is 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 G+2, though with a low probability (14.3% and 11.3% of conformations).

In addition to our hydrogen bonding analysis, we sought to investigate whether other factors are contributing to sequence selectivity. Consequently, we conducted a probability distribution analysis for ionic and hydrophobic interactions. Upon examining the probability distributions of other interaction types, we observed that hydrophobic and ionic interactions exhibit similar distributions in DNMT3A for both cognate CGC and non-cognate CGG motifs. As for ionic bond distribution, CGC and CGG displayed a wider range and typically formed an average of five ionic bonds per frame (Supplementary Figure S7A). To assist these findings, we finally calculated the mean and standard deviation values of DNA shape parameters for each DNMT3A complex (major groove width and base pair step parameters, Supplementary Figure S8A). The major groove width of the cognate CGC is smaller (varied between 12-25 Å) compared to other sequences. This analysis also highlighted that the most variation in base pair step parameters was observed for CGG and CGT sequences, aligning with the loss of +2 selectivity observation, as explained above (Supplementary Figure S8B).

Table 1. Summary of base-specific hydrogen bonds formed between DNMT3A with different DNA sequences. The hydrogen bond forming atoms and their observation frequencies are explicitly stated (over the whole 1.6 μs MD trajectory for a given DNMT3A complex).

DNMT3A	CGC	CGT	CGA	CGG
Arg836	**G+2'** NE-O6 77.4% NH2-N7 74.3% NH2-O6 20.2% NH1-O6 18%	**G+1** NH1-O6 17% NH2-O6 13.3% NE-O6 12.2% NE-N7 11.7% NH1-N7 11.4%	**T+2'** NH1-O4 81.84%	**G+2** NH2-O6 14.3% NE-N7 11.3% **G+1** NH1-N7 12.9%
Asn838	*	*	*	**C+2'** OD1-N4 38.9% **A+3'** ND2-N7 30.5% OD1-N6 17.2%

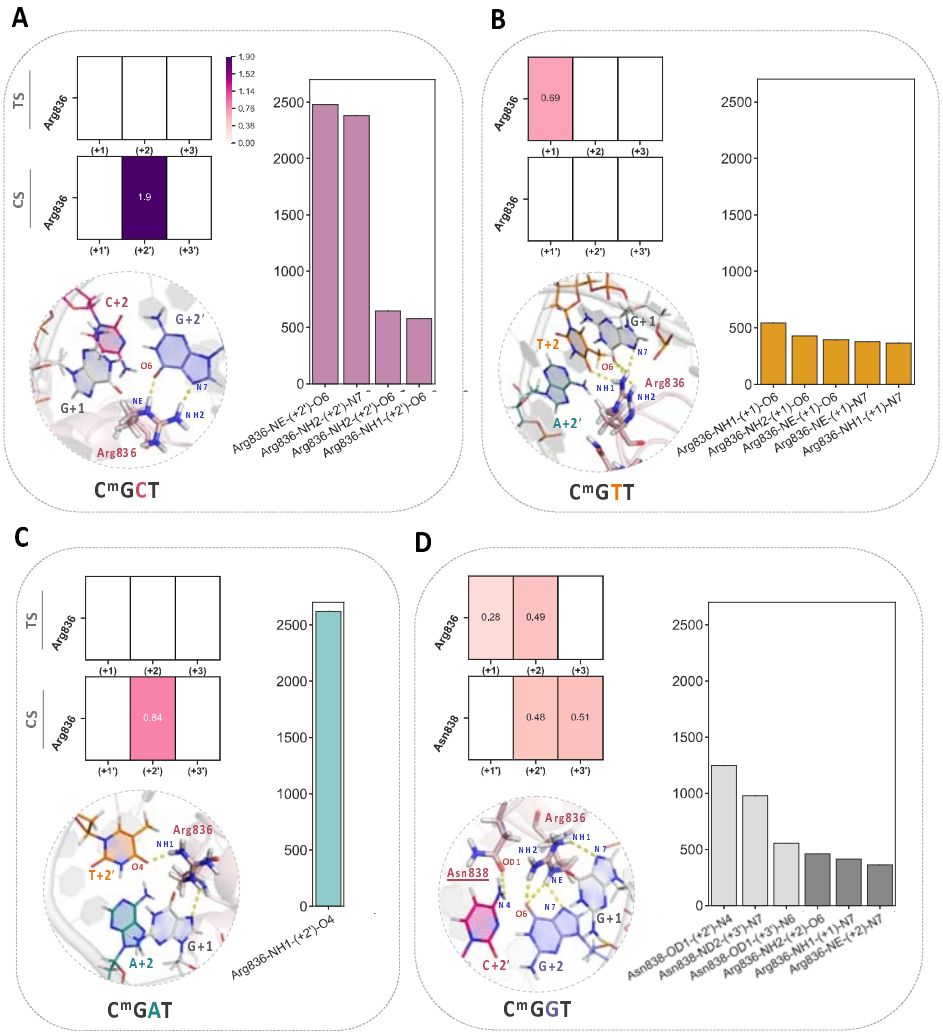

Figure 4. Interaction analysis of DNMT3A cognate (CGC), cognate-like (CGT) and noncognate systems (CGA/G). Flanking nucleotide interactions of DNMT3A complex with **(A)** CGC, **(B)** CGT, **(C)** CGA, **(D)** CGG. Each panel contains a heatmaps, showing the interaction intensities of the important amino acids with +1-to-+3 and its complementary strand (Methods). In the heatmaps, the first row represents the target strand (TS) interactions, while the second row shows the complementary strand (CS) ones. The panels also contain the distribution of base-specific hydrogen bonds analyzed at the atomistic scale (over 1.6 μs MD trajectory), covering the range of interactions depicted by the heatmap. Light colored bars represent the complementary strand interactions, whereas dark colored bars represent target strand interactions. The panels lastly present the structural snapshots of the important base-specific interactions highlighted in the heatmap and barplot.

The loss of hydrogen donation capacity due to the Arg-to-Lys substitution in DNMT3B is compensated by a conserved Asn

We then applied our IDF protocol to DNMT3B complexes. As an outcome, 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 densities (eight hydrogen bonds per frame, Supplementary Figure S6B). The CGT complex showed a higher probability and median value compared to other systems. Still, these probability distributions did not identify essential factors for DNMT3B sequence selectivity.

When we moved to the residue-specific analyses, we observed that, in the cognate CGG motif, Lys777 forms base-specific hydrogen bonds with +1, +2, and +3 positions on the target strand with moderate to low intensities (0.43 to 0.17). Notably, we did not observe any complementary strand interactions of Lys777 (Figure 5A). In this DNMT3B cognate complex, the +2′ complementary strand position is recognized by conserved Asn779 (corresponding to Asn838 of DNMT3A). Asn779 formed specific hydrogen bonds with +1 and +2′ positions with moderate intensity (0.38). The CGG interaction network spanned by Lys777 and Asn779 is strikingly very similar to DNMT3A-CGG, since complementary C+2′ cannot accept hydrogen through the major groove (Figures 3E and 5D). 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 30.1% of conformations, Figure 3A, Table 2). The first dominant interaction pattern is formed between Lys777 NZ and G+1's N7 (37.5% of conformations). G+1 is further stabilized by Asn779 hydrogen donor ND2 and G+1's free hydrogen acceptor O6. Importantly, Lys777 forms base-specific hydrogen bonds with G+1 and G+2 nucleotides through its hydrogen donor NZ (Figure 5A). Thus, in the DNMT3B cognate sequence, cooperative recognition between Lys777-G+2 and Asn779-C+2′ describes the specificity of this complex (Supplementary Movie 2).

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 5B). In this complex, G+1 recognition is handed to Asn779 (0.65 intensity) while Asn779 retains its role for +2′ recognition (same intensity as in the CGG complex). Asn779 acts as a hydrogen donor to +2′ position during

33.9% of conformers (Table 2). 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 65.4% of conformers. O4 of T+3 is reached by Lys777 in 34.2% of conformers, making it the only interaction of Lys777 with the target strand (Figure 5B). Interaction between Asn779 and T+2' was also highlighted by Gao et al. 2020 (14).

When we moved to the CGT motif, Asn779 forms specific hydrogen bonds with A+2′ with the highest intensity (0.66) compared to other complexes. Asn779 also forms 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 forming hydrogen bonds (Figure 3E, Figure 5C). Ser778 comes into play, forming hydrogen bonds through its OG atom, both accepting hydrogen from N6 atom of A+3′ and donating hydrogen to N7 atom of A+3′. Lys777 forms hydrogen bonds with G+1-O6 atom (Figure 5C). When we compared the independent simulations, we saw that Lys777 reads CGT in two different conformations: it reads +4' guanine and +3 thymine, or it reads +1 guanine (Supplementary Figure S5). The structural snapshots of the MD simulation showed that Asn779 forms bidentate hydrogen bonds with A+2′ by both donating and accepting hydrogen, so the interaction intensity increases (Figure 5C).

In the non-cognate case of DNMT3B-CGC complex, Lys777 interactions are distributed to +1 and +3 (0.19-0.25 intensities). It also forms hydrogen bonds with +2' position, possibly due to the attractive field posed by G+2′. This interaction has very low intensity (0.17). Asn779 continues forming hydrogen bonds with +2′, but with decreased intensity (2.4-fold compared to cognate motif and 4.1-fold compared to CGT). Interactions of Asn779 with the target strand are entirely lost and shifted to G+2′/A+3′ (Figure 5D). At the atomistic detail, we observed that Lys777 forms hydrogen bonds with 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 hydrogen bonds during 75% of the total simulation time (Supplementary Figure S5). Notably, one of the interesting interactions of Lys777 is a hydrogen bond with O4 atom of T+3. In this instance, Asn779 reads +2' guanine and +3' adenine. In addition, during the remaining 25% of the total simulation time, Lys777 interacts with +2' guanine and +1 guanine simultaneously (Supplementary Figure S5).

Following the hydrogen bond analysis at atomistic resolution, we calculated ionic bonds and hydrophobic interactions of the DNMT3B complexes. The interaction profile per frame for ionic and hydrophobic interactions exhibit a similar distribution profile in both cognate and non-cognate motifs (Supplementary Figure S7B). We examined major groove width and base-pair step parameters of DNMT3B-bound DNAs to correlate interaction patterns with DNA shape too. Here, major groove width of CGG (cognate sequence) was observed to be larger than other motifs (17-19 Å) (Supplementary Figure S8A). Base pair step parameters at the DNA ends show similar error ranges for CGG and CGA motifs, as well as for CGC and CGT (Supplementary Figure S8B).

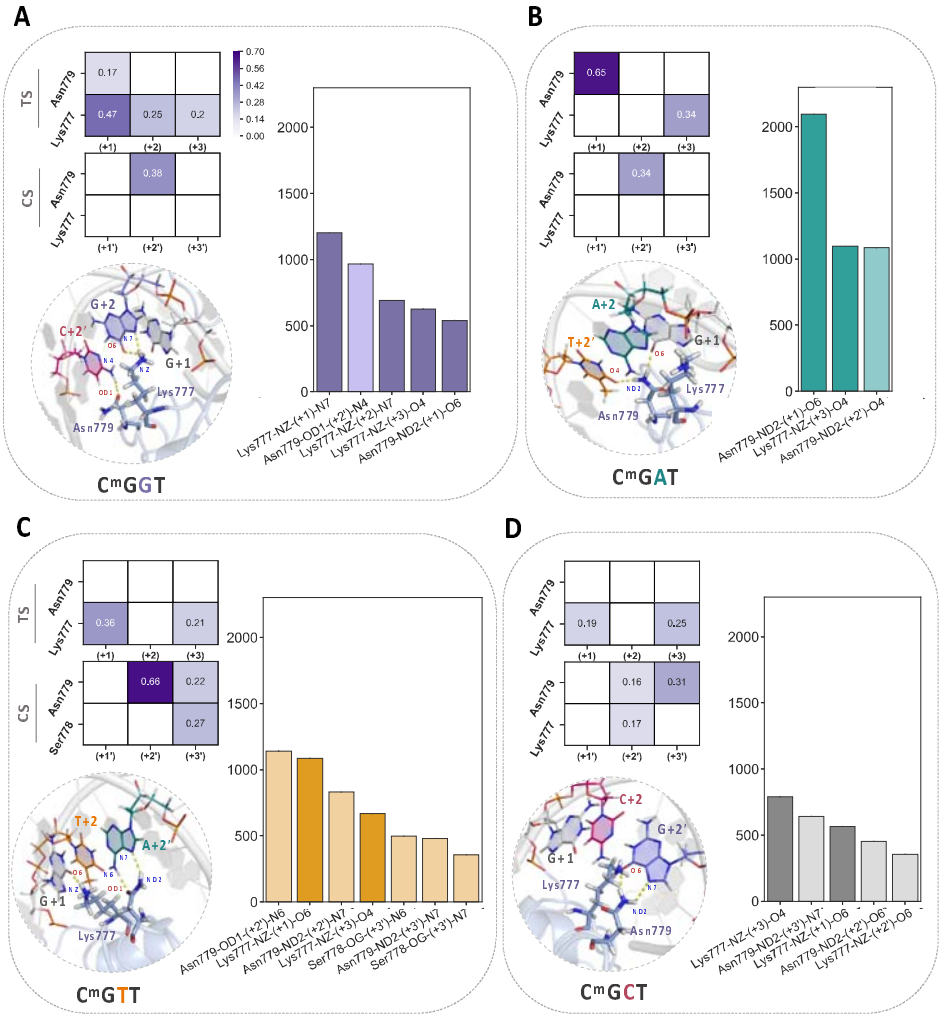

Figure 5. Interaction analysis of DNMT3B cognate (CGG), cognate-like (CGA) and noncognate systems (CGT/C). Flanking nucleotide interactions of DNMT3B complex with **(A)** CGG, **(B)** CGA, **(C)**

CGT, **(D)** CGC. Each panel contains a heatmaps, showing the interaction intensities of the important amino acids with +1-to-+3 and its complementary strand (Methods). In the heatmaps, the first row represents the target strand (TS) interactions, while the second row shows the complementary strand (CS) ones. The panels also contain the distribution of base-specific hydrogen bonds analyzed at the atomistic scale (over 1.6 μs MD trajectory), covering the range of interactions depicted by the heatmap. Light colored bars represent the complementary strand interactions, whereas dark colored bars represent target strand interactions. The panels lastly present the structural snapshots of the important base-specific interactions highlighted in the heatmap and barplot.

Table 2. Summary of base-specific hydrogen bonds formed between DNMT3B and different DNA sequences. The hydrogen bond forming atoms and their observation frequencies are explicitly stated (over the whole 1.6 μs MD trajectory for a given DNMT3A complex).

DNMT3B	**CGG**	**CGA**	**CGT**	**CGC**
Lys777	**G+1** NZ-N7 37.5%	**T+3** NZ-O4 34.2%	**G+1** NZ-O6 33.9%	**T+3** NZ-O4 24.66%
	G+2 NZ-N7 21.5%		**T+3** NZ-O4 20.8%	**G+1** NZ-O6 17.7%
	T+3 NZ-O4 19.6%			**G+2'** NZ-O6 11.11%
Asn779	**C+2'** OD1-N4 30.1%	**G+1** ND2-O6 65.4%	**A+2'** OD1-N6 35.6% ND2-N7 25.9%	**A+3'** ND2-N7 20.07%
	G+1 ND2-O6 16.9%	**T+2'** ND2-O4 33.9%	**A+3'** ND2-N7 15%	**G+2'** ND2-O6 14.14%
Ser778	*	*	**A+3'** OG-N6 15.5% OG-N7 11.1%	*

DISCUSSION

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 in their enzymatic domains, these proteins preferentially methylate close but distinct sequences. In this work, we unraveled the +2 (CGX) DNA sequence selectivity mechanisms of these enzymes by using extended MD simulations and performing a comparative interaction dynamics framework, i.e., IDF, analysis. As an outcome, we observed that in the case of cognate DNMT3A-CGC complex, the long and branched side chain of DNMT3A-Arg836 is directly reading G+2′ (C+2 complementary) by forming stable bidentate hydrogen bonds with N7 and O6 atoms of G+2′. In the case of CGA, a single stable bond O4 of T+2′ and NH1 of Arg836 in approximately 70% of conformations, maintaining selectivity at the +2 position. In contrast, the CGT motif shows less stable hydrogen bonding patterns distributed over different atom pairs, resulting in a decreased selectivity at the +2 position. Interestingly, in the non-cognate CGG sequence, the most frequent interaction occurs between the OD1 atom of Asn838 and the N4 atom of C+2′. When comparing these interaction network analyses with the experimental results, we saw that, like in our observation, higher enzyme activity was noted in the CGA sequence than in CGT by Gao et al, 2020 (14).

DNMT3B, on the other hand, uses its Lys777 (the DNMT3A-Arg836 counterpart) and Asn779 to recognize the cognate CGG. In this binding mode, Asn779 plays a significant role in recognizing C+2′ through its hydrogen acceptor OD1, while Lys777 formed base-specific hydrogen bonds with G+1 and G+2. As mentioned above, the same interaction pattern was surprisingly observed for Arg836 and Asn838 when DNMT3A is bound to CGG. On the other hand, in the case of CGA, Asn779 primarily acts as a hydrogen donor to G+1 and T+2′. In CGT, Lys777 forms hydrogen bonds with the O6 atom of G+1, and Asn779 forms bidentate hydrogen bonds with A+2′ both donating and accepting hydrogen. In non-cognate CGC, Lys777 interacts with G+1 and even extends to G+4′, forming hydrogen bonds with O6 of G+1 and O4 of T+3, while Asn779 primarily interprets G+2' and A+3'. All these observations aligns well with the findings discovered by various cell-scale methylation and enzymatic activity assays, as well as structural studies, which 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 the DNA sequence selectivity of DNMT3A/B enzymes (10, 13–16, 45, 46).

So, the DNMT3A/B +2 sequence selectivity process relies on stable hydrogen bond formation between arginine or lysine residues and guanine nucleotides. This was highlighted by the simple, yet elegant early observations made by Luscombe et al. (2001)(41), who demonstrated a strong preference for arginine and lysine residues in major groove guanines, emphasizing the significance of hydrogen bond distribution at the major groove in DNA sequence recognition. In their work, Luscombe et al. showed that arginine and lysine act as multiple hydrogen donors to recognize a neighboring guanine. They also noted two dominant DNA interaction modes for arginine: the "end-on," where arginine interacts with its distal nitrogen atoms, and the "side-on," where arginine uses its NE or NH1 atoms for hydrogen bonding. Our atomic-resolution analyses of the DNMT3A-CGC complex align well with these findings, identifying the same interaction modes for Arg836. Additionally, interactions between arginine and guanine in the CGT and CGG complexes in DNMT3A and also lysine and guanine interactions in all the other complexes of DNMT3B resonate with these interaction modes. Based on these observations, we claim that the unique characteristics of arginine, including its long, branched side chain, conformational flexibility, and ability to form strong hydrogen bonds, are pivotal for its selective and specific interactions with CGC in DNMT3A. Conversely, limited conformational flexibility due to the limited hydrogen donating capacity of lysine restricts its interaction versatility. This fundamental difference explains why Arg836 predominantly directs the interaction network in DNMT3A as an SDP, whereas Lys777 in DNMT3B requires coordination with Asn779, acting as dual SDPs. As a result of our IDF analysis, we also found that single nucleotide substitutions in other CG motifs significantly alter the DNMT3A-DNA interaction profile, whereas DNMT3B appears less affected, reflecting DNMT3A's higher specificity to CpG motifs (14, 16). We further discovered that the highest interaction intensity observed for DNMT3A-DNA is 3-fold higher compared to one of DNMT3B-DNA (1.9 to 0.66). These are in line with earlier experimental studies by Gao et al. (deep enzymology) and Mallona et al. (whole-genome bisulfite) where they observed higher CpG specificity for DNMT3A.

In addition to hydrogen bond distributions, we evaluated the intermolecular ionic bond profiles for both enzymes. As an outcome, we observed more interacting amino acid-nucleotide pairs in DNMT3A (6-8 pairs in DNMT3A, 4-6 pairs in DNMT3B) (Supplementary Figure S9A). So, DNMT3A anchors to the DNA at more positions than DNMT3B (Supplementary Figure S9B). We uncovered that this was due to amino acid differences at the TRD loop. Especially, Arg831, for which there is DNMT3B Gln772 at the same position, leads to more ionic bond formations in DNMT3A. We correlated this finding with an earlier profiled enzyme kinetics of DNMT3A and DNMT3B (2018) (47). According to this work, 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 (47, 48). This enzymatic study illuminates the cooperative nature of DNMT3A and the processive behavior of DNMT3B. Here, lesser pairwise ionic bond interaction in DNMT3B can be attributed to its processive nature, a more freely sliding pattern on DNA.

All these results highlight how minor variations in amino acids can result in significant differences in sequence selectivity. It also shows that to fully understand the selective interactions between amino acids and nucleotides, it is crucial to adopt a dynamic perspective. In this regard, MD simulations emerge as the sole method capable of deciphering the connection between these variations and their functional implications, as we showed in our previous works (19, 20). However, to understand the role of whole sequence space in DNMT3 selectivity, one should scan the selective and differential interactions over a wider flanking nucleotide range, such as -3 to +3, as suggested in recent experimental studies (14, 46, 49). Such studies can be expanded to non-CpG DNA sequences too. Moreover, the proposed allosteric effect introduced by DNMT3L should also be considered in characterizing all factors contributing to selective *de novo* DNA methylation (50). We propose that our IDF platform can be applied in all these scenarios to unravel the selective DNA sequence space of DNA methyltransferases.

All in all, in this work, we present the recognition mechanisms of DNMT3A and DNMT3B at the atomistic scale by using computational approaches. Our study not only provides crucial

knowledge to develop therapeutics that target aberrant and specific DNMT3A/B-DNA interactions but it also provides a computational framework to study other paralogous protein-DNA recognition mechanisms.

DATA AVAILABILITY

All the analysis scripts and files, parameter files of MD simulations, relevant structural models and supplementary movies are deposited on GitHub (under: https://github.com/CSB-KaracaLab/DNMT3AB_specificity).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

AUTHOR CONTRIBUTIONS

Ayşe Berçin Barlas: Conceptualization, Formal analysis, Methodology, Visualization, Validation, Data curation, Writing—original draft, Writing—review & editing. Ezgi Karaca: Conceptualization, Funding acquisition, Investigation, Methodology, Writing—original draft, Writing—review & editing, Supervision, Project administration.

ACKNOWLEDGEMENTS

The preliminary data of this work was 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 would like to thank IBG for HPC resources that enabled our analyses and simulations. We also would like to thank João P. G. L. M. Rodrigues for the critical reading of our manuscript and his invaluable feedback.

FUNDING

E.K. and A.B.B acknowledge EMBO Installation Grant (Project Number: 4421). A.B.B. acknowledges TÜBİTAK 2232/A International Fellowship for Outstanding Researchers Program (Project Number: 118C354) and TÜBİTAK 2211/A National PhD Scholarship Program.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. Schübeler,D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.

2. Chen,Z. and Zhang,Y. (2020) Role of Mammalian DNA Methyltransferases in Development. *Annu. Rev. Biochem.*, **89**, 135–158.

3. Dor,Y. and Cedar,H. (2018) Principles of DNA methylation and their implications for biology and medicine. *The Lancet*, **392**, 777–786.

4. Kaneda,M., Sado,T., Hata,K., Okano,M., Tsujimoto,N., Li,E. and Sasaki,H. (2004) Role of de novo DNA methyltransferases in initiation of genomic imprinting and X-chromosome inactivation. *Cold Spring Harb. Symp. Quant. Biol.*, **69**, 125–129.

5. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M., *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

6. Schmitz,R.J., Lewis,Z.A. and Goll,M.G. (2019) DNA Methylation: Shared and Divergent Features across Eukaryotes. *Trends Genet.*, **35**, 818–827.

7. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.

8. Song,J., Teplova,M., Ishibe-Murakami,S. and Patel,D.J. (2012) Structure-Based Mechanistic Insights into DNMT1-Mediated Maintenance DNA Methylation. *Science*, **335**, 709–712.

9. Gowher,H. and Jeltsch,A. (2002) Molecular Enzymology of the Catalytic Domains of the Dnmt3a and Dnmt3b DNA Methyltransferases*. *J. Biol. Chem.*, **277**, 20409–20414.

10. Lin,C.-C., Chen,Y.-P., Yang,W.-Z., Shen,J.C.K. and Yuan,H.S. (2020) Structural insights into CpG-specific DNA methylation by human DNA methyltransferase 3B. *Nucleic Acids Res.*, **48**, 3949–3961.

11. Bestor,T.H. (2000) The DNA methyltransferases of mammals. *Hum. Mol. Genet.*, **9**, 2395–2402.

12. Vukic,M. and Daxinger,L. (2019) DNA methylation in disease: Immunodeficiency, Centromeric instability, Facial anomalies syndrome. *Essays Biochem.*, **63**, 773–783.

13. Zhang,Z.-M., Lu,R., Wang,P., Yu,Y., Chen,D., Gao,L., Liu,S., Ji,D., Rothbart,S.B., Wang,Y., *et al.* (2018) Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature*, **554**, 387–391.

14. Gao,L., Emperle,M., Guo,Y., Grimm,S.A., Ren,W., Adam,S., Uryu,H., Zhang,Z.-M., Chen,D., Yin,J., *et al.* (2020) Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nat. Commun.*, **11**, 3355.

15. Mao,S.-Q., Cuesta,S.M., Tannahill,D. and Balasubramanian,S. (2020) Genome-wide DNA Methylation Signatures Are Determined by DNMT3A/B Sequence Preferences. *Biochemistry*, **59**, 2541–2550.

16. Mallona,I., Ilie,I.M., Karemaker,I.D., Butz,S., Manzo,M., Caflisch,A. and Baubec,T. (2021) Flanking sequence preference modulates *de novo* DNA methylation in the mouse genome. *Nucleic Acids Res.*, **49**, 145–157.

17. Emperle,M., Adam,S., Kunert,S., Dukatz,M., Baude,A., Plass,C., Rathert,P., Bashtrykov,P. and Jeltsch,A. (2019) Mutations of R882 change flanking sequence preferences of the DNA methyltransferase DNMT3A and cellular methylation patterns. *Nucleic Acids Res.*, **47**, 11355–11367.

18. Anteneh,H., Fang,J. and Song,J. (2020) Structural basis for impairment of DNA methylation by the DNMT3A R882H mutation. *Nat. Commun.*, **11**, 2294.

19. Karakulak,T., Rifaioglu,A.S., Rodrigues,J.P.G.L.M. and Karaca,E. (2021) Predicting the Specificity- Determining Positions of Receptor Tyrosine Kinase Axl. *Front. Mol. Biosci.*, **8**.

20. Ozden,B., Boopathi,R., Barlas,A.B., Lone,I.N., Bednar,J., Petosa,C., Kale,S., Hamiche,A., Angelov,D., Dimitrov,S., *et al.* (2023) Molecular Mechanism of Nucleosome Recognition by the Pioneer Transcription Factor Sox. *J. Chem. Inf. Model.*, **63**, 3839–3853.

21. Rausell,A., Juan,D., Pazos,F. and Valencia,A. (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 1995–2000.

22. Sloutsky,R. and Naegle,K.M. (2016) High-Resolution Identification of Specificity Determining Positions in the LacI Protein Family Using Ensembles of Sub-Sampled Alignments. *PloS One*, **11**, e0162579.

23. Madeira,F., Park,Y. mi, Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D., *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.

24. Robert,X. and Gouet,P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.*, **42**, W320–W324.

25. Li,S., Olson,W.K. and Lu,X.-J. (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.*, **47**, W26–W34.

26. Takeshita,K., Suetake,I., Yamashita,E., Suga,M., Narita,H., Nakagawa,A. and Tajima,S. (2011) Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). *Proc. Natl. Acad. Sci.*, **108**, 9055–9059.

27. van Zundert,G.C.P., Rodrigues,J.P.G.L.M., Trellet,M., Schmitz,C., Kastritis,P.L., Karaca,E., Melquiond,A.S.J., van Dijk,M., de Vries,S.J. and Bonvin,A.M.J.J. (2016) The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.*, **428**, 720–725.

28. Van Der Spoel,D., Lindahl,E., Hess,B., Groenhof,G., Mark,A.E. and Berendsen,H.J.C. (2005) GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, **26**, 1701–1718.

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

30. Sousa da Silva,A.W. and Vranken,W.F. (2012) ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes*, **5**, 367.

31. Mark,P. and Nilsson,L. (2001) Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A*, **105**, 9954–9960.

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

33. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.

34. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577–8593.

35. Hess,B. (2008) P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.*, **4**, 116–122.

36. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J., *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.

37. The pandas development team. (2024) pandas-dev/pandas: Pandas. 10.5281/zenodo.10957263.

38. Rodrigues,J., Valentine,C. and Jimenez,B. (2019) JoaoRodrigues/interfacea: First beta version of the API. 10.5281/zenodo.3516439.

39. Waskom,M.L. (2021) seaborn: statistical data visualization. *J. Open Source Softw.*, **6**, 3021.

40. Lu,X. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.

41. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.

42. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.*, **79**, 233–269.

43. Deepa,P., Kolandaivel,P. and Senthilkumar,K. (2011) Hydrogen-bonding studies of amino acid side-chains with DNA base pairs. *Mol. Phys.*, **109**, 1995–2008.

44. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **73**, 804–808.

45. Dukatz,M., Adam,S., Biswal,M., Song,J., Bashtrykov,P. and Jeltsch,A. (2020) Complex DNA sequence readout mechanisms of the DNMT3B DNA methyltransferase. *Nucleic Acids Res.*, **48**, 11495–11509.

46. Dukatz,M., Dittrich,M., Stahl,E., Adam,S., de Mendoza,A., Bashtrykov,P. and Jeltsch,A. (2022) DNA methyltransferase DNMT3A forms interaction networks with the CpG site and flanking sequence elements for efficient methylation. *J. Biol. Chem.*, **298**, 102462.

47. Norvil,A.B., Petell,C.J., Alabdi,L., Wu,L., Rossie,S. and Gowher,H. (2018) Dnmt3b Methylates DNA by a Noncooperative Mechanism, and Its Activity Is Unaffected by Manipulations at the Predicted Dimer Interface. *Biochemistry*, **57**, 4312–4324.

48. Gowher,H. and Jeltsch,A. (2018) Mammalian DNA methyltransferases: new discoveries and open questions. *Biochem. Soc. Trans.*, **46**, 1191–1202.

49. Jeltsch,A., Adam,S., Dukatz,M., Emperle,M. and Bashtrykov,P. (2021) Deep Enzymology Studies on DNA Methyltransferases Reveal Novel Connections between Flanking Sequences and Enzyme Activity. *J. Mol. Biol.*, **433**, 167186.

50. Wienholz,B.L., Kareta,M.S., Moarefi,A.H., Gordon,C.A., Ginno,P.A. and Chédin,F. (2010) DNMT3L modulates significant and distinct flanking sequence preference for DNA methylation by DNMT3A and DNMT3B in vivo. *PLoS Genet.*, **6**, e1001106.

TABLE AND FIGURES LEGENDS

Table 1. Summary of base-specific hydrogen bonds formed between DNMT3A with different DNA sequences. The hydrogen bond forming atoms and their observation frequencies are explicitly stated (over the whole 1.6 μs MD trajectory for a given DNMT3A complex).

Table 2. Summary of base-specific hydrogen bonds formed between DNMT3B and different DNA sequences. The hydrogen bond forming atoms and their observation frequencies are explicitly stated (over the whole 1.6 μs MD trajectory for a given DNMT3A complex).

Figure 1. Differential amino acid positions in DNMT3A and DNMT3B MTase domains. (A) A pairwise sequence alignment (with EBLOSUM62 matrix) was performed for the MTase domains of DNMT3A and DNMT3B over their 283 amino acids. The conserved residues are highlighted in red, while the non-conserved residues with similar physicochemical properties are marked in yellow. Black residues with a white background indicate non-conserved amino acids. The sequence identity and similarity values are 80.2% and 91.2%, respectively. **(B)** The non-conserved amino acids of DNMT3A (upper) and DNMT3B (down) are located on the TRD-loop (Arg831-to-Gln772, Arg836-to-Lys777, Asp845-to-Asn786, His847-to-Leu788) and on the catalytic loop (Ile715-to-Asn656). Yellow residues represent the TRD loop, while orange residues represent the catalytic loop.

Figure 2. 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). 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 (13)). In this crystal structure, Arg836 forms a base specific hydrogen bond with G+1 of CGC (yellow) (13). Right: DNMT3B is bound to CGG (PDB ID: 6KDA (10)). In this crystal structure, Lys777 forms van der Waals interactions with G+2 (green) (10) and Asn779 forms a hydrogen bond with G+1 of CGG (yellow) (14).

Figure 3. An interaction dynamics framework (IDF) to identify DNMT3 selectivity +2 flanking 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 hydrogen 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.

Figure 4. 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. Each panel contains a heatmaps, showing the interaction intensities of the important amino acids with +1-to-+3 and its complementary

strand (Methods). In the heatmaps, the first row represents the target strand (TS) interactions, while the second row shows the complementary strand (CS) ones. The panels also contain the distribution of base-specific hydrogen bonds analyzed at the atomistic scale (over 1.6 μs MD trajectory), covering the range of interactions depicted by the heatmap. Light colored bars represent the complementary strand interactions, whereas dark colored bars represent target strand interactions. The panels lastly present the structural snapshots of the important base-specific interactions highlighted in the heatmap and barplot.

Figure 5. 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. Each panel contains a heatmaps, showing the interaction intensities of the important amino acids with +1-to-+3 and its complementary strand (Methods). In the heatmaps, the first row represents the target strand (TS) interactions, while the second row shows the complementary strand (CS) ones. The panels also contain the distribution of base-specific hydrogen bonds analyzed at the atomistic scale (over 1.6 μs MD trajectory), covering the range of interactions depicted by the heatmap. Light colored bars represent the complementary strand interactions, whereas dark colored bars represent target strand interactions. The panels lastly present the structural snapshots of the important base-specific interactions highlighted in the heatmap and barplot.