

T.C.
DOKUZ EYLUL UNIVERSITY
İZMİR INTERNATIONAL BIOMEDICINE AND GENOME
INSTITUTE

**MOLECULAR DETERMINANTS OF DNA
METHYLATION MECHANISMS IN
MAMMALS**

DENİZ DOĞAN

MOLECULAR BIOLOGY AND GENETICS

MASTER OF SCIENCE THESIS

İZMİR-2020

TEZ KODU: DEU.IBG.MSc-2017850032

T.C.
DOKUZ EYLUL UNIVERSITY
IZMIR INTERNATIONAL BIOMEDICINE AND GENOME
INSTITUTE

**MOLECULAR DETERMINANTS OF DNA
METHYLATION MECHANISMS IN
MAMMALS**

MOLECULAR BIOLOGY AND GENETICS
MASTER OF SCIENCE THESIS

DENİZ DOĞAN

Supervisor: Asst. Prof. Dr. Ezgi Karaca

TEZ KODU: DEU.IBG.MSc-2017850032

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü

Genom Bilimleri ve Moleküler Biyoteknoloji Anabilim Dalı,

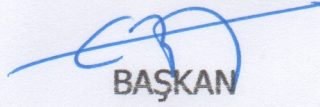
Moleküler Biyoloji ve Genetik Yüksek Lisans programı öğrencisi

Deniz Doğan

'MOLECULAR DETERMINANTS OF DNA METHYLATION

MECHANISMS IN MAMMALS' konulu Yüksek Lisans tezini 28.09.2020

tarihinde başarılı olarak tamamlamıştır.


BAŞKAN

Dr. Öğr. Üyesi Ezgi Karaca

Dokuz Eylül Üniversitesi,

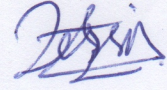
İzmir Uluslararası Biyotıp ve Genom Enstitüsü

ÜYE

Dr. Öğr. Üyesi İsmail Hakkı Akgün

Ege Üniversitesi

Biyomühendislik Anabilim Dalı



ÜYE

Dr. Öğr. Üyesi Zerrin Işık

Dokuz Eylül Üniversitesi

Bilgisayar Yazılımı Anabilim Dalı

YEDEK ÜYE

Dr. Athanasia Pavlopoulou

Dokuz Eylül Üniversitesi,

İzmir Uluslararası Biyotıp ve Genom Enstitüsü

YEDEK ÜYE

Dr. Öğr. Üyesi Burcu Kaplan

Ege Üniversitesi

Gıda Teknolojisi Ana Bilim Dalı

Dokuz Eylül University Izmir International Biomedicine and Genome Institute

Department of Genomics and Molecular Biotechnology,

Molecular Biology and Genetics graduate program, Master of Science student

Deniz DOĞAN has successfully completed his Master of Science thesis titled

'MOLECULAR DETERMINANTS OF DNA METHYLATION

MECHANISMS IN MAMMALS' on the date of 28/09/2020.



CHAIR

Asst. Prof. Dr. Ezgi KARACA

Dokuz Eylül Üniversitesi, İzmir Uluslararası Biyotıp ve Genom Enstitüsü

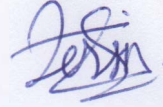
MEMBER



Asst. Prof. Dr. İsmail Hakkı AKGÜN

Ege University,

Department of Bioengineering



MEMBER

Asst. Prof. Dr. Zerrin IŞIK

Dokuz Eylül University,

Department of Computer Engineering

SUBSTITUTE MEMBER

Dr. Athanasia PAVLOPOULOU

Dokuz Eylül University,

Izmir International

Biomedicine and Genome Institute

SUBSTITUTE MEMBER

Asst. Prof. Burcu KAPLAN

Ege University,

Department of Food Engineering

Table of Contents

INDEX OF TABLES.....	iii
INDEX OF FIGURES.....	iv
LIST OF ABBREVIATIONS.....	vi
ACKNOWLEDGEMENTS.....	vii
ABSTRACT	1
ÖZET	2
1. INTRODUCTION AND AIM.....	3
1.1. STATEMENT AND IMPORTANCE OF THE PROBLEM.....	3
1.2. AIM OF THE STUDY	3
1.3. HYPOTHESIS OF THE STUDY	3
2. GENERAL INFORMATION	5
2.1. The most abundant epigenetic mark: DNA Methylation	5
2.2. C5-DNA Methylation from an Evolutionary Perspective.....	6
2.3. DNA Methylation in Bacteria	8
2.4. A structural insight into Bacterial DNA Methyltransferases	8
2.5. DNA Methylation in Mammals	9
2.5.1. <i>The working principles of DNMT3A/B enzymes</i>	10
2.5.2. <i>The sequence specificity of DNMT3A/Bs towards distinct methylation sites...</i>	12
2.5.3. <i>The state-of-the-art structural knowledge on DNMT3A/Bs</i>	13
2.6. A computational perspective to study protein-DNA interaction specificities.....	15
3. MATERIALS AND METHODS	18
3.1. Type of Study	18
3.2. Time and Location of Study.....	18
3.3. Population and Sample of Study	18
3.4. Materials of Study	18
3.5. Variables of Study.....	18
3.6. Tools for Data Collection.....	18
3.6.1. <i>Web-servers & resources used for data collection</i>	18
3.6.2. <i>Methodology</i>	21
3.7. Study Plan and Calendar	38
3.8. Data Evaluation.....	39
3.9. Limitations of Study.....	39
3.10. Ethic Committe Approval	39

4. RESULTS.....	40
4.1. Building Homology Models of bacterial <i>HhaI</i> methyltransferases	40
4.2. Investigation of the working principles of bacterial <i>HhaI</i> methyltransferases	40
4.2.1. Interface comparison of different <i>HhaI</i> MTase states.....	40
4.2.2. Determining the crucial amino acids mediating <i>HhaI</i> MTase specificity enzyme 41	
4.2.3. Linking <i>HhaI</i> information with the mammalian <i>de novo</i> methyltransferases ..	45
4.3. Exploring the working principles of mammalian <i>de novo</i> methyltransferases	46
4.3.1. Template-based modelling the DNMT3B.....	46
4.3.2. Investigation of the working principles of DNMT3A/Bs towards hemi- methylated and non-methylated DNA	48
4.3.3. Constructing Homology Models of DNMT3A/B:DNA	49
4.3.4. Preparing models for MD	50
4.3.5. Water Refinement of DNMT3A/B:DNA Models.....	50
4.3.6. Deciding the best initial structure for the MD simulations.....	50
4.3.7. Assessment of the convergence of MD simulations.....	51
4.3.8. The persistency of trajectories	51
4.3.9. The convergence of trajectories.....	51
4.3.10. DNMT3B binds more stably to any DNA sequences compared to DNMT3A. .	57
4.3.11. Comprehensive analysis of interaction network	58
4.3.12. The Distribution of Total Hydrophobic bonds	67
4.3.13. The evaluation of the loop dynamics via solvent accessible surface areas (SASA) 70	
4.3.14. The investigation of internal dynamics of DNMT3A/B complexes	72
4.3.15. Essential amino acids.....	75
5. DISCUSSION	76
6. CONCLUSION AND SUGGESTIONS	82
REFERENCES.....	83
APPENDIX A.	86
APPENDIX B	90
APPENDIX C.	92
APPENDIX D.	95

INDEX OF TABLES

Table 1: Schematic representation of DNA sequences utilized in modelling.....	25
Table 2: Adjusted parameters for the optimization of hemi-methylated and non-methylated CpG bound DNA methyltransferases.....	28
Table 3: Adjusted parameters for the water refinement.....	29
Table 4: Summary of HADDOCK-based scores for each HhaI models.....	41
Table 5: The list of critical amino acids dissected for the HhaI enzyme (5MHT) and their corresponding amino acids in DNMT3A/B.	45
Table 6: Summary of DNA-binding Interface Statistics.....	48
Table 7: Summary of HADDOCK statistics is listed below.....	50
Table 8: Time percentage of identified intermolecular hydrophobic bonds in DNMT3A:DNA complexes.....	68
Table 9: Time percentage of intermolecular hydrophobic bonds in DNMT3B:DNAs.....	70

INDEX OF FIGURES

Figure 1: A general perspective into the DNA methylation.....	5
Figure 2: A) Target CYT (pink) and SAM (blue) are enclosed by conserved motifs (white spheres) of the catalytic MTase domain. B) CYS of motif IV, two ARGs of motif VIII and GLU of motif IV approach to SAM and target CYT to mediate methylation reaction.	7
Figure 3: DNMT3s mediates de novo methylation after flipping out target cytosine to be methylated on the target strand. After this, DNMT1 and/or DNMT3B carries out hemi-methylation to acquire fully methylated DNA.	10
Figure 4: Indicative states of DNMT3A/B-mediated oligomerization..	11
Figure 5: Critical amino acids contributing to working principles of DNMT3A..	14
Figure 6. The workflow of minnie and the schematic representation of folder organization..	35
Figure 7: Dissecting critical amino acids for the sequence specificity of HhaI enzyme.	42
Figure 8: Comparison of contact maps to discriminate the sequence specificity of HhaI enzyme towards hemi-methylated (5CM) and non-methylated (CYT) DNA.....	44
Figure 9: Pairwise alignment of methyltransferase domains of human DNMT3A and human DNMT3B..	47
Figure 10: Homology models of DNMT3A/Bs bound to DNA and SAM..	49
Figure 11: The RMSD of DNMT3A:DNA complexes throughout simulations.....	52
Figure 12: The RMSD of DNMT3B:DNA complexes throughout simulation.....	53
Figure 13: The Rg of DNMT3A:DNA complexes throughout simulation.	55
Figure 14: The Rg of DNMT3B:DNA complexes throughout simulation..	56
Figure 15: The distribution of complex surface areas of DNMT3A:DNA-SAM complexes..	58
Figure 16: Distribution of total number of intermolecular hydrogen bonds. e.	60
Figure 17: The distribution complex-specific hydrogen bonds in DNMT3A/B:DNA-SAM complexes.....	62
Figure 18: Time-dependent behaviours of intermolecular hydrogen bonds for DNMT3A:DNA-SAM complexes.	64
Figure 19: Time-dependent behaviours of intermolecular hydrogen bonds for DNMT3B:DNA-SAM complexes.	66
Figure 20: The distribution of intermolecular hydrophobic bonds in the DNMT3A:DNA-SAM complexes..	68
Figure 21: The distribution of intermolecular hydrophobic bonds in the DNMT3B:DNA-SAM complexes.....	69
Figure 22: SASA-based Distribution of TRD and catalytic loops in DNMT3A/B:DNA complexes.....	72
Figure 23: Signaling propensity of the catalytic loop with respect to the rest of DNMT3A/B:DNA complexes..	73

Figure 24: Signaling propensity of the TRD loop with respect to the rest of DNMT3A/B:DNA complexes.....	74
---	----



LIST OF ABBREVIATIONS

Å	Angstrom
5CM	hemi-methylated DNA (methylated cytosine)
5MHT	HhaI methyltransferase
CpG	Cytosine followed by Guanine
CSA	Complex Surface Area
CYT	non-methylated DNA (non-methylated cytosine)
DNMT	Mammalian DNA Methyltransferases
DNMT3A/B	Mammalian de novo DNA methyltransferase 3A/B
H	Hypothesis
HADDOCK	High Ambiguity Driven Biomolecular DOCKing
MD	Molecular Dynamics
minnie	Molecular INTERaction N fIngerprints
MTase	Catalytic domain of DNA methyltransferase
PDB	Protein Data Bank
SAH	S-adenosylhomocysteine
SAM	S-Adenosyl methionine
SASA	Solvent accessible surface area
TRD	Target Recognition Domain
Xtal	Crystal

ACKNOWLEDGEMENTS

First and foremost, I would like to thank to my thesis advisor, Dr. Ezgi Karaca, for her invaluable mentorship, constant support and patience throughout this research. I am extremely grateful for her insightful and farseeing guidance of academic career. Special thanks to support me for improving writing skills.

I also would like to thank to Dr. Seyit Kale, for always encouraging to go outside of my scope and expanding my scientific thinking. I am deeply grateful for his supportive and motivational approach.

I would like to also thank to SurfSARA, TRUBA and IBG-HPC for offering many great opportunities to run molecular dynamics simulations and analysis.

Many thanks to my office mates, Burcu Özden, Eda Şamiloğlu, Işıl Takan and İbrahim Toy, for such a pleasant work environment, coffee breaks and their friendship.

I would like to thank to Burcu Ekinci, Tevfik Hatipoğlu, Esra Esmeray, Yavuz Mercan and Yağmur Toktay for all the cherishing memories and for always helping me to overcome any obstacles throughout this journey.

I would like to also thank to my beloved friend Necip Kuru for his understanding and for adding many colors to my life. Special thanks for all the white lilies.

I would like to thank to my sister for always being there for me and especially for her patience while writing this thesis. Last but definitely not least, I would like to express my gratitude to the most great-hearted man in my life, my father, for his never-ending support to follow my dreams. This thesis is dedicated to my family.

MOLECULAR DETERMINANTS OF DNA METHYLATION MECHANISMS IN MAMMALS

Deniz Doğan, Izmir International Biomedicine and Genome Institute, Dokuz Eylul
University Health Campus, Balçova, 35340, Izmir / Turkey

ABSTRACT

DNA methylation is intrinsic to the integrity of diverse organisms due to its impact on the genomic stability and gene expression. The most common type of DNA methylation occurs by transferring the methyl group of SAM to the C5 atom of cytosine nucleotide. The *anew* (*de novo*) and/or maintenance of DNA methylation patterns are imposed by the catalytic domain of DNA methyltransferases. The critical mutations occurring at this domain lead to developmental defects or cancer biogenesis. Thereby, discovering the working principles of DNA methyltransferases is crucial towards developing potent therapeutics combatting DNA methyltransferase malfunction.

Mammals have three paralogous catalytically active DNA methyltransferases: DNMT1, DNMT3A and DNMT3B. DNMT3A/Bs predominantly take role in *de novo* methylation, while DNMT1 and/or DNMT3B maintain the pre-established DNA methylation pattern. Methylation site (CpG or non-CpG) preference of these enzymes is encoded in the flanking region surrounding the target cytosine. As an example, DNMT3A prefers -CAC- over -CAG- sequence, while the reverse is true for DNMT3B.

Expanding on this notion, in this work, we aimed to investigate the mechanistic principles of DNMT3A/B enzymes with computational structural biology approaches. For this, we initially studied the structure of a DNA methyltransferase enzyme from a simple organism. This allowed us to define the evolutionary conserved structural elements of the DNA methyltransferases and to transfer them to mammalian DNMT3A/Bs. Besides, we also explored the stability, time-dependent intermolecular interaction networks and dynamics of DNMT3A/B:DNA complexes.

Keywords: DNA methyltransferases, Specificity, Molecular Dynamics, HADDOCK, Structural Modelling

Memelilerde DNA Metilasyon Mekanizmasını Belirleyen Moleküler Faktörler

Deniz Doğan, İzmir Uluslararası Biyotıp ve Genom Enstitüsü, Dokuz Eylül Üniversitesi,
Sağlık Kampüsü, Balçova, 35340, İzmir / Turkey

ÖZET

DNA metilasyonu, genomik kararlılığı ve gen ifadesini şekillendirdiği için canlıların ayrılmaz bir parçasıdır. En yaygın görülen C5-DNA metilasyonu, SAM'in metil grubunun, hedeflenen sitozin nükleotidinin C5 atomuna transfer edilmesi ile gerçekleşir. DNA metilasyonun yayınlanması ve/veya devamlılığı, DNA metiltransferaz enzimlerinin katalitik domaini tarafından sağlamaktadır. DNA metil transferazların katalitik domainlerinde meydana gelen mutasyonlar, kansere ve/veya gelişimsel bozukluklara yol açmaktadır. Etkili tedavi yöntemleri geliştirebilmek adına, DNA metiltransferazların çalışma mekanizmalarını ortaya çıkartmak hayati bir öneme sahiptir.

Memelilerde katalitik olarak aktif olan üç adet DNA metiltransferaz enzimi bulunmaktadır: DNMT1, DNMT3A ve DNMT3B. DNMT3A/B DNA metilasyon desenlerinin yayınlanmasından sorumluyken, DNMT1 ve/veya DNMT3B yayınlanmış olan metilasyon desenlerinin devamlılığını sağlamaktadır. Ayrıca, DNMT3A/B enzimleri potansiyel metilasyon noktasına ve metilasyon noktasını çevreleyen sekanslara bakarak hangi metilasyon noktasının hedefleneceğine karar verir.

Bu bilgiler ışığında, hesaplamalı biyoloji tekniklerini kullanarak DNMT3A/B'lerin çalışma mekanizmasını ortaya koymayı hedefledik. İlk olarak, sıkça çalışılmış olan bakteriyel HhaI metiltransferaz enzimini araştırdık. Bu sayede, memeli DNMT3A/B enzimlerinin çalışma mekanizmalarını araştırırken kullanmak için bir temel oluşturduk. Sonrasında, DNMT3A/B:DNA komplekslerinin kararlılığına, etkileşim örüntüsüne ve komplekslerin içsel dinamiklerine baktık.

Keywords: DNA metiltransferaz, Özgüllük, Moleküler Dinamik, HADDOCK, Yapısal modelleme

1. INTRODUCTION AND AIM

1.1. STATEMENT AND IMPORTANCE OF THE PROBLEM

Mammalian DNA methylation occurs in two distinct steps: i) the establishment of the DNA methylation pattern and ii) the maintenance of the available DNA methylation pattern. DNMT3A/B enzymes carry de novo DNA methylation while DNMT1 and/or DNMT3A maintains already established DNA methylation pattern. The activity of DNMTs is primarily dictated by the highly conserved catalytic methyltransferase domain. However, DNMT3A/Bs show differential sequence specificity with respect to the methylation site (CpG vs non-CpG) and flanking region. DNMT3A/Bs predominantly methylate CpG sites, to a lesser extent, CpA, CpT and CpC methylation sites. Among the CpA sites, DNMT3A favors -CAC- sequence over -CAG- sequence while DNMT3B majorly targets -CAG- sequence. Additionally, DNMT3A/Bs also impose CpG methylation differently: DNMT3As utilize either distributive or processive mechanism, while DNMT3Bs follow processive mechanism. Having this background as basis, we aim to explore the discriminative factors governing differential methylation activity and sequence specificity of DNMT3A/B enzymes. Our results would pave the way for the development of novel therapeutic approaches targeting developmental defects or cancer biogenesis, occurring due to DNMT malfunction.

1.2. AIM OF THE STUDY

In this thesis, we aimed at exploring the factors defining the differential working principles of DNMT3A/B enzymes.

1.3. HYPOTHESIS OF THE STUDY

H1: The models of DNA methyltransferases bound to DNA and methyl donor (SAM) might be constructed with confidence due to the conserved binding mode of homolog/paralog enzymes.

H2: Dissecting the crucial factors of *HhaI* methyltransferases enlighten the working principles of mammalian DNA methyltransferases.

H3: DNA-oriented structural approaches can illuminate the sequence specificity of protein:DNA complexes.

H4: Systematic comparison of intermolecular interactions can shed the light on the sequence specificity of DNA methyltransferases towards distinct methylation sites.



2. GENERAL INFORMATION

2.1. The most abundant epigenetic mark: DNA Methylation

Epigenetics describe the modifications that regulate gene expression without altering the genetic code. Epigenetic regulation is modulated through DNA methylation and/or histone modifications, such as methylation, phosphorylation, acetylation etc.. Among these, DNA methylation is a prevalent regulatory mechanism. DNA methylation was initially thought as a defense mechanism against extraneous foreign DNA, as, it was foremost detected in phage-infected bacteria (Vasu & Nagaraja, 2013).

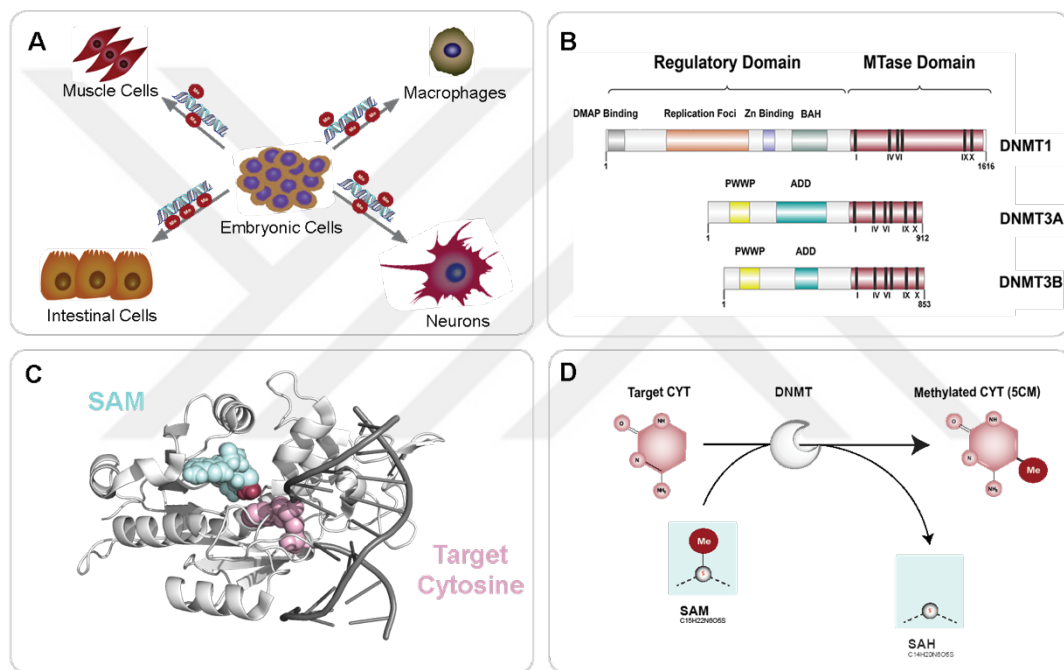


Figure 1: A general perspective into the DNA methylation. **A)** DNA methylation pattern dictates cellular differentiation of eukaryotic cells. **B)** DNA Methyltransferases (DNMTs) have variant regulatory domain and highly conserved MTase domain with six most conserved motifs. **C)** Structural representation of DNMT bound to the CpG site (PDB ID:6F57). Pink color represents target/flipped cytosine to be methylated. Blue one demonstrates the SAM that functions as a methyl donor. The reaction center is where flipped cytosine and SAM is placed **D)** DNMTs transfers a methyl group of methyl donor (SAM) to C5 atom of target cytosine.

DNA methylation occurs via transferring a methyl group (of SAM) to a target nucleotide, e.g. an adenine or a cytosine. The DNA methylation mechanism occurs via the following mechanism: (i) The DNA methyltransferase (MTase) (**Figure 1B**) wraps the target DNA with the aid of its target recognition loops (**Figure 1B-C**); (ii) The target nucleotide flips out of the DNA double-helix, such that it faces to the catalytic center of the MTase (**Figure 1C**); (iii) The methyl donor S-AdenosylMethionine (SAM), sitting at the catalytic center of the enzyme, transfers a CH₃ group to the specific atom of the target nucleotide (predominantly the C5 of target cytosine) (Fig 1D).

2.2. C5-DNA Methylation from an Evolutionary Perspective

Across species, C5-DNA MTases are evolutionary conserved. Bacterial DNMT offers a simplistic model to study DNA methylation mechanism. This enzyme is composed of only an MTase domain, while the mammalian DNMTs contain multiple domains. For instance, the mammalian *de novo* DNA methyltransferases involve a PWWP (proline–tryptophan–tryptophan–proline), an ADD (ATRX–DNMT3–DNMT3L-type) zinc finger, and a MTase domain. The PWWP and ADD domains mediate methylation activity of *de novo* DNA methyltransferases. On the other hand, *in vitro* studies have shown that a single MTase domain is enough to explain the activity and specificity of *de novo* DNA methyltransferases.

The MTase domains consist of a conserved catalytic subdomain and a variant target recognition subdomain (TRD) (Sankpal & Rao, 2002). The catalytic subdomain involves the SAM-binding site and a DNA-binding site. On the other hand, TRD has only a DNA-binding site. The MTase contains four motifs that are conserved at a moderate level and six strictly conserved motifs (I, IV, VI, VIII, IX, X). All SAM-dependent methyltransferases share motif I (FXGXG) and motif IV (PC). The motif I is responsible of contacting SAM, while the motif IV plays a crucial role during catalysis. SAM-binding is further supported by motif X (GN). Mutagenesis analysis showed that the region between motif VIII (QXRXR) and motif IX (RE) determines TRD's boundaries (Sankpal & Rao, 2002). TRD is a major determinant in DNA sequence specificity. The motif VI (ENV) and the motif VIII also contribute to DNA binding and catalysis.

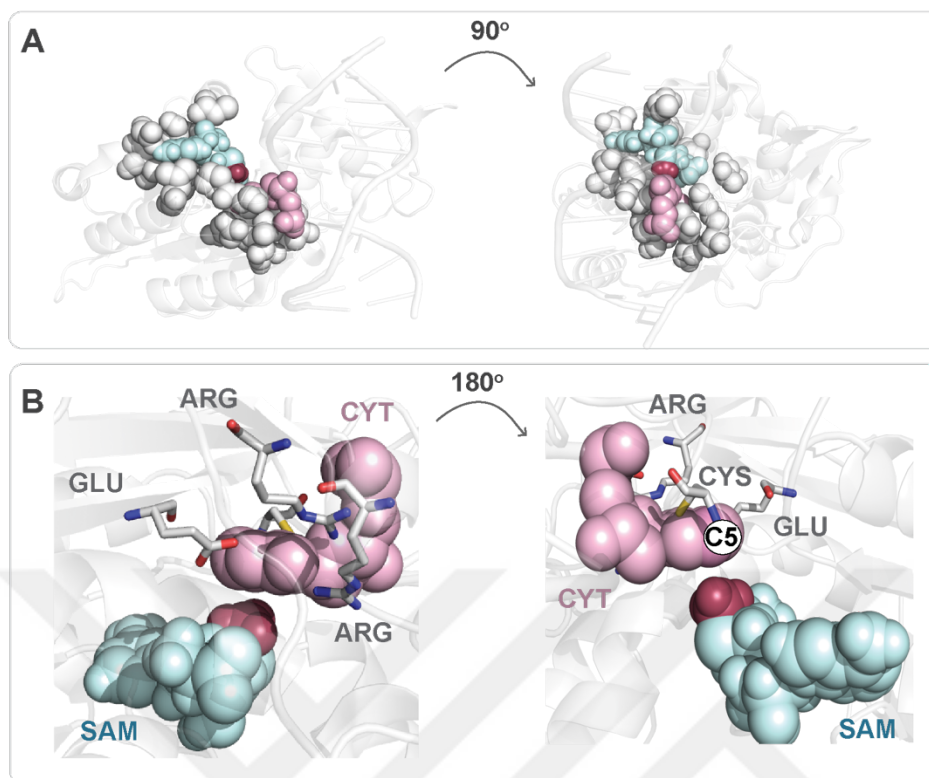


Figure 2: **A)** Target cytosine (CYT) (pink) and SAM (blue) are enclosed as conserved motifs (white spheres) of the catalytic MTase . **B)** CYS of motif IV, two ARGs of motif VIII and GLU of motif IV approach to SAM and target CYT to mediate methylation reaction.

The catalytic activity initially requires a conserved cysteine amino acid (in motif IV). This cysteine carries out a nucleophilic attack on the C6 atom of cytosine. The nucleophilic attack results in a covalent linkage between protein and DNA. This is followed by transferring $-CH_3$ group of SAM to C5 atom of target cytosine. Mutating this conserved cysteine abolishes methylation activity. The arginine of motif VIII is responsible of deprotonation of cytosine at the C5 atom, which follows the β -elimination of the cysteine. During this process, the second arginine in the motif VIII and glutamic acid in the motif IV helps cytosine positioning. This glutamic acid interacts with flipped cytosine at the N4 atom.

2.3. DNA Methylation in Bacteria

DNA methylation is a key factor to regulate cell cycle, DNA replication and DNA repair in bacteria (Moore, Le, & Fan, 2013). A bacterial DNA, *HhaI*, methyltransferase is specialized to target 5'-GCGC-3' DNA motif (Sankpal & Rao, 2002). Thereby, they show highest specificity towards this sequence, when compared to mammals. The most well-known bacterial DNA methyltransferase is *HhaI* methyltransferase (Sankpal & Rao, 2002). This enzyme methylates C5 atom of the target cytosine within 5'-GCGC-3' DNA sequence.

2.4. A structural insight into Bacterial DNA Methyltransferases

The first structure of DNA methyltransferase, *HhaI* methyltransferase, was resolved in 1994 (PDB ID: 5MHT). Its MTase domain can be dissected in two distinct subdomains, as large and small subdomains. The large subdomain is mostly composed of conserved residues. It is responsible of binding to a flipped/target cytosine and SAM, which carries out the methylation reaction. The small subdomain, TRD, is less conserved and responsible of recognizing DNA binding site.

In the DNA-free form, the conserved ARG165 interacts with another well-conserved GLU119. As a result of DNA binding, this salt bridge breaks to form a hydrogen bond between ARG165 and O atom of the target cytosine. ARG165 further makes a salt bridge by contacting the backbone of the target cytosine. In this DNA-bound state GLU119 makes a hydrogen bond with N3 and N4 atoms of target cytosine. Another conserved amino acid, PHE79, forms a hydrogen bond with the N4 atom of target cytosine. Further studies have shown that mutating GLU119 might alter the specificity of *HhaI* methyltransferases. Thereby, it has been stated that GLU119 mediated interaction with target cytosine is one of the factors governing the enzyme's specificity (Sankpal & Rao, 2002).

HhaI enzyme's small subdomain contacts with the DNA's major groove while a large subdomain interacts with minor groove (Sankpal & Rao, 2002). A gly-rich loop of small subdomain makes base-specific interactions with the 5'-GCGC-3' DNA motif. *HhaI* MTase uses GLN237 and ARG240 to recognize the flanking region surrounding the

target DNA motif (Dong et al., 2004). The *HhaI*:DNA complex is further stabilized by THR250, TYR254, and GLY257.

2.5. DNA Methylation in Mammals

As different from bacterial systems, mammals require multiple DNA methyltransferase proteins (DNMT) to establish the required DNA methylation pattern on the genome (**Figure 3**). Among the evolutionary mechanisms, in gene duplication, the organism aggregates a selective set of mutations to enhance its functional diversification. As a result, novel homologous proteins, i.e., *paralogs*, are generated. The novelty is identified either as a new function or as an increased/decreased activity/specificity towards one of proteins' substrates. The MTase domains of the mammalian DNMTs are evolved as a result of gene duplication.

DNA is at first methylated at *de novo* level. *De novo* methylation is responsible of establishing DNA methylation pattern during embryogenesis (Hervouet, Peixoto, Delage-Mourroux, Boyer-Guittaut, & Cartron, 2018). This is a key factor of cellular differentiation. *De novo* methylation is primarily carried out by *de novo* methyltransferase family: DNMT3A/B/L. DNMT3L only plays a regulatory role without showing any catalytic activity unlike DNMT3A/B (Nimura et al., 2006). It enhances DNMT3A/B's methylation activity. It has been shown that DNMT3L is predominantly critical to establish the methylation pattern during gametogenesis by managing the activity of DNMT3A enzyme. *De novo* methylation yields a single strand methylated DNA, hemi-methylated DNA. Another class of DNA methyltransferase, maintenance methyltransferase, carries out maintenance methylation. As a result of this step, the methylation information is transferred onto both DNA strands.

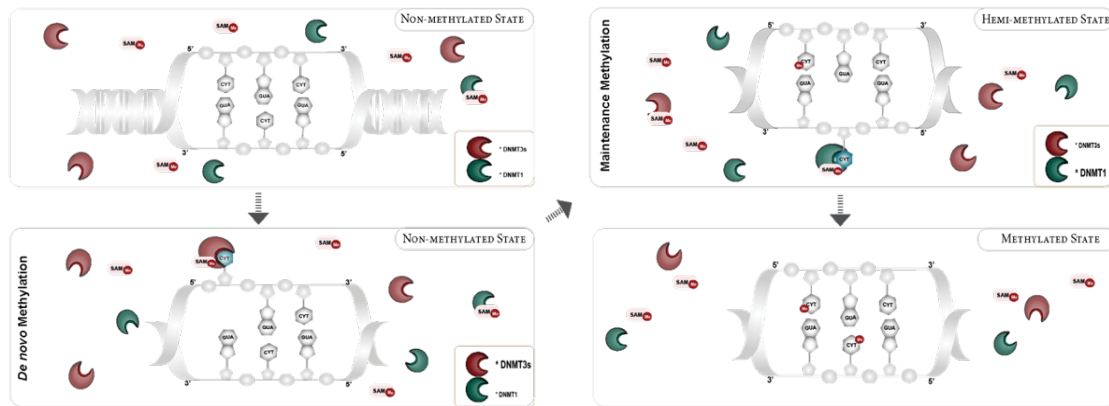


Figure 3: DNMT3s mediates de novo methylation after flipping out target cytosine to be methylated on the target strand. After this, DNMT1 and/or DNMT3B carries out hemi-methylation to acquire fully methylated DNA.

2.5.1. The working principles of DNMT3A/B enzymes

DNMT3A/Bs has three distinct interfaces: a DNA binding, a hydrophobic, and a hydrophilic interface (**Figure 4**)(Zhang et al., 2018). The DNA binding interface encompasses the catalytic center of the enzyme. The hydrophobic interface governs the efficiency of methylation activity either by enhancing the processivity of DNMT3s or by triggering the oligomerization of DNMT3s (Holz-Schietinger, Matje, & Reich, 2012), please see **Figure 4**. The DNMT3 can form higher-order oligomeric states through interacting by their hydrophobic and hydrophilic interfaces (Emperle, Rajavelu, Reinhardt, Jurkowska, & Jeltsch, 2014). As an example, the hydrophobic-interface-mediated interactions could yield to homo- or heterodimer DNMT3 complexes bound to a single methylation site (**Figure 4**). The hydrophilic interface is needed to form homo- or heterotetramer DNMT3 complexes (**Figure 4**). As structural studies have shown, heterodimers bind to two methylation sites separated by approximately 14 base pairs. The hydrophilic interface of DNMT3A also affect the sequence specificity by altering the affinity towards flanking region. Furthermore, it has been shown that DNMT3A enzyme further oligomerize on the naked DNA to methylate multiple sites (Emperle et al., 2014; Rajavelu, Jurkowska, Fritz, & Jeltsch, 2012).

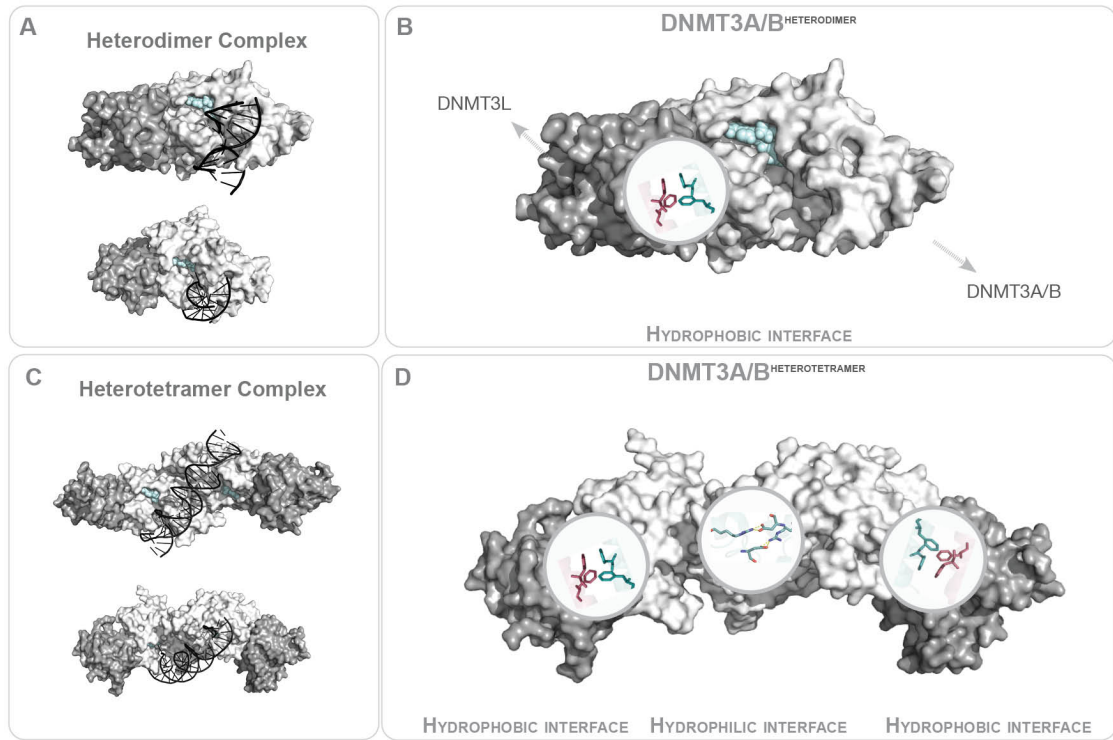


Figure 4: Indicative states of DNMT3A/B-mediated oligomerization. **A-B)** DNMT3A/B forms heterodimers through hydrophobic interface (FF-mediated interaction). **C-D)** DNMT3A/B:DNMT3L heterodimers further generates heterotetramers through hydrophilic interface (RD-mediated interactions).

DNMT3A/Bs impose DNA methylation by following similar steps as stated at the section 2.1. But they diverge in terms of the processivity and the cooperativity. DNMT3A oligomerize on unpacked, nucleosome-free DNA either by binding to homo-oligomer DNMT3As or by interacting with hetero-oligomer DNMT3A:DNMT3L complexes (C. C. Lin, Chen, Yang, Shen, & Yuan, 2020; Rajavelu et al., 2012; Veland et al., 2019). In this way, DNMT3A can methylate multiple methylation sites that are separated by 14 base pairs. During gametogenesis, on the other hand, the activity of DNMT3A solely depends on its oligomerization with DNMT3L. For DNMT3B, it also has been shown that hydrophobic-interface-mediated DNMT3B:DNMT3L complex formation impacts the methylation rate. Though, this is never as dramatic as in the case of DNMT3A. For instance, comparison of DNMT3A knockout and DNMT3L knockout studies revealed that DNMT3A is not enough to publish methylation pattern during gametogenesis. It has

been also showed that hydrophobic-interface- and hydrophilic-interface-mediated interactions contribute to the cooperativity/processivity of DNMT3A/B enzymes (Emperle et al., 2014; Holz-Schietinger et al., 2012; Rajavelu et al., 2012). It has been shown that hydrophobic-interface-mediated interactions enhance the processivity of DNMT3B enzyme while hydrophilic-interface- and hydrophobic-interface-mediated contacts guide oligomerization of DNMT3As which follow either distributive or processive mechanism in a cooperative manner (Holz-Schietinger et al., 2012).

2.5.2. The sequence specificity of DNMT3A/Bs towards distinct methylation sites

DNMT3A/B proteins impose a methyl group at CpG islands (5'-CG-3') and, to a lesser extent, at non-CpG sites, i.e. CpA, CpT and CpC sites (Guo et al., 2014; C. C. Lin et al., 2020; Oka, Rodic, Graddy, Chang, & Terada, 2006; Suetake, Miyazaki, Murakami, Takeshima, & Tajima, 2003; Zhang et al., 2018). The sequence specificities of DNMT3A/B enzymes towards distinct methylation sites are shaped by the flanking nucleotides surrounding the target sites. Multiple studies have compared the methylation activity of DNMT3A/Bs towards CpG sites with different flanking nucleotide contents (Handa & Jeltsch, 2005). For instance, a study run a comparative analysis on DNMT3A/B's sequence preference towards a group of NNNCGNNN (N: A, C, G, T) sequences (C. C. Lin et al., 2020). This study has revealed DNMT3A enzyme prefers pyrimidines right after the CpG site (NTNCG(C/T)NN), while DNMT3Bs predominantly targets CpG followed by a purine nucleotide (NTNCG(G/A)NN). Another study has revealed the impact of flanking region on the methylation of CpA sites. In this study, it has been shown that DNMT3A enzymes methylates preferentially CAC over CAG sites, while DNMT3B enzymes predominantly target CAG sequences.

Further knockout experiments have revealed that DNMT3A/B might also work interchangeably, since they might target the similar sequences. However, DNMT3A is crucial for genomic imprinting and methylation of major satellite regions, while DNMT3B is critical for methylation of minor satellite regions and embryonic development. It has been shown that aberrant mutations at the DNMT3A leads to hematological cancers (e.g. AML) (Holz-Schietinger et al., 2012; T. V. Nguyen et al., 2019). The mutations of DNMT3B might cause facial abnormalities (ICF) and

centromeric instability. Thereby, both DNMT3A/Bs are necessary to maintain healthy state of cells and organisms.

2.5.3. The state-of-the-art structural knowledge on DNMT3A/Bs

The first structure of DNMT3A bound to DNA, was resolved in 2018 (PDB ID: 6F57). In this study, they showed that DNMT3A involves a TRD loop (R831-F848) and a catalytic loop (G707-K721) (**Figure 5A**)(Zhang et al., 2018). In the DNA-free forms, TRD loop could not be solved due to its flexible nature. TRD is a major groove binding element, while the catalytic loop targets DNA's minor groove. Thanks to these interactions, DNMT3A mediates its DNA binding mode. A rather recent crystal structure has also revealed that DNMT3L does not make any contact with DNA however it enhances DNMT3A DNA binding (Jurkowska et al., 2008).

In the bound state, the target cytosine is found in its catalytically competent flipped state, where it is stabilized by its interactions with CYS710, SER714, GLU756, ARG790 and ARG792 (Zhang et al., 2018). Regarding the conserved motifs, as introduced in section 2.2, GLU756 and CYS719 belong to motif VI and motif IV, respectively, while ARG790 and ARG792 belongs to motif VIII. The OG atom of Ser714 interacts with OP1 atom of flipped cytosine. The NH1 atom of ARG790 and OE2 atom of GLU756 makes base-specific interaction with flipped cytosine. Finally, ARG792 makes multiple interactions with the flipped cytosine. ARG836 of DNMT3A contacts with guanine next to cytosine (orphan nucleotide) to methylated at the target strand. This interaction is further supported by the VAL716 and PHE718. On the other hand, initial analysis on the Xtal structure could not depict any interaction with the complementary strand. To that end, the authors have stated that this is why DNMT3A can target both hemi-methylated and non-methylated DNA. The study also showed that DNA binding is further supported by various residues of catalytic subdomain (ASN711, SER714, ILE715) and TRD (ARG831, THR832, THR835, ASN838 and LYS841) (**Figure 5**). Since these amino acids are predominantly conserved also in DNMT3B, they have stated that DNMT3A/B share similar working mechanisms. The mutation of R836A was shown that ARG836 is critical to discriminate CpG methylation site from non-CpG methylation site. On the other hand, VAL716ALA mutant revealed that VAL716 is essential for DNMT3A's methylation activity. Further mutational studies showed that VAL716GLU, ARG792HIS

and LYS841GLU both diminish methylation activity and also interfere with the DNMT3A WT activity. The tested mutations are common among the AML patients (Zhang et al., 2018).

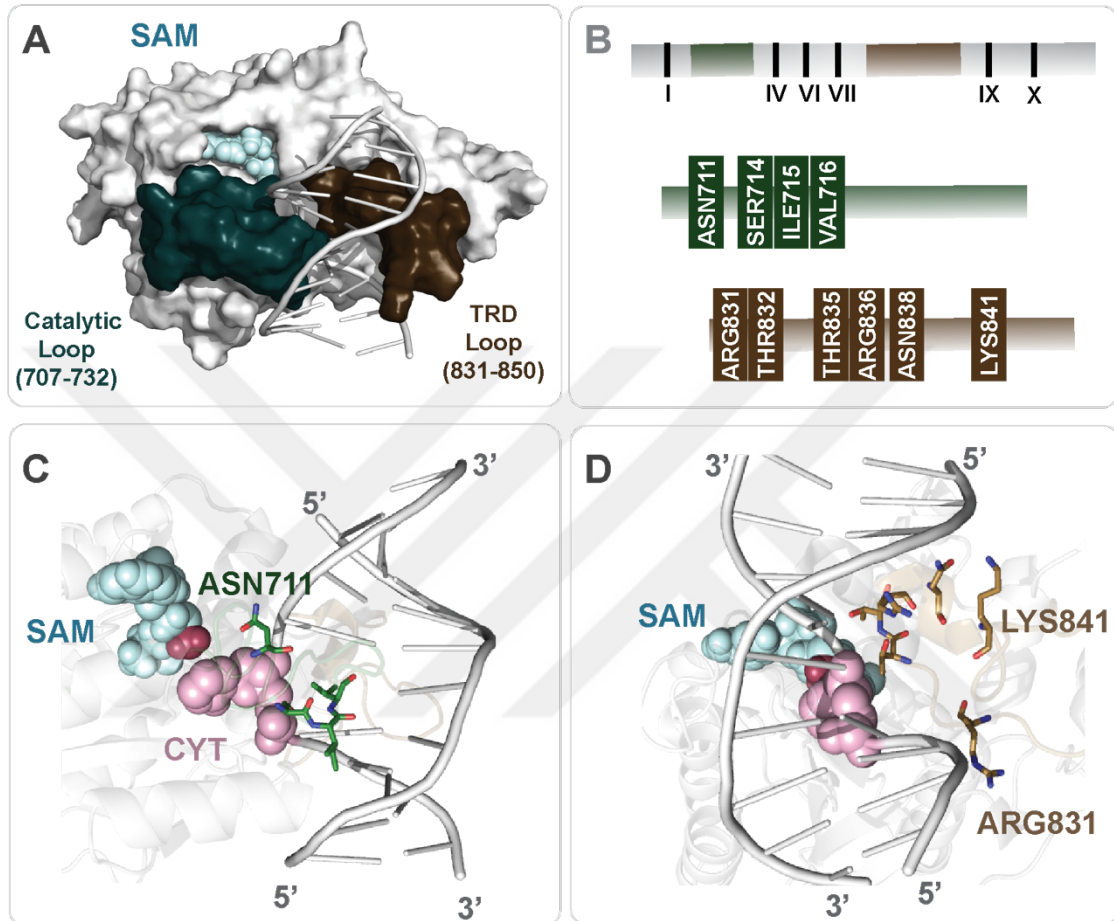


Figure 5: Critical amino acids contributing to working principles of DNMT3A. **A)** DNMT3A enzymes enclose DNA sequence through their catalytic and TRD loops. **B)** Critical amino acids contribute to the working principles of DNMT3A enzyme. **C)** Structural representation of catalytic loop amino acids contacting to DNA. **D)** Structural representation of TRD loop amino acids interacting DNA.

In 2020, the crystal structures of DNMT3A/B bound to different DNA sites have also been published (PDB IDs: 6U8V, 6UW8, 6U8X, 6U8P, 6U91, 6U90, 6W8J, 6W8B,

6W8D, 6W89, 6KDL, 6KDB, 6KDA, 6KDP, 6KDT) (Gao et al., 2020; C. C. Lin et al., 2020; Zhang et al., 2018).

The crystal structures of DNMT3B have showed that the TRD and catalytic loops enclose DNA binding region as seen in DNMT3A. The flipped cytosine is stabilized by CYS651 in motif IV, SER649, GLU697 in motif VI, ARG731 in motif VIII and ARG733 in motif VIII. It has been further suggested that the first time a water channel has been observed in the DNMT3 family bound to DNA. The water channel has been proposed to take a role in the final stage of methylation by deprotonating C5 atom of flipped cytosine (C. C. Lin et al., 2020).

Further mutational analysis revealed the potential role of residues on the activity and specificity of DNMT3B (Gao et al., 2020). Based on the structural data, VAL657 residue insert itself into the minor groove to fill the emptiness originating from flipping out. V657G and T775S mutant cause partially diminished methylation activity. DNMT3B uses ASN779 in the TRD loop to recognize O6 atom of orphan guanine. They have stated that this guanine is recognized by ARG836 in DNMT3A (Zhang et al., 2018). Further N779A, N779V, N779D and N779Q mutations have showed that N779 is essential to target CpG sites (C. C. Lin et al., 2020). They also showed that N779 interacts with flanking region, which mediates DNMT3B's flanking region preference. Another mutation was made on LYS777 corresponding to ARG836 in DNMT3A. The K777R mutant significantly increased CpG methylation while the K777A mutant showed slightly decreasing methylation towards the both CpG and non-CpG sites (C. C. Lin et al., 2020).

2.6. A computational perspective to study protein-DNA interaction specificities

Biomolecular recognition is an integral step of biological processes. It is driven by the conformational changes to enhance chemical and geometrical complementarity at the binding region. The DNA recognition is proposed to be carried out through shape and/or by base readout mechanisms. **The shape readout mechanism** (SRM) is an indirect recognition mechanism, as it depends on the structural variations of DNA topology (Abe et al., 2015; Golovenko et al., 2018). The structural diversity is identified by comparing the degree of backbone-dependent rotations of nucleotides. The DNA shape variations might be implied as global (e.g. DNA bending) or local (e.g. DNA kinks or minor groove narrowing).

The DNA bending is originated from the interrelated alteration of the angles among the adjacent nucleotides. This implies the formation of a curvature. The DNA-binding proteins, then, adapts their conformation to bind the defined curvature. Alternatively, the DNA-binding proteins force conformational adaptation of the defined curvature on the DNA. The DNA kinks is a small form of DNA bending since it occurs where stacking interactions between base-pairs are disrupted. The kinks are predominantly stabilized by the hydrophobic residues of proteins. The last example of shape variation is minor groove narrowing/widening. The altered hydrogen bonds between base-pairs and stacking interactions between adjacent nucleotides might trigger minor groove narrowing/widening. These alterations help DNA to be targeted by distinct protein families.

Base readout mechanism (BRM), which is a direct recognition mechanism, is the key factor in the recognition of DNA sequences specifically (leading to high affinity binding) (M. Lin & Guo, 2019). They are predominantly mediated by a number of contacts between nucleotides and amino acids. Interactions between the base region of a nucleotide and the side chain of amino acids are BRM's main contributor. A prominent way to define these interactions relies on the characterization of intermolecular contacts. A well-defined interaction network sheds light on the DNA sequence specificity of the DNA-binding enzymes.

The temporal evolution of inter- or intramolecular interaction network is fundamental to understand the BRM-driven biomolecules. Since computational approaches can enlighten the time-dependent behaviors of the biomolecular systems, they become indispensable to explore the DNA recognition mechanisms. One of the such prominent computational methodologies is molecular dynamics (MD) simulations. MD simulations aids to explore the conformational space of protein-DNA complexes.

The comprehensive analysis of the intra-/intermolecular interactions reveals the contacts crucial for functioning of the proteins. The persistently similar contacts across different MD simulation trajectories can be used to define the dominant interaction mode. The impact of a mutation or key differences of homolog systems might be explored through comparing investigated interaction modes. For instance, homolog proteins make use of conserved amino acids while carrying out their function. On the other hand, variant

amino acids should be the key directing the specificity of homologs. While dissecting the critical interactions, hydrogen bonds are the predominant tool to investigate the specificity of protein-DNA interactions (M. Lin & Guo, 2019). It also has been shown that electrostatic interactions, as well, contribute to the specificity and/or stability of system (Bardwell & Bardwell, 2015; Faltejskova, Jakubec, & Vondrasek, 2020; Ivanov, Cawley, Huber, Bond, & Warwicker, 2017; Sinha & Smith-Gill, 2002). Thereby, the definition of interactions should have been expanded as hydrogen, hydrophobic, and electrostatic bonds.



3. MATERIALS AND METHODS

3.1. Type of Study

This thesis involves a theoretical study, driven by high-end molecular simulations. During this work, we used web-based modeling tools, molecular visualization software, and molecular simulation packages.

3.2. Time and Location of Study

This study has been performed at the Izmir International Biomedicine and Genome Institute, Izmir Biomedicine and Genome Center, and University of Groningen during September 2018 and August 2020.

3.3. Population and Sample of Study

We explored the sequence specificity of mammalian *de novo* methyltransferases towards their preferential methylation sites.

3.4. Materials of Study

In this thesis, we used the Xtal structures of HhaI methyltransferase (PDB ID: 5MHT) and DNMT3A (PDB ID: 6F57) for modelling, simulations and analysis.

We run molecular dynamics simulations on the SURFsara, TRUBA and IBG-HPC via GROMACS 2018.1. 1 GPU and 40 cores of akya-cuda in TRUBA, 3 GPU of Cartesius in SURFsara and 1 GPU and 40 cores in IBG-HPC were utilized.

3.5. Variables of Study

5MHT or 6F57 were used as a template to model complexes i) 5MHT bound to non-methylated DNA, and SAH or SAM ii) DNMT3A/B bound to non-methylated DNA and hemi-methylated DNA, iii) DNMT3A/B bound to CpG, -CAC- or -CAG- sequences). SAH is the form of SAM giving away its methyl group. It was used to describe the unproductive state of complexes.

3.6. Tools for Data Collection

3.6.1. Web-servers & resources used for data collection

During our research, we used the below-listed web services, simulation and visualization packages.

UNIPROT is a freely accessible database offering extensive information on protein families, sequences, and the impact of mutations on the phenotype (UniProt, 2019).

RCSB PDB is the protein databank, storing all the available three-dimensional (3D) coordinates of biomolecular complexes (Burley et al., 2019).

NEEDLE is a pairwise alignment webserver. Needle performs a global sequence alignment using Needleman-Wunsch algorithm.

CONSURF is a web-based tool for calculating the evolutionary conservation of amino acids upon comparing the query sequences with the homologous ones. Consurf uses HMMER homolog searching algorithm to collect the homologous sequences. MAFFT method is used to align the collected sequences. By default, 150 number of sequences are used to measure conservation by keeping the threshold for sequence identity as between 35% and 95% (Ashkenazy et al., 2016).

PHYRE2 is a threading-based webserver used to predict the 3D structure of a protein. Threading assists for the template-based modeling of homolog proteins by taking the structure of a homolog (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015).

3DNA is a software to analyze and model the 3D structure of nucleic acids. It mutates available DNA structures by using the given the DNA topology as template. It further helps to analyze DNA shape dimensions, such as major groove width, minor groove width (Lu & Olson, 2003).

PYMOL is a molecular visualization tool. It was used to align DNMT3B onto the crystal structure of DNMT3A to model the DNMT3B:DNA complexes. Further SAM was superimposed onto the SAH molecule. Lastly, it was used to visualize the snapshots generated by the MD trajectories.

HADDOCK is a docking webserver, which involves multiple modules from rigid body docking to interface refinement. HADDOCK was used to refine all of the modelled

structures before molecular dynamic simulations (Dominguez, Boelens, & Bonvin, 2003).

GROMACS is an open-source molecular dynamics package, simulating time-dependent behavior of biological systems. It was used for the MD purposes.

MOLPROBITY is a webserver to validate the reliability and quality of modelled and crystal structures (Chen et al., 2010).

PROCHECK is another webserver to control reliability and quality of modelled and crystal structures.

dr_sasa calculates solvent accessible surface area (SASA) and buried surface area (BSA) for given coordinate file (in the form of pdb). dr_sasa supports the calculation of SASA and BSA regarding nucleic acids, proteins and ligands. It also gives residue-based BSA values (Ribeiro, Rios-Vera, Melo, & Schuller, 2019).

FREESASA calculates solvent accessible surface area of a given pdb file. It implements Lee & Richards' algorithm while allowing a choice of residue range (Mitternacht, 2016).

SURFsara is the Dutch High Performance Computing (HPC) center. We have access to SURFsara since I got accepted as a HPC-Europa3 visitor. It was used to run molecular dynamics simulations.

TRUBA is our national HPC center. It was used to run analyses of MD trajectories.

IBG HPC is the local HPC center of IBG. It was used for the simulation of mutated DNMT3B bound to -CAG- sequence.

R and Python packages were used to carry out analysis. The detailed information regarding the packages are provided under the Methodology section.

3.6.2. Methodology

3.6.2.1. Obtaining the experimentally determined structures of DNMT complexes

SAH-bound, DNA- and SAH-bound or non-bound forms of DNA methyltransferases were experimentally determined and deposited to RSCB PDB. During our study, we investigated following complexes:

- The complex of bacterial DNA methyltransferase bound to DNA and SAH (PDB ID: 5MHT),
- The complex of DNMT1 bound to DNA and SAH (PDB ID: 4DA4),
- The complex of DNMT3A:DNMT3L bound to DNA and SAH (PDB ID: 6F57). This complex will be referred as the “dimer”.

Recently more structures were resolved, however they could not be involved in our investigation due to the time constraints.

3.6.2.2. Pairwise Sequence Alignment of DNMT3A/B enzymes

Sequence alignment is crucial to reveal similar regions among nucleic acids or proteins. Homolog enzymes utilize the conserved regions to carry out their activity while variant regions might aid to explain the divergence of multiple paralog enzymes. Furthermore, sequence similarity at the amino acid level would help to choose a reference template for the construction of homology model in case there is no experimentally resolved structure. When the sequence similarity is higher than 80%, it is safe to say that they could be modeled by taking another one as a reference template

To understand whether DNMT3B could be modelled by taking the crystal structure of DNMT3A as a reference template (PDB ID: 6F57), we run pairwise alignment of DNMT3 enzymes through EMBOSS NEEDLE.

3.6.2.3. Construction of homology-based models

3.6.2.3.1. Homology Modelling

Homology modelling aids to construct 3D protein structures by taking homologous protein structure(s) as template. Homology modelling generates a model based on the assumption of that homolog proteins, carrying out similar function, should have similar folding domain. Phyre2 is one of the top modelling webservers allowing model construction by using single template (one-to-one threading protocol). While generating plausible homology models, Phyre2 builds HMMs (Hidden Markov Model) as depending on given template structure. Then, it aligns the ensemble of HMMs by applying HHsearch algorithm. Finally, it gives a single model.

In this study, we made use of the Xtal structure of monomer DNMT3A methyltransferase domain (PDB ID: 6F57) to model the structure of MTase domain of DNMT3B enzyme. The sequence of methyltransferase domain of DNMT3B enzyme were taken from the UNIPROT (Q9UBC3). The steric clashes and the geometrical rotation of DNMT3B model was controlled with Molprobability.

3.6.2.3.2. Modelling non-CpG Methylation sites

3DNA is a well-known software developed to build, analyze or visualize nucleic acids. 3DNA implements mutate_base to mutate a DNA sequence without changing the given DNA backbone conformation.

To model non-CpG sites, we used DNA in the Xtal structure of DNMT3A:CpG as a template (PDB ID: 6F57). To acquire -CAC- sequence, we converted GUA370 and CYT359 to the ADE370 and CYT359THY, accordingly. To obtain -CAG- sequence, we further made CYT371GUA and GUA358CYT mutations in the -CAC- sequence. At the final stage, we had crude models of three distinct DNA sequences: i) Xtal structure of CpG site, ii) -CAC- sequence and iii) -CAG- sequence.

3.6.2.3.3. Template-based modelling

After acquiring DNMT3B model and DNA sequences, we further constructed crude models of DNA methyltransferase enzymes bound to different methylation sites.

For modelling, we utilized superposition command of Pymol by taking Xtal structures of 5MHT:5CM-SAH and DNMT3A:CpG-SAH as a template (PDB IDs: 5MHT, 6F57). Since SAM is the natural substrate of DNA methyltransferases, we also replaced SAH in the template structures, if necessary, with the previously resolved SAM structure (PDB ID: 3AV6) through superimposition. Finally, we renumbered amino acids of HhaI and DNMT3A/B enzymes based on the multiple sequence alignments to confidently compare HhaI and mammalian DNMT3A/Bs enzymes. For the amino acid numbers, please see APPENDIX D.

In this way, we constructed multiple crude models of HhaI-, DNMT3A- and DNMT3B-bound to variant DNA sequences:



- i. **Building HhaI DNA methyltransferase models:** We used Xtal structure of HhaI complex (PDB ID: 5MHT) as template. 5MHT is composed of HhaI enzyme bound to hemi-methylated DNA (5CM) and SAH (5MHT:5CM-SAH). To obtain crude models of HhaI enzyme bound to non-methylated DNA (CYT), we deleted the methyl group of the methylated cytosine at the complementary strand of hemi-methylated DNA. After modeling, we acquired: a) MTase bound to hemi-methylated DNA and SAM (5MHT:5CM-SAM), b) MTase bound to non-methylated DNA and SAM (5MHT:CYT-SAM), c) MTase bound to non-methylated DNA and SAH (5MHT:CYT-SAH), d) Xtal structure of MTase bound to hemi-methylated DNA and SAH (5MHT:5CM-SAH). To keep track of DNA sequences and where methylated cytosine (5CM) and target cytosine (red C) are located, please check **Table 1**.
- ii. **Construction of DNMT3A/B Models bound to non-methylated or hemi-methylated CpG sites:** We used Xtal structure of DNMT3A bound to non-methylated CpG site and SAH, as template (DNMT3A:CYT-SAH). We built crude model of hemi-methylated DNA by changing name of CYT (base-pair of orphan nucleotide next to target cytosine) to 5CM at the complementary strand. At the end, we had one Xtal structure and three models: a) Xtal structure of DNMT3A bound to non-methylated CpG site (DNMT3A:CYT-SAH) b) DNMT3A bound to hemi-methylated DNA (DNMT3A:5CM-SAM), c) DNMT3B bound to hemi-methylated DNA (DNMT3B:5CM-SAM), d) DNMT3B bound to non-methylated DNA (DNMT3B:CYT-SAM). For DNA sequences, please see **Table 1**
- iii. **Construction of DNMT3A/B Models bound to non-CpG sites:** We used Xtal structure of DNMT3A:CpG-SAH (PDB ID: 6F57) as template. Crude models of DNA sequences, generated at the section 3.6.2.3.2, were superimposed onto the DNA of template. Next, DNMT3B model was also superimposed onto the DNMT3A of template. At the end, we obtained six crude models of i) DNMT3A:CpG-SAM, ii) DNMT3A:CAC-SAM, iii) DNMT3A:CAG-SAM, iv) DNMT3B:CpG-SAM, v) DNMT3B:CAC-SAM, vi) DNMT3B:CAG-SAM. For DNA sequences, please see **Table 1**.

Table 1: Schematic representation of DNA sequences utilized in modelling.

Symbol to indicate DNA sequence	Where DNA motif is located	DNA Motif
5CM	Target Strand	5'- G C GC -3'
	Complementary Strand	3'- CG 5 CMG -5'
CYT or CpG	Target Strand	5'- G C GC -3'
	Complementary Strand	3'- CGCG -5'
-CAC-	Target Strand	5'- G C AC -3'
	Complementary Strand	3'- CGTG -5'
-CAG-	Target Strand	5'- G C AG -3'
	Complementary Strand	3'- CGTC -5'

As a result, we obtained multi crude models, which we optimized with HADDOCK.

3.6.2.4. HADDOCK Protocol

HADDOCK is a molecular docking program to investigate the plausible binding modes of biomolecular systems. As different from common docking algorithms, it follows three distinct steps to carry out docking:

1. it0, rigid body energy minimization: The docking begins with the generation of randomly oriented structures by taking center of mass of given structures as reference point. Then, translational/rotational space is searched after grouping the

specified atoms together, as rigid bodies. The randomly oriented structures, composed of rigid bodies, are docked while their energy scores are calculated. Usually, 1000 number of possible conformations are generated in this step. These structures are ranked based on their HADDOCK scores.

2. *it1*, semi-flexible simulated annealing: The best 200 number of structures generated at the *it0* are further simulated by following i) rigid body search at high temperature, ii) rigid-body simulated annealing, iii) semi-flexible simulated annealing with flexible side chains at the binding interface, iv) semi-flexible simulated annealing with fully flexible atoms at the interface.

3. water, flexible refinement: The generated structures at the end of the *it1* stage are further solved within the explicit solvent (either TIP3P or DMSO is used). Finally, a short MD simulation is applied to optimize structures.

For the protein-DNA systems, this protocol has been extended by also taking DNA flexibility into account (van Dijk, van Dijk, Hsu, Boelens, & Bonvin, 2006). The intrinsic flexibility of DNA is dynamically altered due to the protein binding. To define the conformational space of DNA, global and local parameters are implemented. In the *it1* stage, while side chains of proteins are flexible, the backbone and bases of DNA are kept flexible.

3.6.2.4.1. Adjusted Docking Parameters to understand selectivity towards hemi-methylated and non-methylated DNA sequences

To improve crude models generated at the section 3.6.2.3.3 (except models generated at the part iii of section 3.6.2.3.3), we optimized the DNA binding interface through simulated annealing followed by short MD simulations. Since methylated cytosine is not defined in the HADDOCK webserver, we initially defined the topologies of methylated cytosine, SAM and SAH molecules in the local version of HADDOCK 2.2. PRODRG was used to acquire aforementioned topologies. PRODRG generates molecular topologies by using the coordinates and atom names defined in the given PDB file. We used PRODRG since it is fast and also used by the HADDOCK to define topologies of ligands. Before any processing, HADDOCK completes the missing atoms of the crude

models by using predefined topologies of residues/ligands. The further refinement HADDOCK parameters we used are given in **Table 2**.



Table 2: Adjusted parameters for the optimization of hemi-methylated and non-methylated CpG bound DNA methyltransferases.

	Definition of Parameter	Parameter	Value
DNA-RNA restraints	Use DNA/RNA restraints (DNA-rna_restraints.def in data/sequence)?	dnarest_on	true
Docking protocol	temperature for rigid body high temperature TAD	tadhigh_t	20000
	initial temperature for rigid body first TAD cooling step	tadinit1_t	2000
	final temperature after first cooling step	tadfinal1_t	500
	initial temperature for second TAD cooling step with flexible side-chain at the interface	tadinit2_t	2000
	final temperature after second cooling step	tadfinal2_t	50
	initial temperature for third TAD cooling step with fully flexible interface	tadinit3_t	2000
	final temperature after third cooling step	tadfinal3_t	50
	time step	timestep	0.002
	factor for timestep in TAD	tadfactor	4
	number of MD steps for rigid body high temperature TAD	initisteps	20000
	number of MD steps during first rigid body cooling stage	cool1_steps	20000
	number of MD steps during second cooling stage with flexible side-chains at interface	cool2_steps	1000

3.6.2.4.2 *Applied HADDOCK protocol while investigating the structural mechanism of de novo methylation*

The water refinement module of HADDOCK webserver was utilized to improve crude models generated at the part iii of 3.6.2.3.3. For the used parameters, please check **Table 3**.

Table 3: Adjusted parameters for the water refinement.

	Definition of Parameter	Parameter	Value
Distance restraints	Do you want to define center of mass restraints to enforce con-tact between the molecules?	cmrest	true
Docking protocol	Cross-dock all combinations in the ensembles of starting structures?	crossdock	false
	Randomize start-ing orientations?	randorien	false
	Perform initial rigid body minimisation?	rigidmini	false
	Allow transla-tion in rigid body minimisa-tion?	rigidtrans	false
	factor for timestep in TAD	tadfactor	4

3.6.2.5. *HADDOCK analysis*

HADDOCK utilize distinct evaluation metrics to compare the generated conformations. For each plausible conformation, HADDOCK score is calculated as a function of these metrics. HADDOCK score comprises van der Waals and electrostatic energy scores, and buried surface area (BSA, which is not used at the water stage). HADDOCK score is used to rank the generated conformations. Technically, the structure having the lowest HADDOCK score is accepted as the best possible binding mode. However, this might not be the case for every system. Because of this, we further evaluated the optimized conformations with respect to their binding mode and the

topology of DNA in combination with the HADDOCK scoring terms. For instance, flipped cytosine should be in close proximity with the conserved motifs (defined in the introduction) and the methyl group of SAM. While comparing the models, we used the HADDOCK and energy scores, and the BSA. The lower energy scores and HADDOCK score, and higher buried surface area are general sign of a better complex formation.

HADDOCK also calculates the contribution of each interface residue to the total energy of complex. This assists to understand the most important amino acids for the system. For the further evaluation, we used the “heatmaply” library in R to dissect the most critical amino acids of each complex and to distinguish the impact of methylation state on the bacterial DNA methyltransferases.

3.6.2.6. Quality control of models

After selecting the best structures, we used two tools (PROCHECK and Molprobit) to evaluate the stereochemical quality of models. PROCHECK assess structures based on the conformation of amino acids. It gives statistics of amino acids as depending on the Ramachandran plot by calculating torsional angles of amino acids. The higher percentage of amino acids in the most favored regions indicates the greater reliability of model. Molprobit was further used for more detailed evaluations since it also supports assessment of model quality as residue-wise manner.

3.6.2.7. Molecular Dynamics Simulation

Molecular dynamics (MD) simulation is a computational method utilized to investigate the dynamics of biomolecular complexes. The dynamics defines the transition from current state to next state by solving Newton’s equation of motion within a defined time step. The iterative solvation of Newton’s equation of motion generates a trajectory composed of multiple states.

MD requires an initial state and parameters defining the given system. The initial state might be a crystal structure or a homology model. The parameters are defined in the molecular mechanics force fields. Then, the state is put into a box which is filled with an explicit solvent. As a next step, ions are added to system to mimic the cellular condition. Thermodynamical equilibrium following energy minimization is carried out.

The energy minimization assists to optimize the initial state while searching for the state with global minimum. Steepest descent and/or conjugate gradient algorithms might be applied for the energy minimization. Then, the system is gradually heated. After equilibration of pressure, production run is started.

3.6.2.7.1. Applied Molecular Dynamics Protocol

To dissect the dynamic behavior of DNMT3A/B:DNA, we run MD simulations with Gromacs 2018.1 on SURFsara, Dutch high-performance computing center. The simulation temperature and pressure values were adjusted as 298 K and 1 atm, respectively. Long range electrostatic interactions were treated by using PME. The simulations were performed under the effect of the parmbsc1 force field (Ivani et al., 2016).

A dodecahedron box with a 1.4 nm diameter was used to solvate system. 0.15 M KCl ions were added to neutralize system. After minimizing water molecules with restraints on solute, entire system was minimized. Next, whole system was heated to 298K throughout 100 ps while applying harmonic restraints of 25 kcal/mol/ Å on the solute atoms via V-rescale algorithm. Then, the restraints were relaxed from 5 to 1 kcal/mol/ Å through series of 5 segments of 1000 steps energy minimization. Next, the system was further equilibrated at constant T (298K) and P (1 atm) for 50 ps. During equilibration, harmonic restraints of 0.5 kcal/mol/ Å were applied to the system. Right before production run, 50 ps unrestrained equilibrium were carried out at constant T (298K) and P (1 atm). Production run was carried out for 300 ns. 1500 were extracted conformers per complex.

Same protocol was followed to simulate the mutated DNMT3B:DNA complexes in TÜBİTAK ULAKBİM High Performance and Grid Computing Center (<https://www.truba.gov.tr/index.php/en/main-page/>) and IBG-HPC.

3.6.2.7.2. Assessment of the convergence of MD simulations

The quality control of MD trajectories should be checked before running any analysis. For this aim, we calculated Root Mean Square Deviation (RMSD) and Radius of Gyration (Rg) values throughout simulations.

- **Root Mean Square Deviation (RMSD)** calculates the average atomic displacement after superimposing onto the given reference state. We measured RMSD with the

GROMACS *rms* tool by giving initial state as a reference point. The equation 1 is for calculating RMSD.

$$\text{RMSD}(t) = \left[\frac{1}{M} \sum_{i=1}^N m_i |\mathbf{r}_i(t) - \mathbf{r}_i^{\text{ref}}|^2 \right]^{1/2} \quad (1)$$

- **Radius of gyration** (Rg) is measured with the GROMACS gyrate tool. It calculates Rg with the following equation 2 based on the given reference state. We used initial state as a reference point

$$R_g = \left(\frac{\sum_i |\mathbf{r}_i|^2 m_i}{\sum_i m_i} \right)^{1/2} \quad (2)$$

The plots of RMSD and Rg were created by using in-house python scripts. Based on the time to reach the plateau, we decided the equilibration time range. First halves of simulations (150 ns) was discarded. Rest were divided into the 3000 number of frames for further evaluations.

3.6.2.8. *Comprehensive interaction network analysis of MD trajectories*

MD simulations produce extremely valuable information on functionally critical interactions. Dissecting crucial contacts relies on the characterization of intermolecular interactions that are persistently similar or different across different simulations. To that end, we calculated intermolecular hydrogen and hydrophobic bonds. A **hydrogen bond** (HB) requires a donor (HBD) with partially positive charge and an acceptor (HBA) with partially negative charge. Atomic spatial orientations are vital to identify the character of

a hydrogen bond. A hydrogen bond becomes stronger when distance between HBD and HBA is lower, and/or degree between HBD and HBA becomes close to 180°. HBs might be further characterized as simple and complex. Simple HB occurs between a single HBD and a single HBA. Complex HBs occurs either between multiple HBA and single HBD. **Hydrophobic bonds** are intrinsic to protein folding, and protein-protein interactions in the certain cases. The side chains of hydrophobic amino acids (ALA, GLY, VAL, LEU, ILE, MET, PRO, PHE and TRP) shield themselves from water. This creates a hydrophobic core inside the protein.

Since there is no other tool to compare the inter-monomer bonding profiles of paralogous proteins, we developed the python package, minnie. minnie takes an ensemble of complex coordinates (in .pdb format) as an input. The ensemble coordinates should be separated with an ENDMDL flag. To compare the simulations of different complexes, residue numbering of complex monomers should correspond to each other. Within minnie, the input ensemble is splitted into single snapshots with splitpdb function. After the ensemble is splitted, findbonds analyzes each snapshot in parallel. At this step, minnie calculates various interaction profiles through with the interfacea python package. The calculated interaction profiles include inter-/intra-molecular hydrogen bonds, ionic, hydrophobic and ring stacking interactions. A hydrogen bonding information, for example, contains the donor and acceptor atoms/residues, the label whether it is an inter/intra-monomer contact, and whether it is formed between protein-protein or protein-DNA etc. Afterwards, the “insignificant” interactions are eliminated with the timefilter tool, which keeps the interactions, if they are observed more than the user-defined frequency filter (by default, 25%). The complex-specific (unique to each trajectory) and common (observed in both trajectories) interaction profiles are documented with the comparebonds tool. Here, comparebonds lists the common and distinct interactions are detected by comparing a unique pairwise interaction list composed of chainIDs, residue names and residue numbers. At this stage, the hydrogen bonds, e.g. bidentate and bifurcated ones, are counted as a single one. Finally, the common and complex-specific interaction distributions are plotted by using box-and-whisker plots with the corresponding histograms. These interactions can be visualized with the PyMol script, provided within pipeline.sh, together with all the command sequences described up to

here. The package flowchart and the folder organizations are provided as Figure 6. The source code of minnie is given in <https://github.com/CSB-KaracaLab/minnie> .



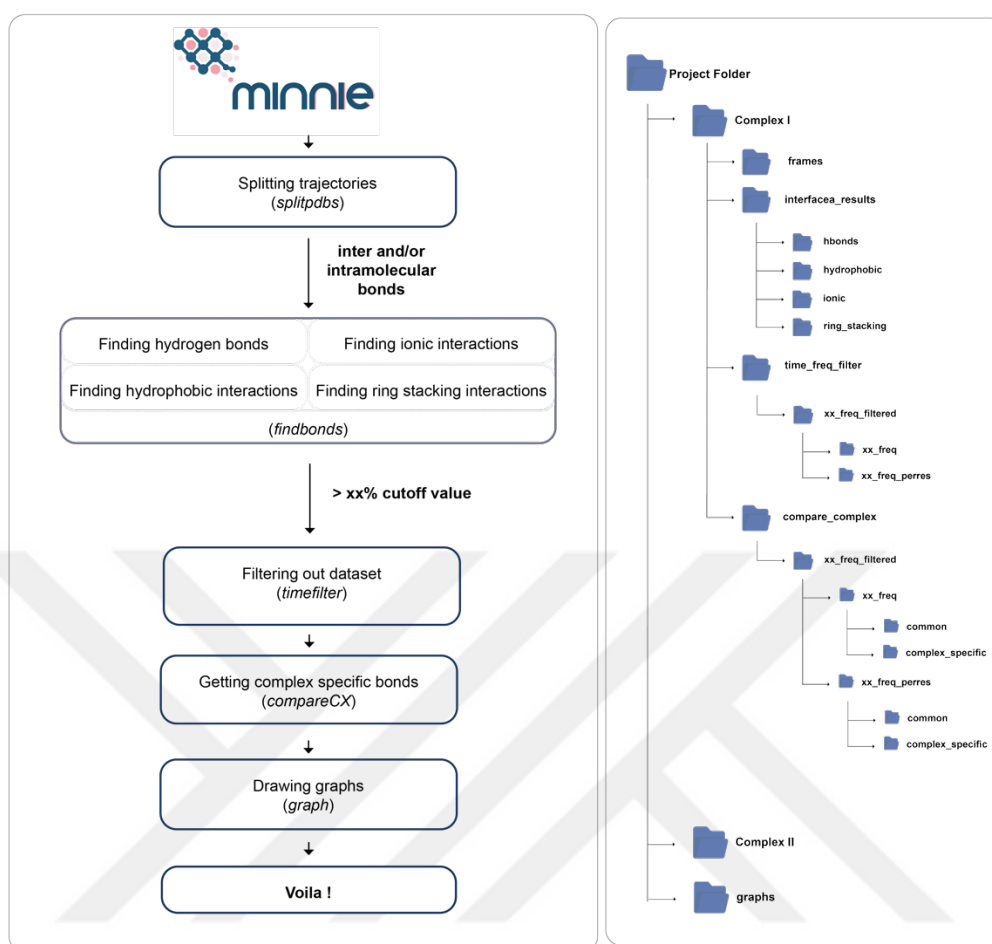


Figure 6. The workflow of minnie and the schematic representation of folder organization. Inside each project folder, two subfolders are created for each complex (Complex I and Complex II). Splitting trajectories are saved into “frames” subfolder. The interaction profile lists are written under “interfacea_results” into “hbonds”, “hydrophobic”, “ionic” and “ring_stacking” subfolders. “time_freq_filter” folder contains “xx_freq_filtered” subfolder, which has the interaction profiles filtered according to the user-supplied frequency cut-off, “is created. Interaction summaries are listed under “xx_freq” while the whole lists are saved within “xx_freq_perres” subfolder. After comparing the interaction profiles between complexes, “common” and “complex_specific” are saved in the corresponding subfolders. The interaction distribution graphs are saved into “graph” subfolder.

3.6.2.9. *Dissecting crucial intermolecular contacts*

To understand which intermolecular contacts are essential, we further considered intermolecular contacts according to their:

- **Consistency of interactions:** To eliminate the role of randomness, we run two simulations having different initial velocity. We considered an interaction to be important, if it persistently occurs for more than 25% the time.
- **Complex specificity of interactions:** To dissect how DNMT3A/Bs target their sequences, we made a systematic comparison. Since experimental data predominantly evaluates the sequence specificity of each enzyme towards different methylation sites, rather than direct comparison of DNMT3A/B enzymes, we compared DNMT3A complex with each other (CpG, -CAC- and -CAG-) while comparing DNMT3B:DNA complexes with each other. If an interaction only occurs in one complex, then this interaction is labelled as complex-specific interaction.
- **Base specificity of hydrogen bonds** is another factor enhancing specificity of enzymes. An interaction between base of a nucleotide and side chain of an amino acid defines the sequence specificity of DNA-binding enzymes.
- **Major groove vs minor groove interactions:** The interaction network is further categorized as depending on the type of contacted DNA groove (major and minor grooves). The major groove is almost twice as wider and relatively deeper (8.5 Å vs 7.5 Å) compared to minor groove. Interactions through the major groove further contribute to protein's selectivity. Because the type of base pairs might be differentiated by the unique pattern of hydrogen bond donor and acceptors of base pairs. Thereby, major groove mediated interactions are primary reason of protein's selectivity/specificity. The minor groove mediated interactions are shown to mediate specificity predominantly through water-mediated interactions (B. Nguyen, Neidle, & Wilson, 2009). In addition to hydrogen bonds, hydrophobic, electrostatic and stacking interactions also contribute to enzyme's activity and selectivity.

3.6.2.10. *Metrics to evaluate conformational dynamics of DNMT3A/B:DNAs*

- **Complex surface area (CSA):** As its name implies, CSA reveals the total surface area of biomolecular complexes. The tighter complexes have lower complex surface area.

- **Distribution of number of contacts:** To explore stability of complexes, we looked for the distribution of total number of intermolecular hydrogen bonds since consistently greater number of intermolecular contacts implies greater stability. To investigate sequence specificity of DNMT3A/B:DNA towards distinct methylation sites, we explored the distribution of complex-specific intermolecular contacts per time frame. A greater number of complex-specific contacts implies higher specificity.
- **Solvent accessible surface area:** In line with the BSA, solvent accessible surface area (SASA) is utilized to understand the flexibility of biomolecular complexes. As different from the BSA, it measures the total interface area exposed to solvent. It is also an indirect method to evaluate the role of water on the working principles of biomolecular complexes.

dr_sasa was used to calculate SASA and complex surface of DNMT3A/B:DNA complexes.

3.6.2.11. *Signal Propagation*

To analyze the residue-based communication efficiency, we calculated signal propagation as a function of inter-residue displacement (Chennubhotla & Bahar, 2007; Morra, Verkhivker, & Colombo, 2009). To that end, we calculated the fluctuations of C α -C α distance throughout simulations. The less signal intensity implies greater communication efficiency.

3.7. Study Plan and Calendar

Investigating the structural mechanism of *de novo* and maintenance methylation

Construction of <i>HhaI</i> and DNMT3A/B Models	February 2018 - March 2018
Optimization of Constructed Models	March 2018
Analysis of Optimized Models	April 2018 - June 2018

Investigating the structural mechanism of *de novo* methylation

Modelling <i>de novo</i> DNMT3A/B:DNA complexes	February 2018 - March 2019
Interface Refinement of Models	February 2018 - March 2019
Writing HPC-Europa3 Proposal	May 2018 - September 2018
Optimization of MD setups and running MD Simulations	May 2018 - February 2019
Analysis of MD trajectories	May 2018 - February 2019
Writing TUBITAK1002 Project	March 2019 - September 2019
Preparing Minnie Package	September 2019 - May 2020
Further analysis of MD trajectories	May 2020 - July 2020
Writing thesis	July 2020 - September 2020

3.8. Data Evaluation

We have applied following steps for quality controls:

- Xtal structures (PDB IDs: 5MHT and 6F57) were evaluated as depending on their resolution.
- Homology models were assessed with PROCHECK and Molprobit.
- The convergences of MD trajectories were controlled with RMSD and Rg.

3.9. Limitations of Study

Since we had limited computational source, we did not investigate rest of oligomerization states.

3.10. Ethic Committe Approval

Ethic committe approval was not requested since we only used computational tools.

4. RESULTS

4.1. Building Homology Models of bacterial *HhaI* methyltransferases

Mammalian DNA methylation is a complex molecular mechanism. Its regulation strictly depends on the type of DNA methyltransferase, the sequence of the methylation sites, and the methylation state of DNA. As a result, each DNA methyltransferase has distinct preferences towards different methylation sites and states.

To establish a common ground on fundamental components of DNA methylation, we have initially explored the working principles of a simple C5-DNA DNA methyltransferase system, bacterial *HhaI* methyltransferase. To that end, we have modelled *HhaI* binding to non-methylated DNA (5MHT:CYT-SAH), based on the available *HhaI* methyltransferase structure, containing a hemi-methylated DNA and SAH (5MHT:5CM-SAH). To obtain a realistic structural state, we have modelled i) *HhaI* bound to hemi-methylated DNA and SAM (5MHT:5CM-SAM), ii) *HhaI* bound to non-methylated DNA and SAM (5MHT:CYT-SAM).

4.2. Investigation of the working principles of bacterial *HhaI* methyltransferases

HhaI enzymes show similar methylation rates for non-methylated and hemi-methylated DNA, while displaying differential DNA binding affinity towards non-methylated and hemi-methylated DNA. To understand the fundamental factors, distinguishing *de novo* and maintenance *HhaI* C5-DNA methylation, we have run short HADDOCK water refinement on 5MHT:CYT-SAH, 5MHT:5CM-SAH, 5MHT:5CM-SAM, and 5MHT:CYT-SAM complexes. This has allowed us to compare the HADDOCK energies, as well as intermolecular hydrogen bond networks of different *HhaI* MTase states.

4.2.1. Interface comparison of different *HhaI* MTase states

After running a short water refinement on different *HhaI* MTase states, we have compared their HADDOCK energies, the individual scores that make up HADDOCK energies, and BSA values (**Table 4**). The statistics are calculated on the top four scoring HADDOCK-refined complexes. According to this interface analysis, 5MHT:CYT-SAH has slightly higher energy scores and BSA compared to 5MHT:CYT-SAM. 5MHT:5CM-SAH, on the other hand, has slightly lower HADDOCK energy, van der Waals and electrostatic energy scores, and lower BSA compared to 5MHT:5CM-SAM. This implies that ligand type has no significant impact on any of models. On the other hand,

5MHT:CYT-SAM has lower electrostatic energy score and BSA compared to 5MHT:5CM-SAM. Electrostatic energy is mainly driven by the intermolecular hydrogen bonds. Since both BSA and electrostatic energy score drops in the presence of methylated cytosine, considering additional methyl group on the cytosine, it is safe to assume that methylated cytosine significantly alters intermolecular interaction network. Thereby, we could utilize electrostatic energy score in the light of interaction network to assess the sequence specificity of DNA methyltransferases as long as HADDOCK confidently generates model-based conformations.

Table 4: Summary of HADDOCK-based scores for each HhaI models.

Models	5MHT:CYT-SAM	5MHT:CYT-SAH	5MHT:5CM-SAM	5MHT:5CM-SAH
Haddock Score (a.u.)	-272	-252	-250	-242
Total energy (kcal/mol)	-937.68 +/-65.85	-911.74 +/-70.65	-812.55 +/- 71.86	-806.04 +/-77.12
van der Waals (kcal/mol)	-130.07 +/-7.81	-118.49 +/-8.23	-125.73 +/- 9.35	-117.44 +/-10.75
Electrostatics (kcal/mol)	-807.61 +/-60.92	-793.24 +/-64.88	-686.81 +/- 65.64	-688.60 +/-69.03
BSA (Å ²)	2966.95 +/-105.75	2954.68 +/-127.89	2892.18 +/- 142.66	2899.9 +/-160.77

4.2.2.Determining the crucial amino acids mediating *HhaI* MTase specificity enzyme

To reveal the crucial interface residues, which contribute to the interface energetics the most, we have gathered per-residue van der Waals and electrostatic scores and calculated their sum, Etot. As a result, we have obtained six amino acids, which contribute to the protein-DNA interactions the most: LYS102, ARG185, ARG293, ARG229, ARG260 and ARG248 (**Figure 7**). Among these, i) LYS102 inserts itself into the minor groove for DNA geometry stabilization, ii) ARG229 and LYS248 interacts with the DNA backbone of the complementary strand through the major groove, iii) ARG248 and ARG260 recognize DNA target strand via major groove, and i) ARG185 is in close proximity to the flipped cytosine. The comparative energy analysis also has shown that

ARG185 might aid to differentiate SAM-bound complexes from SAH-bound ones. While ARG248 is essential to recognize hemi-methylated DNA, LYS102 plays a vital role to stabilize the productive state.

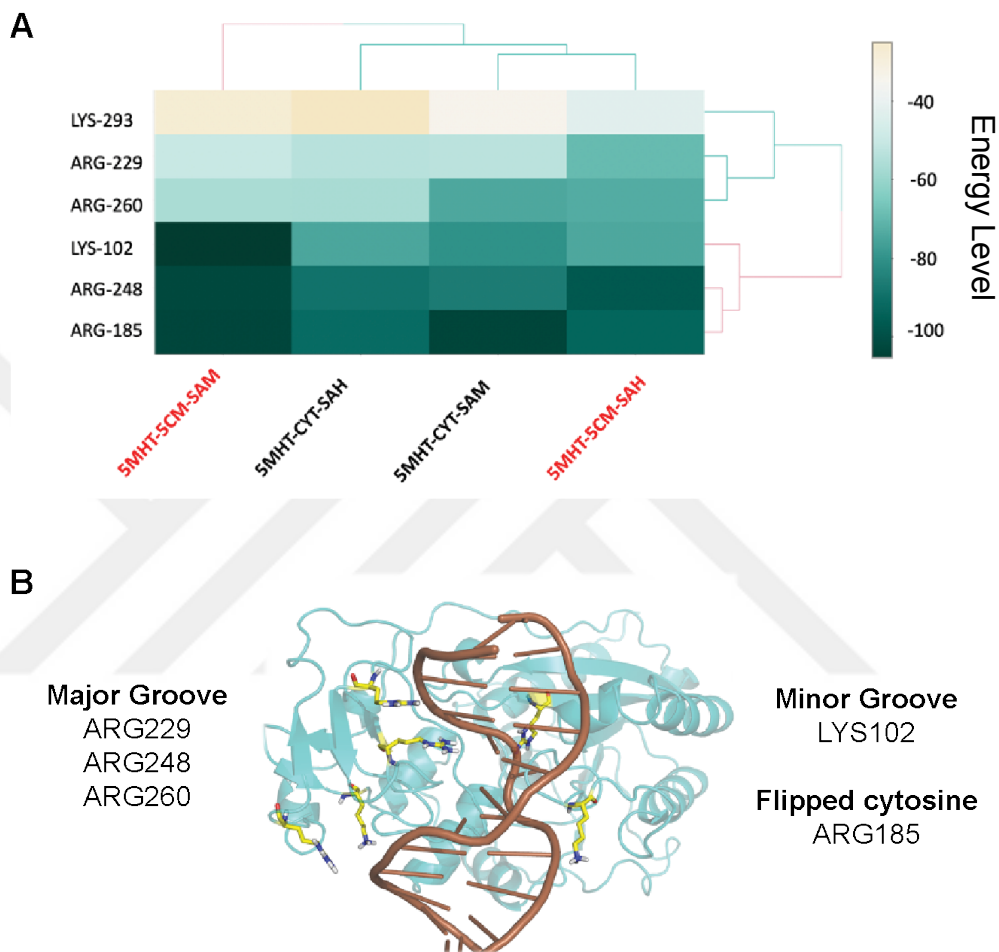


Figure 7: Dissecting critical amino acids for the sequence specificity of HhaI enzyme. **A)** Energy contribution of each residue to the total energy of system was revealed that five amino acids are essential to discriminate hemi-methylated DNA from non-methylated DNA. **B)** The crucial amino acids were mapped onto the 3D structure of HhaI:DNA system.

We have further explored the role of these energetically important amino acid via interaction network analysis. For this, we have compared the contact maps of the 5MHT:CYT-SAM and the 5MHT:5CM-SAM. In this analysis, we have used the top ranking HADDOCK-refined structures. The overlaid contact maps are given in Figure

8. This comparison has revealed that LYS89, ARG97, ARG165, ARG228, LYS234 and ARG240 might take an essential role to recognize the DNA motif and/or to stabilize the complex. In addition to this, another amino acid, GLN237 might be important for both DNA recognition and the activity of enzyme. It has been previously shown that the target recognition loop and catalytic loops of maintenance methyltransferases contact with each other by inserting themselves into the DNA grooves from both side of DNA. Accordingly, in the case of hemi-methylated bacterial system, GLN257 insert itself into the major groove, while SER99 approaches to DNA through the minor groove. The interaction between GLN257 and SER99 fill the emptiness originated from the flipped cytosine. Previously it has also been shown that the mammalian maintenance methyltransferases use a similar mechanism, which differs them from *de novo* methyltransferases. As another important point, GLN257 recognizes the second GUA of 5'-GmCGC-3' motif at the complementary strand, while ARG260 contacts to the first GUA of 5'-GCGC-3' motif at the target strand. So, the recognition of both strands might be essential in regulating the function of *HhaI* enzyme.

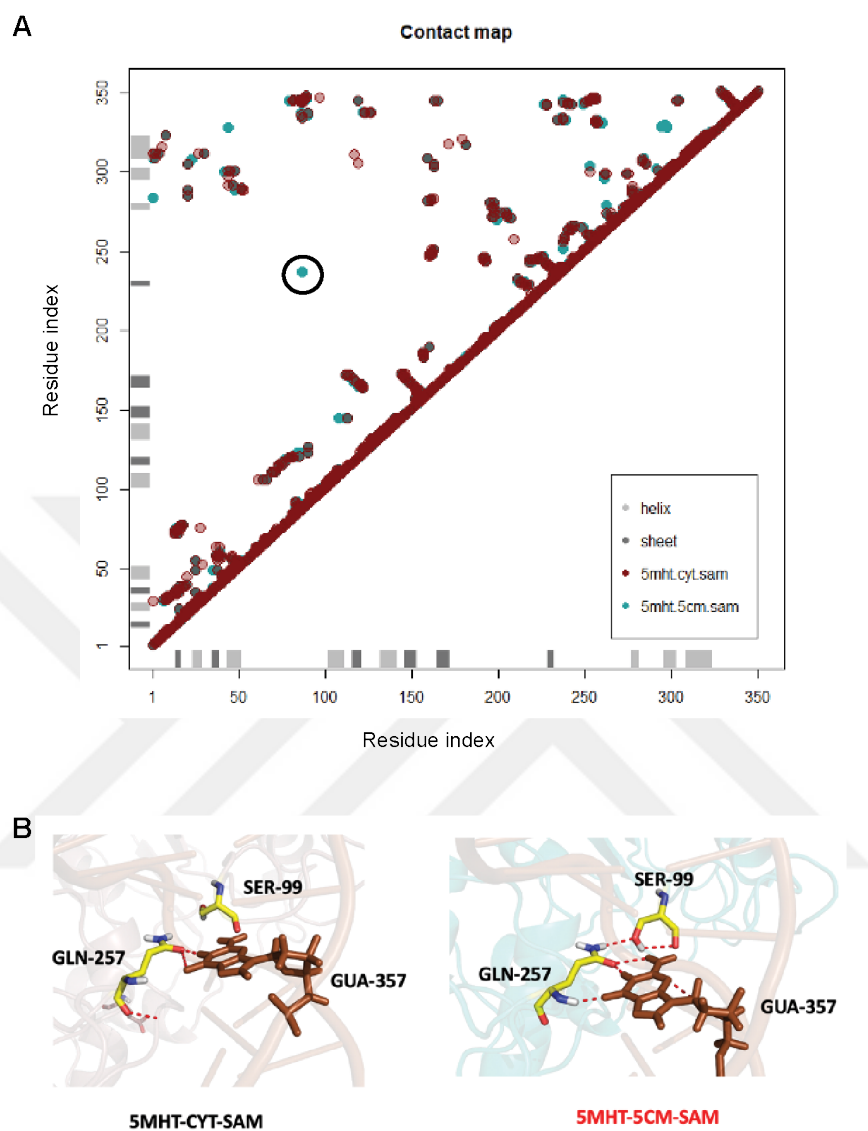


Figure 8: Comparison of contact maps to discriminate the sequence specificity of HhaI enzyme towards hemi-methylated (5CM) and non-methylated (CYT) DNA. Among variant interactions, the encircled contact occurs between SER99 and GLN257.

4.2.3. Linking *HhaI* information with the mammalian *de novo* methyltransferases

The critical amino acids of *HhaI* enzyme and their corresponding mammalian amino acids in DNMT3s are listed in **Table 5**. They will be further used to enlighten the working principles of DNMT3s.

Table 5: The list of critical amino acids dissected for the *HhaI* enzyme (5MHT) and their corresponding amino acids in DNMT3A/B. (?) indicates that corresponding residue is not clear at the structural level.

5MHT	DNMT3A	DNMT3B
SER87	ILE98	ASN98
LYS89	ASN100	ASN100
ARG97	THR110?	THR110?
ARG165	ARG185	ARG185
ARG228	LYS275	LYS275
LYS234	GLN281	GLN281
GLN237	ARG270/ASN272?	ARG270/ASN272?
ARG240	LYS278	LYS278

4.3. Exploring the working principles of mammalian *de novo* methyltransferases

As introduced earlier, DNMT3's show differential specificity towards CpG and non-CpG sites, which is further shaped by the flanking sequence. Additionally, DNMT3B also takes role in the maintenance methylation as similar to *HhaI* enzyme. To this end, we have initially modeled the DNMT3B enzyme by taking the crystal structure of DNMT3A as template. Next, to gain the mechanistic insight into *de novo* and maintenance methylation mechanisms of monomer DNMT3A/Bs, we have modeled and analyzed the interfaces of four different DNMT3 states (as we have performed for the bacterial MTase): i) DNMT3A:CYT-SAM, ii) DNMT3A:5CM-SAM, iii) DNMT3B:CYT-SAM and iv) DNMT3B:5CM-SAM. Next, we have focused on dissecting the sequence specificity and molecular dynamics of DNMT3A/Bs upon using i) DNMT3A:CpG-SAM structure, ii) DNMT3A:CAC-SAM model, iii) DNMT3A:CAG-SAM model, iv) DNMT3B:CpG-SAM model, v) DNMT3B:CAC-SAM model, and vi) DNMT3B:CAG-SAM model. By the end of this, we have also mutated two plausibly critical amino acids of DNMT3B to their corresponding amino acids in DNMT3A, to see whether we can turn one enzyme into another.

4.3.1. Template-based modelling the DNMT3B

To elucidate the working principles of DNMT3B, we initially modeled DNMT3B since its Xtal structure was not resolved at the time of the beginning of this project. To control whether the Xtal structure of DNMT3A (PDB ID: 6F57) is ideal for being a reference template, we run a pairwise alignment of sequences of DNMT3A/B's MTase domains with EMBOS NEEDLE tool (**Figure 9**). Since they share more than 80% sequence identity and they are also functionally relevant proteins, we have decided to model DNMT3B by taking the Xtal structure of DNMT3A as template. For the modelling, Phyre2 webserver was used. Phyre2 built the DNMT3B model by only using the Xtal structure of DNMT3A (PDB ID: 6F57) as a template.

At the last stage, PROCHECK was used to evaluate the quality assessment of DNMT3B model. It is an ingrained tool, evaluating the stereochemical aspects of amino acids. The report of PROCHECK supports the reliability of DNMT3B model since 90% of DNMT3B residues are at their most favorable conformations. Thereby, we decided to use the DNMT3B model in the next steps.

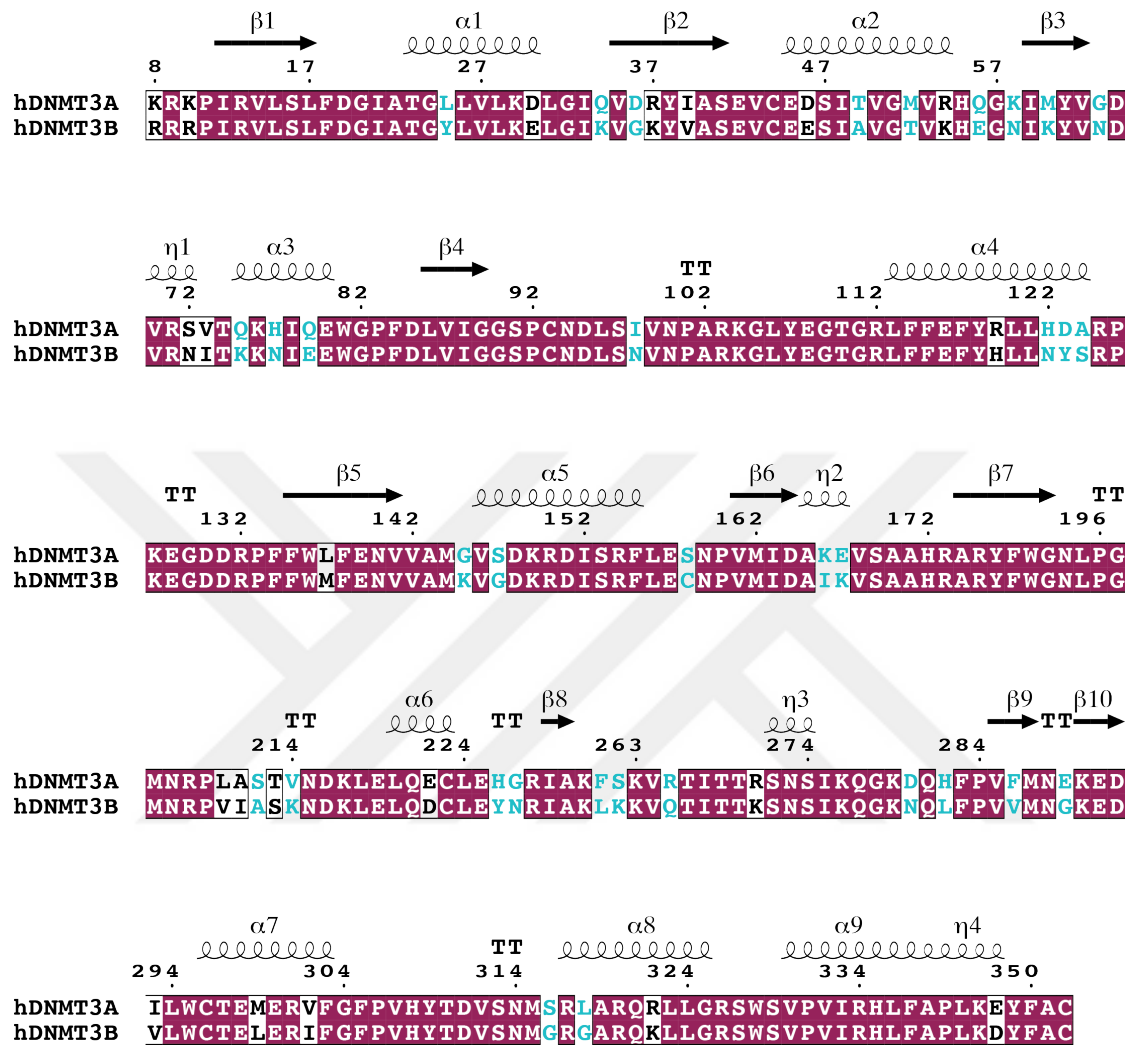


Figure 9: Pairwise alignment of methyltransferase domains of human DNMT3A and human DNMT3B. Coloring scheme follow the conservation and physicochemical aspects of amino acids. Conserved residues are colored as white with magenta background. Altered residues with similar physicochemical features are colored in bold black with the white background. Blue residues are not conserved.

4.3.2. Investigation of the working principles of DNMT3A/Bs towards hemi-methylated and non-methylated DNA

DNMT3A/B enzymes are primary factors governing mammalian *de novo* methylation. However, DNMT3B takes role in the maintenance methylation as similar to HhaI MTase enzyme. On the other hand, the regulation of DNMT3A/Bs selectivity towards hemi-methylated and non-methylated DNA is yet to be explored. To gain the mechanistic insight into *de novo* and maintenance methylation mechanisms of monomer DNMT3A/Bs, we have modeled productive states of four different cases: i) DNMT3A:CYT-SAM, ii) DNMT3A:5CM-SAM, iii) DNMT3B:CYT-SAM and iv) DNMT3B:5CM-SAM. After modeling, we have used HADDOCK for a short interface refinement as presented for HhaI MTase enzyme, please see **Table 6**. This analysis has revealed that DNMT3B shows no significant selectivity towards any of the sequences, while DNMT3A prefers non-methylated DNA over hemi-methylated DNA. DNMT3A:CYT-SAM shows the greatest electrostatic stability while DNMT3A:5CM-SAM is the least stable complex. To dissect the differences of DNMT3A/B, we have decided to use a more sophisticated tool, molecular dynamics.

Table 6: Summary of DNA-binding Interface Statistics

Models	DNMT3A:CYT	DNMT3A:5CM	DNMT3B:CYT	DNMT3B:5CM
Total energy (kcal/mol)	-702.17 +/- 41.98	-603.42 +/- 79.73	-645.02 +/- 87.82	-644.23 +/- 77.45
van der Waals (kcal/mol)	-76.96 +/- 6.91	-79.47 +/- 13.64	-81.21 +/- 9.23	-74.22 +/- 9.47
Electrostatics (kcal/mol)	-625.21 +/- 39.87	-523.94 +/- 70.55	-563.81 +/- 82.36	-570.01 +/- 71.73
BSA (Å ²)	2562.46 +/- 99.57	2461.22 +/- 254.67	2333.37 +/- 163.70	2351.11 +/- 154.36

4.3.3. Constructing Homology Models of DNMT3A/B:DNA

As different from DNMT1, DNMT3A/Bs methylate both CpG and non-CpG sites during *de novo* methylation. Their sequence specificity is also altered with the effect of flanking region surrounding methylation sites. For instance, DNMT3A prefers -CAC- sequence over -CAG- sequence while DNMT3B predominantly methylates -CAG- sequence. To understand DNA sequence preference of DNMT3A/B's enzymes towards non-CpG sites, we imposed *in silico* mutations to orphan guanine and/or nucleotide adjacent to orphan guanine (towards the 3' end of the forward strand). As a result of this, we obtained -CAC- and -CAG- sequences from the original -CGC- sequence. The nucleotides at the complementary strand were mutated accordingly. At the final stage, we had six different models to be investigated: i) DNMT3A:CpG-SAM, ii) DNMT3A:CAC-SAM, iii) DNMT3A:CAG-SAM, iv) DNMT3B:CpG-SAM, v) DNMT3B:CAC-SAM, and vi) DNMT3B:CAG-SAM. For the simplicity, “-SAM” label will not be used in the next sections.

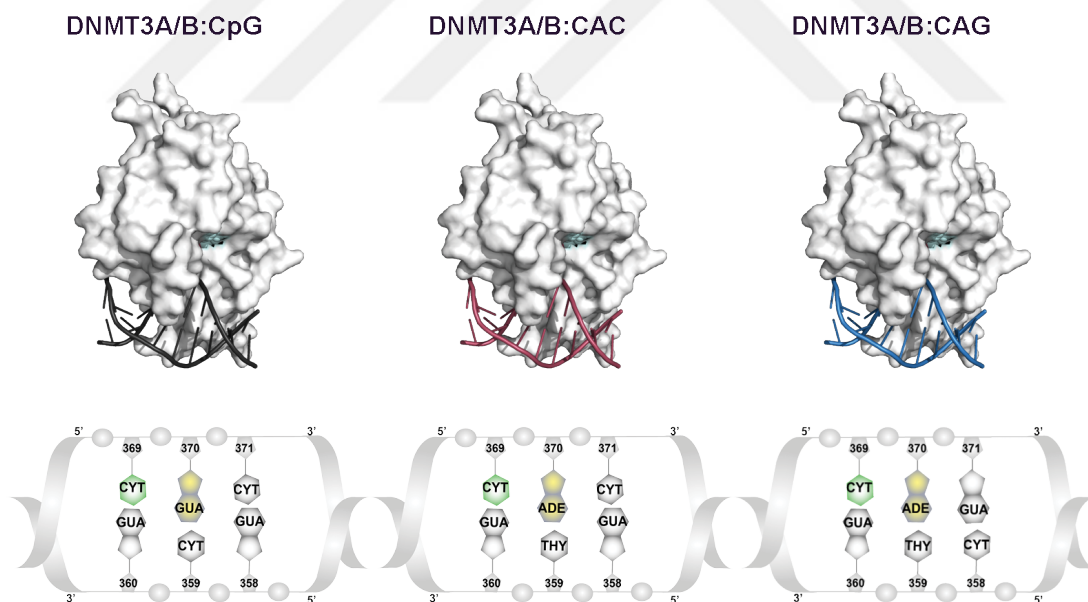


Figure 10: Homology models of DNMT3A/Bs bound to DNA and SAM. Gray DNA shows CpG sites, pink one indicates -CAC- sequence while blue shows -CAG- sequence.

4.3.4. Preparing models for MD

To reveal the structural mechanism of mammalian *de novo* methylation, we have run MD simulations of each DNMT model (generated at the section iii). To overcome the problems regarding missing atoms and steric clashes, we have initially run a short water refinement of these models with HADDOCK 2.2. Next, we have manually selected the most realistic and plausible conformation of each model for MD simulations.

4.3.5. Water Refinement of DNMT3A/B:DNA Models

To improve the accuracy of crude models, we used the water refinement module of HADDOCK 2.2. Since we just refined the models, we did not expect to see any significant difference among models. This was in line with the results recorded in the **Table 7**.

Table 7: Summary of HADDOCK statistics is listed below.

Models	DNMT3A:CpG	DNMT3A:CAG	DNMT3A:CAC	DNMT3B:CpG	DNMT3B:CAG	DNMT3B:CAC
Total energy (kcal/mol)	-346.1	-321.4	-322.7	-321.9	-312.6	-320.9
van der Waals (kcal/mol)	-166.4	-150.1	-151.6	-155.6	-150.7	-151.4
Electrostatics (kcal/mol)	-898.8	-856.2	-855.8	-831.6	-809.5	-847.7
BSA (Å ²)	3878.1	3665.3	3638.8	3609.2	3637.3	3629.5

4.3.6. Deciding the best initial structure for the MD simulations

HADDOCK 2.2 gives top four conformations after ranking all conformations based on HADDOCK score. We indicated that not only the HADDOCK score or the components of HADDOCK score, but also intermolecular interaction network is crucial to discriminate models. In line with this, it has been shown that the conformations of strictly conserved residues, SAM and flipped cytosine are necessary to carry out DNA methylation. Thereby, we chose the best structure of each case by considering: i) the presence of interactions between target cytosine and GLU139, ARG183, ARG185, ii) small distance (less than 4 Å) between C5 atom of target cytosine and the CE atom of SAM iii) buried ring region of SAM inside the DNMT3A/B.

4.3.7. Assessment of the convergence of MD simulations

The quality of MD trajectories depends on two aspects: i) the persistency of trajectories, and ii) the convergence of simulations. MD simulations are expected to sample all possible states. For this, longer simulation time is required. However, the proper simulation time changes with respect to the system of interest. For instance, tracing the intermolecular interactions of a protein:DNA systems requires a minimum run length of ~100-150 ns. Therefore, we generated two parallel simulations, totaling 600 ns for each case.

4.3.8. The persistency of trajectories

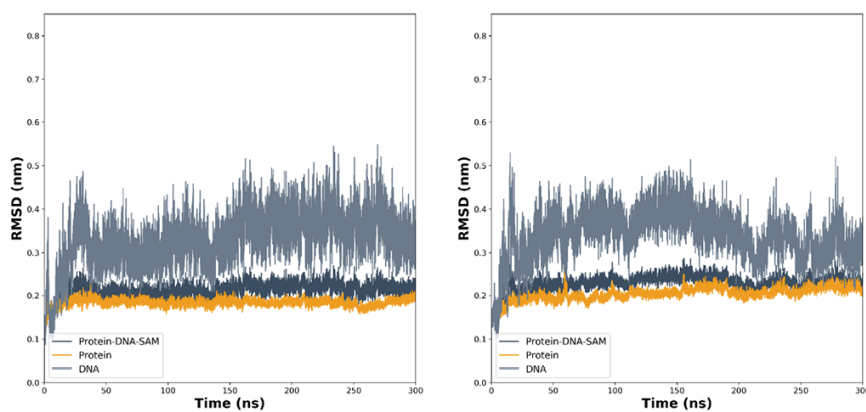
MD is a deterministic approach to simulate the time-dependent behavior of systems. To add the stochastic perspective, multiple simulations should be run with different initial velocities. This aids to distinguish chance-dependent and consistent key factors regarding characteristic aspects of system of interest. For this aim, we had run two parallel simulations with different initial velocities. While exploring the simulations, we had initially looked for the persistent parameters in parallel simulations.

4.3.9. The convergance of trajectories

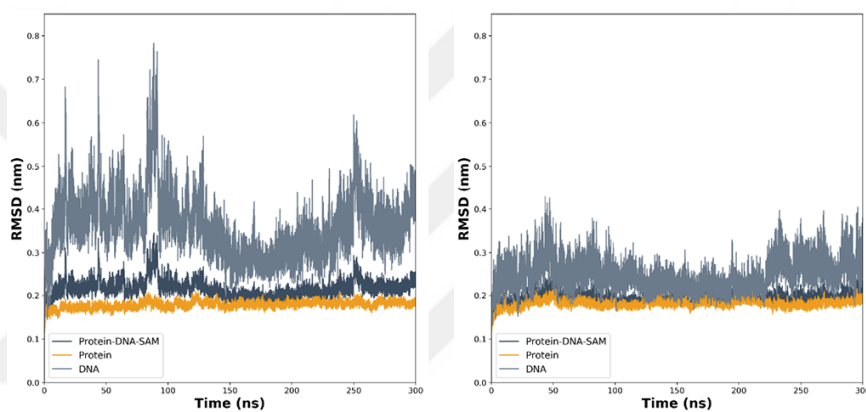
4.3.9.1. Root mean square deviation (RMSD)

The convergence of trajectories was initially controlled with the RMSDs. The first frames of trajectories were taken as a reference to calculate RMSDs. In almost all cases, the fluctuations of protein backbone, DNA and protein:DNA-SAM complexes reached a plateau at 150ns. The highest fluctuations are observed for DNA. This is expected due to the short size of the DNA.

CpG Site



-CAC- Site



-CAG- Site

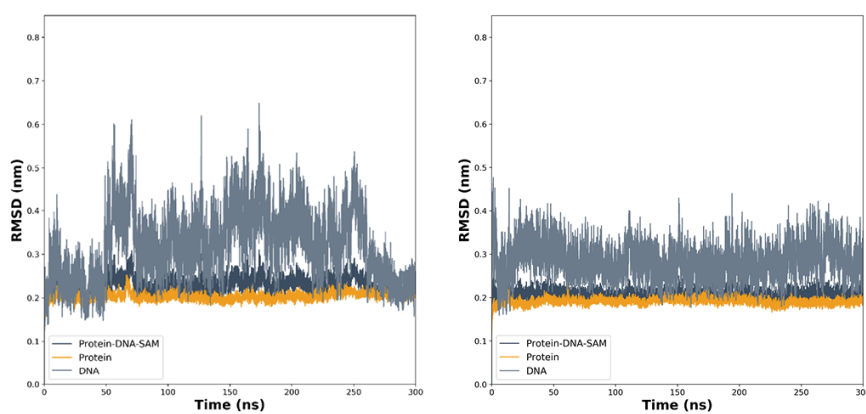
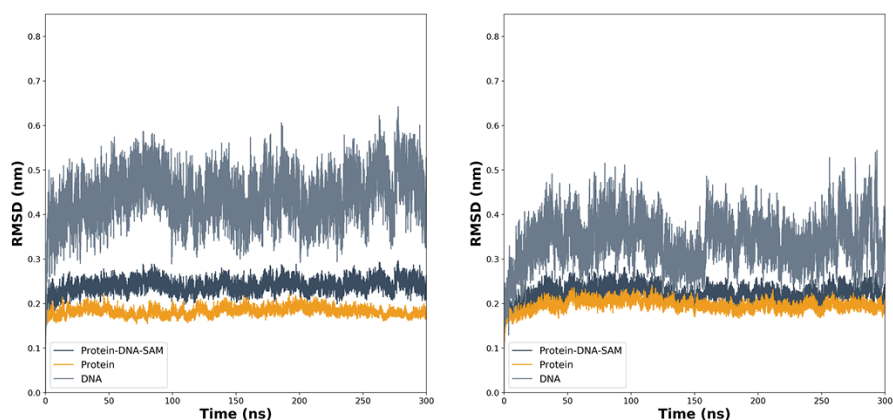
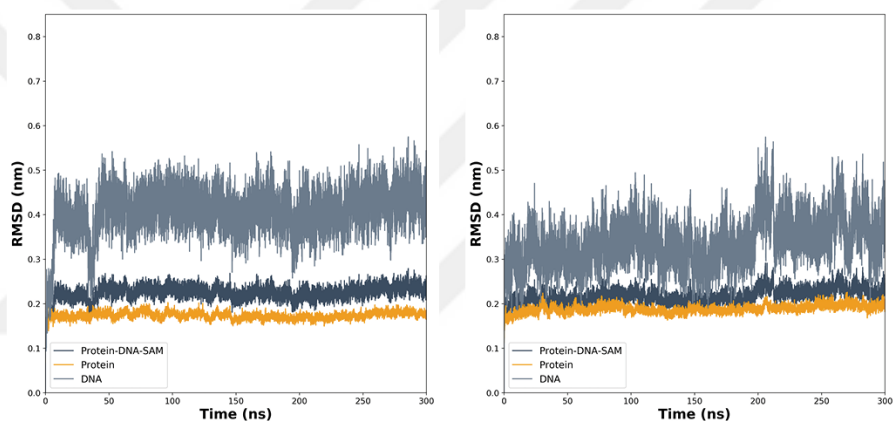


Figure 11: The RMSD of DNMT3A:DNA complexes throughout simulations. Light blue lines show the RMSD of all-atom of DNMT3A:DNA-SAM, yellow ones indicate the RMSD of protein backbone and dark blue lines shows the RMSD of DNA.

CpG Site



-CAC- Site



-CAG- Site

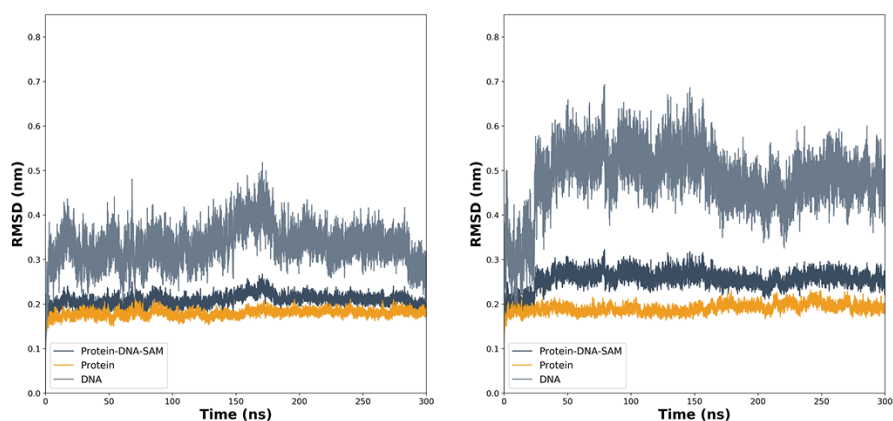


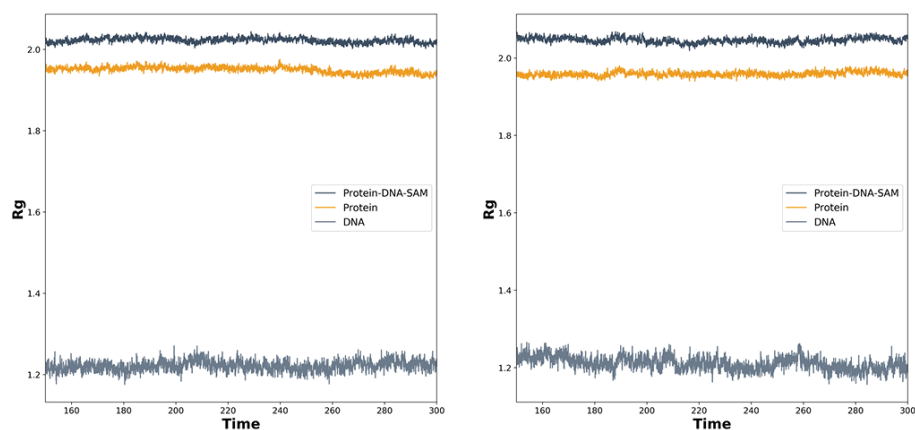
Figure 12: The RMSD of DNMT3B:DNA complexes throughout simulation. Light blue lines show the RMSD of all-atom of DNMT3B:DNA-SAM, yellow ones indicate the RMSD of protein backbone and dark blue lines shows the RMSD of DNA.

4.3.9.2. *Radius of gyration (Rg)*

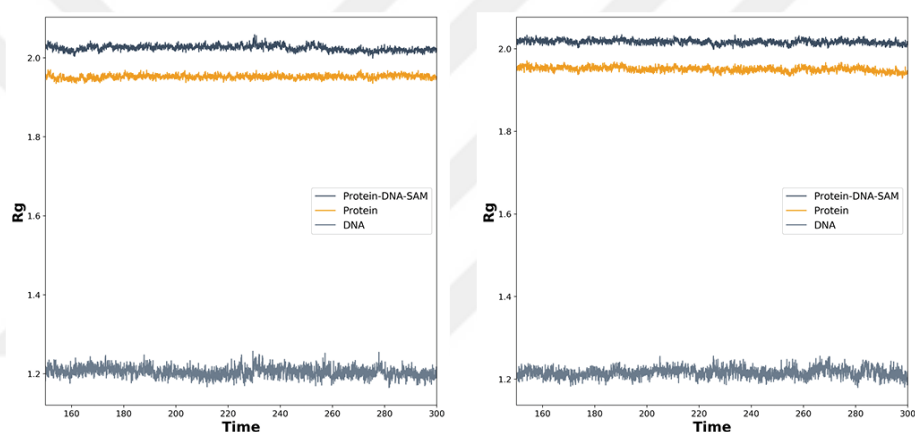
Radius of gyration is another parameter to evaluate the stability of simulations. Rg shows the volumetric alterations throughout the simulations. In line with this, it also reports the compactness of systems. The time evolution of Rg was evaluated for each complex. The compactness of DNMT3A/B enzymes, DNAs and Protein-DNA-SAM (DNMT3A/B:DNA-SAM) was stable throughout the simulations.



CpG Site



-CAC- Site



-CAG- Site

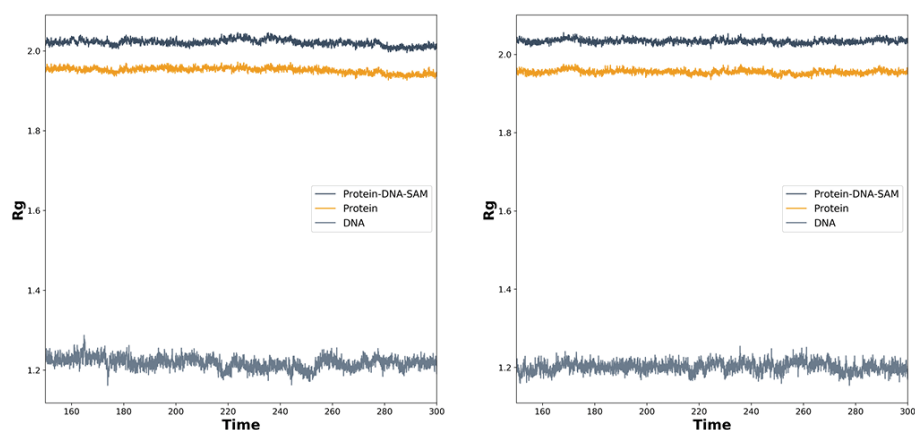


Figure 13: The Rg of DNMT3A:DNA complexes throughout simulation. Light blue lines show the Rg of all-atom of DNMT3A:DNA-SAM, yellow ones indicate the Rg of protein backbone and dark blue lines shows the Rg of DNA.

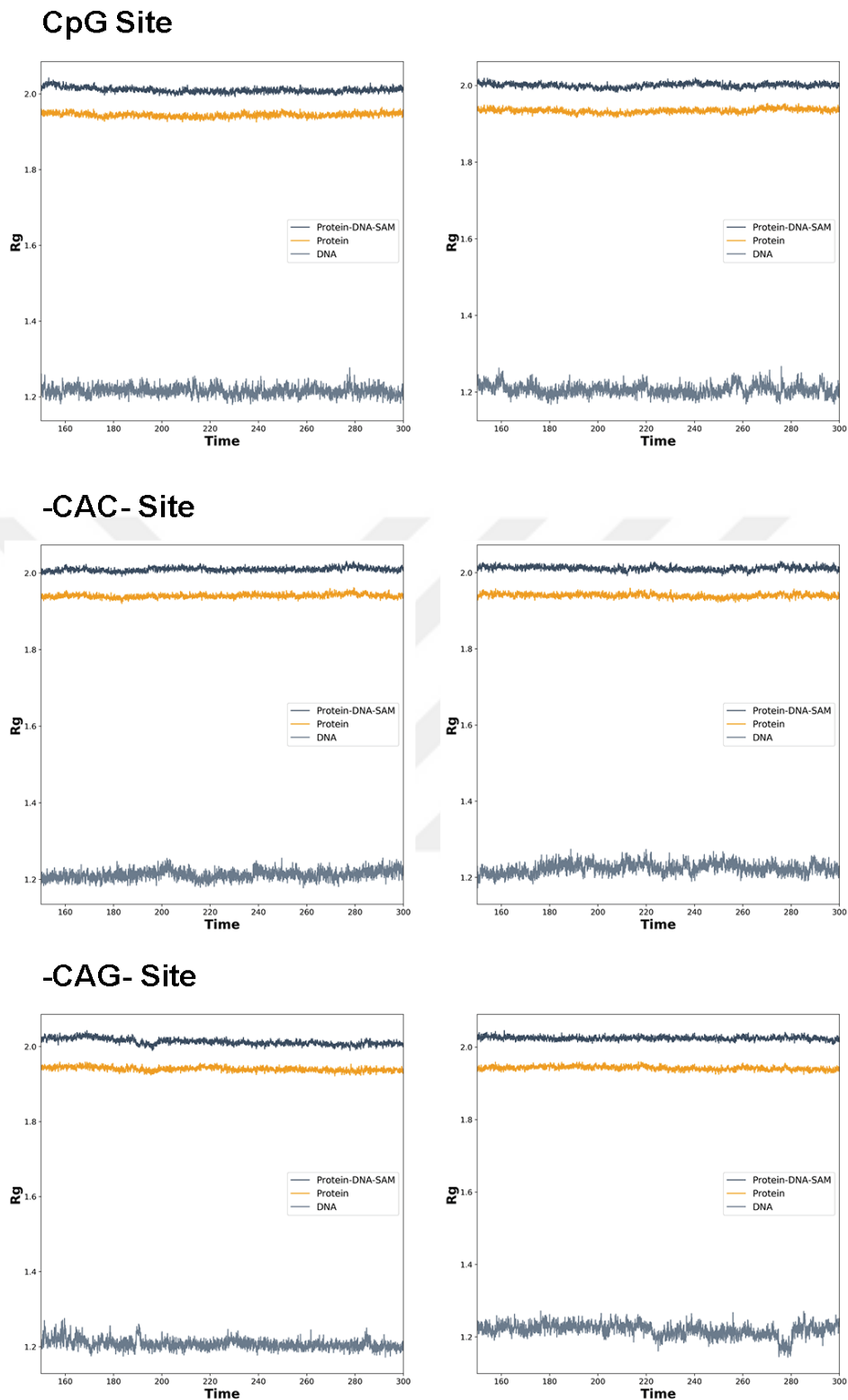


Figure 14: The Rg of DNMT3B:DNA complexes throughout simulation. Light blue lines show the Rg of all-atom of DNMT3B:DNA-SAM, yellow ones indicate the Rg of protein backbone and dark blue lines shows the Rg of DNA.

As a result of RMSD and Rg evaluations, we have decided to exclude first halves of simulations since the trajectories are converged around 150ns. For the further analysis, last 150ns was evaluated.

4.3.10. DNMT3B binds more stably to any DNA sequences compared to DNMT3A.

To investigate the compactness of the DNMT3A/B:DNA models, we have compared the distribution of complex surface area (CSA) of each model. The initial DNMT3A/B:DNA models are similar in terms of their size. However, when we compare the CSA distributions of each complex, observed during the simulation (**Figure 15**), we have seen that DNMT3B:DNAs complexes have lower CSAs compared to DNMT3A:DNA ones. This implies that the DNMT3B enzyme forms tighter complexes. Thereby, we have assumed that the DNMT3B enzyme makes a greater number of intermolecular contacts, compared to the DNMT3A enzyme. Also, DNMT3A complexes reflects the greatest variation with respect to different DNA sequences. This might imply that the internal dynamics of DNMT3B enzyme is not impacted by its partner DNA sequence, while the stability of DNMT3A complex strictly depends on the DNA sequence.

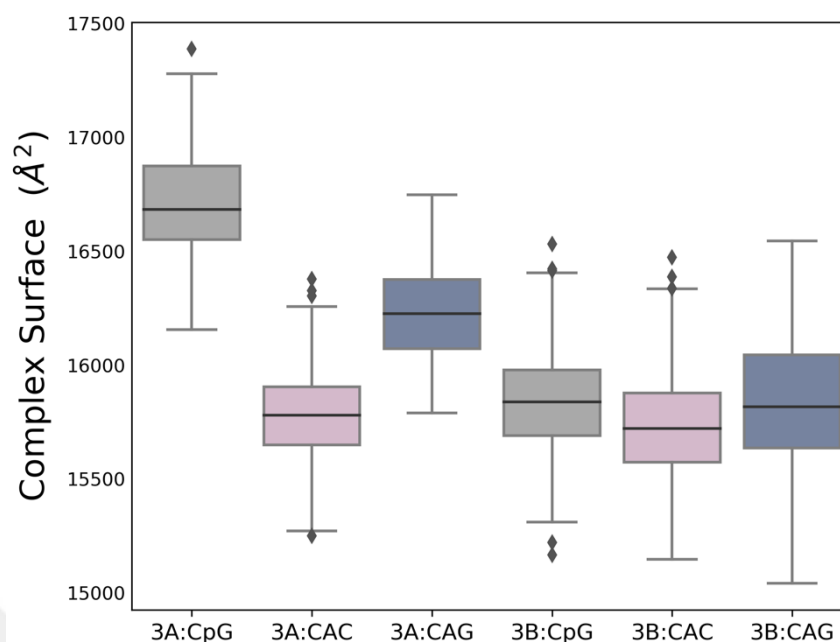


Figure 15: The distribution of complex surface areas of DNMT3A:DNA-SAM complexes. Color code implies type of methylations sites: grey for CpG site, pink for -CAC- sequence and blue for -CAG- sequence.

4.3.11. Comprehensive analysis of interaction network

To explore the stability of DNMT3A/B:DNA complexes, we have calculated the distribution of total hydrogen and hydrophobic bonds throughout the simulations. Compared to the CSA calculations, the calculation of intermolecular interaction is more direct way to understand the stability of DNMT3A/B:DNA complexes. Expanding on our CSA results, we have assumed that the total number of intermolecular contacts of DNMT3B:DNA complexes should be greater than the ones in the DNMT3A:DNA complexes. Also, to dissect the base-readout mechanisms of DNMT3A/B enzymes, we have systematically recorded the complex-specific hydrogen bonds formed across different simulations. We should once more highlight that the DNMT3A shows greater selectivity towards the -CAC- sequences over -CAG- ones, while the reverse is true for the DNMT3B. Here, our initial assumption was that this difference should be reflected in the complex-specific hydrogen bonds.

To trace these complex-specific bonds, we have developed the python package minnie (Molecular INteraction fIngErprints). minnie acquires the important system-

specific contacts upon comparing two distinct simulation sets, while (i) calculating the inter-/intra-monomer hydrogen bonds, ionic, hydrophobic and ring stacking interactions with the interfacea python package; (ii) eliminating the non-essential interactions with a user-defined interaction frequency filter; (iii) comparing the essential interactions across the two simulation sets; and (iv) acquiring the simulation-specific interaction fingerprints. minnie is designed to portray the specific interaction profiles stemming from simulation sets of the same complex, highly homologous systems, or wild-type and mutant states of the same system. This flexibility places minnie in a unique position compared to other MD interaction analysis packages, which leave the interaction comparisons to the user. To showcase the capabilities of minnie, we have run it against the MD trajectories of DNMT3A/B:DNA complexes. As a result, minnie could unambiguously reveal the core and flanking sequence interaction profile differences of these systems.

4.3.11.1. Comparison of total hydrogen bonding network

To elucidate the contribution of protein-DNA hydrogen bonds to the stability of DNMT3A/B complexes, we have compared the distribution of the total number of intermolecular hydrogen bonds of the distinct DNMT3A/B complexes with minnie, as given **Figure 16**. This analysis has shown that DNMT3B enzyme forms greater number of hydrogen bonds with all DNA sequences compared to the DNMT3A. This finding is in line with our CSA results (see section 4.3.10). Further, DNMT3A enzyme shows differential distribution of total number of hydrogen bonds, while DNMT3B makes similar number of hydrogen bonds across DNA sequences. Specifically, DNMT3A forms more contacts with the -CAC- sequence, followed by -CAG- and -CpG- (**Figure 16**). The distribution of intermolecular hydrogen bonds formed across DNMT3A:CAC is less skewed (or closer to normal), compared to the DNMT3A:CAG complex. In line with the CSA analysis, DNMT3A forms the most stable complex with the -CAC- sequence while -CpG- is the least contacted methylation site.

In comparison to DNMT3A, the total number of hydrogen bonds at the interface of DNMT3B:DNA complexes displays similar profile (**Figure 16**). Among all, DNMT3B makes a smaller number of contacts with the -CAG- sequence. The distribution of the total intermolecular hydrogen bonds throughout the simulation is also

similar among different methylation sites. The distribution patterns of DNMT3B:DNA complexes are more normal compared to DNMT3A:DNA complexes.

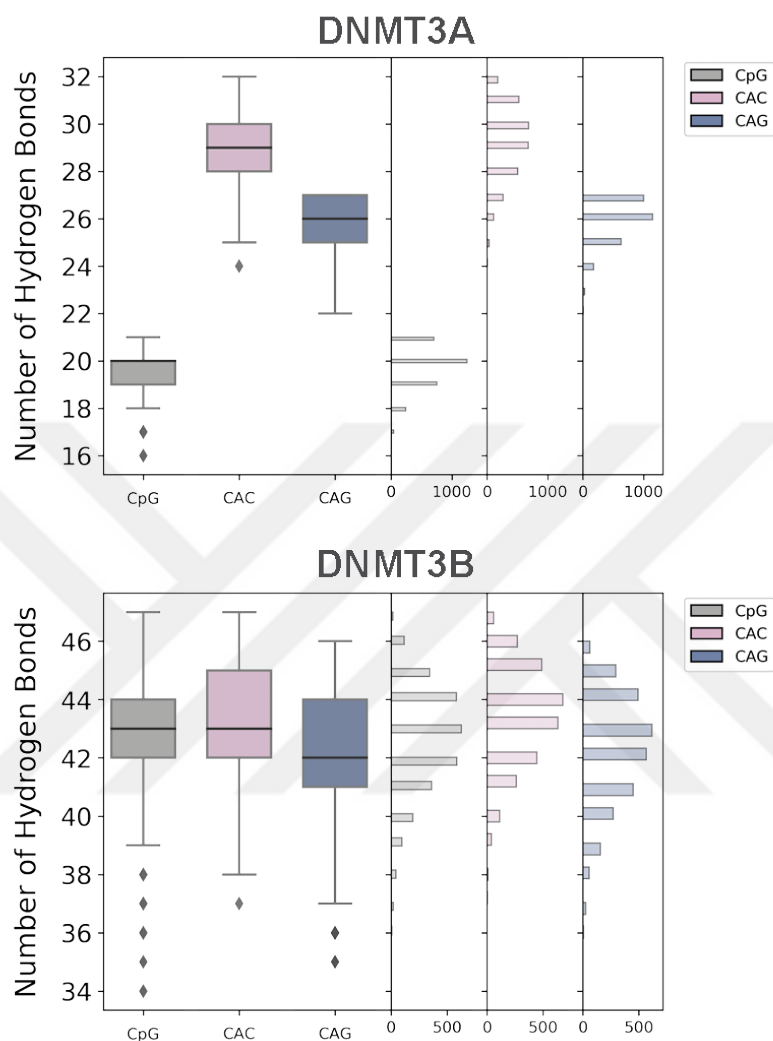


Figure 16: Distribution of total number of intermolecular hydrogen bonds. Color code implies the type of DNA sequence.

All in all, while stability is the discriminate factor for DNMT3A to explain its sequence specificity towards distinct methylation sites, DNMT3B requires further analysis.

4.3.11.2. Complex specific hydrogen bonds

To further enlighten the specificity of DNMT3A/B enzymes, we have systematically compared the complex-specific hydrogen bonding networks of DNMT3A/B complexes. Here, complex-specific indicates the interactions observed only for that particular complex. For the complex-specific hydrogen bonds, DNMT3A:DNA complexes display a drastic difference among CpG and non-CpG sites (**Figure 17**): DNMT3A makes at least 2- to 3-fold smaller number of complex-specific contacts with the CpG compared to the non-CpG sites. We could not detect a statistically significant distribution for the CpG complex. In the case of non-CpG sites, DNMT3A makes eleven complex-specific contacts with the -CAC- sequence on average. It also makes similar number of complex-specific contacts with the -CAG- sequence. Though, the distribution complex specific contacts is more skewed for the DNMT3A:CAG complex.

As similar to DNMT3A, DNMT3B complexes display a significantly distinct profile, depending on the DNA methylation site type (please see **Figure 17**). The highest number of complex-specific hydrogen bonds are detected for the DNMT3B:CAG complex. Thereby, DNMT3B has the greatest specificity towards the -CAG- methylation site followed by the -CpG- and -CAC- sites.

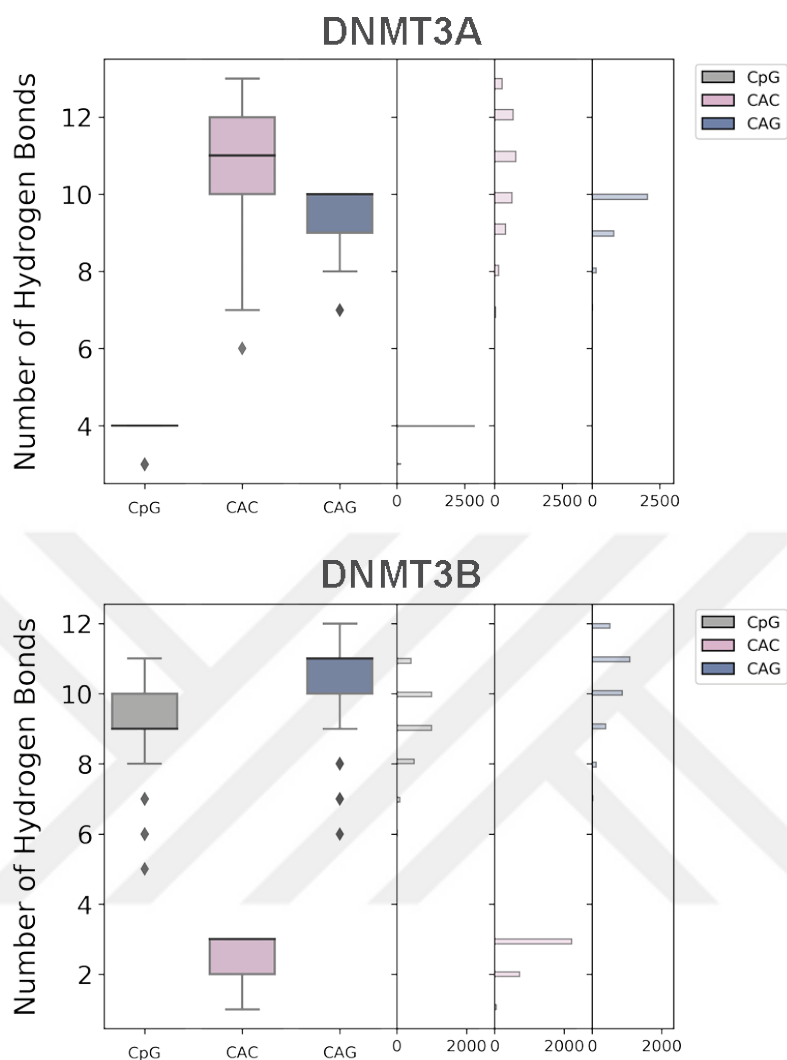


Figure 17: The distribution complex-specific hydrogen bonds in DNMT3A/B:DNA-SAM complexes. DNMT3A makes the greatest number of contacts in DNMT3A-CAC complex while DNMT3B prefers -CAG- sequence over -CAC-.

Overall, this analysis has shown that DNMT3A regulates its sequence specificity through general and complex-specific intermolecular hydrogen bonding networks. On the other hand, DNMT3B's selectivity is encoded only on the complex-specific hydrogen bonds. The amino acids governing the sequence specificity is explored in the following section.

4.3.11.3. *Hydrogen bond mediated Sequence Specificity*

To acquire the fingerprint of the specificity-driving interactions, we further investigated the types of complex specific hydrogen bonds. For this aim, interactions were listed as barcode-like graph in the **Figure 18** and **Figure 19**. DNMT3A/B enzymes use similar amino acids to target the methylation sites. However, the consistency and number of complex specific interactions changes according to their partner DNA sequence.

DNMT3A predominantly targets the complementary strand (355-364) in the case of -CAC-, while it forms similar number of complex-specific interactions with the target strand (365-374) of -CAG-. The flipped cytosine of the target strand is recognized by ARG326 and SER91 in DNMT3A:CpG, ARG185 in DNMT3A:CAC, and TRP328 in DNMT3A:CAG. The orphan nucleotide adjacent to flipped cytosine is interacting with THR269 in DNMT3A:CpG, ASN100, ARG270 and THR268 in DNMT3A:CAG. The pair of flipped cytosine, DG360, is recognized by the ILE98 in the DNMT3A:CAG. As different from CpG site, DNMT3A recognizes the flanking region surrounding the -CAC- and -CAG- sequences. While targeting -CAC- sequence, DC371 in the target strand and DG358 in the complementary strand are interacting with the ASN100, ASN272 and GLN281. On the other hand, corresponding nucleotides in the -CAG- are targeted by ASN94, ASN100 and ARG270. Among these interactions, some further enhances the specificity towards -CAC- and -CAG- sequences, since they occur between the side chain of amino acids and the base region of nucleotides (base specific interactions). For instance, ASN272:DG358 increases the specificity towards the -CAC- sequence while ARG270:DC358 and THR268:DA370 increases the specificity towards the -CAG- sequence.

CpG Site	DNA Groove	DNA Strand	Base Specific
73.98% ARG326:DC369	-	TS	-
94.77% SER91:DC369	Minor G.	TS	-
99.93% THR269:DG370	Major G.	TS	-
-CAC- Site			
99.9% ARG103:DA362	Minor G.	CS	-
100.0% ARG103:DC361	Minor G.	CS	-
91.87% ARG103:DG360	Minor G.	CS	Specific
98.7% ARG103:DG368	Minor G.	TS	Specific
89.44% ARG185:DC369	-	TS	-
94.07% ASN100:DC371	Minor G.	TS	-
66.88% ASN272:DG358	Major G.	CS	Specific
73.51% GLN281:DG358	Major G.	CS	-
-CAG- Site			
97.1% ARG270:DA370	Major G.	TS	Specific
98.63% ARG270:DC358	Major G.	CS	Specific
99.87% ARG270:DG371	Major G.	TS	Specific
88.77% ARG270:DT372	Major G.	TS	Specific
100.0% ASN100:DA370	Minor G.	TS	-
99.77% ASN100:DG371	Minor G.	TS	-
99.43% ASN94:DG371	Minor G.	TS	-
92.44% ILE98:DG360	Minor G.	CS	Specific
78.31% THR268:DA370	Major G.	TS	Specific
98.5% TRP328:DC369	-	TS	-

Figure 18: Time-dependent behaviours of intermolecular hydrogen bonds for DNMT3A:DNA-SAM complexes.

DNMT3B makes a greater number of complex-specific contacts with target strand of the CpG site while it shows a balanced profile in -CAG- sequence. The flipped cytosine in the DNMT3A:CpG is targeted by GLY90 while corresponding residue in the -CAC- is contacted by the SER327. The orphan nucleotide is interacted with the ASN94, ASN100, LY270, THR268 and THR269 in the DNMT3B:CpG complex while same nucleotide is targeted by the LYS278 in the DNMT3B:CAG. ASN98 recognizes the pair of flipped cytosine, DG360, in the DNMT3B:CAG. The sequence specificity of DNMT3B:CAG complex is further enhanced by the LYS270:DG371, LYS278:DA370, ASN272:DG358 and GLN281:DG358.



CpG Site	DNA Groove	DNA Strand	Base Specific
37.62% ASN100:DG370	Minor G.	TS	-
32.06% ASN94:DG370	Minor G.	TS	-
68.81% GLY90:DC369	-	TS	-
93.97% LYS270:DC359	Major G.	CS	-
98.47% LYS270:DG370	Major G.	TS	Specific
83.37% LYS322:DT372	-	TS	-
98.9% THR268:DG370	Major G.	CS	Specific
100.0% THR269:DG370	Major G.	CS	-
-CAC- Site			
25.72% ASN100:DA370	Minor G.	TS	-
99.97% LYS270:DT359	Major G.	CS	Specific
80.21% SER327:DC369	-	TS	-
-CAG- Site			
96.27% ASN272:DC358	Major G.	CS	Specific
95.97% ASN94:DG371	Minor G.	TS	-
92.24% ASN98:DG360	Minor G.	CS	Specific
94.44% GLN281:DC358	Major G.	CS	Specific
59.95% LYS270:DG371	Major G.	TS	Specific
98.5% LYS270:DT372	Major G.	TS	Specific
96.4% LYS278:DA370	Major G.	TS	Specific
99.43% LYS278:DG368	Major G.	TS	-
81.81% LYS278:DT359	Major G.	CS	Specific

Figure 19: Time-dependent behaviours of intermolecular hydrogen bonds for DNMT3B:DNA-SAM complexes.

Lastly, the complex specific interactions are compared as depending on whether they are established through DNA major groove by TRD loop or through DNA minor groove by catalytic loop. They are dominated by the minor groove interactions in the DNMT3A:CAC and by major groove interactions in the DNMT3B:CAG. Five out of seven complex specific interactions are made through the minor groove in the DNMT3A:CAC. On the other hand, seven out of nine complex specific hydrogen bonds are the major groove driven.

4.3.12. The Distribution of Total Hydrophobic bonds

Hydrophobic interaction also contributes to the stability or specificity of biomolecular systems. To dissect their role in the DNMT3A/B:DNA complexes, we have compared the distribution of the total number of intermolecular hydrophobic interactions (please see **Figure 20** and **Figure 21**). As compared to the hydrogen bonds, the total numbers of hydrophobic interactions are much less in number.

Comparison of the DNMT3A:DNA complexes revealed that the CpA sites behave similarly, while DNMT3A:CpG has a distinct profile (please see **Figure 20**). The majority of hydrophobic bonds are made through the minor groove by the catalytic loop of DNMT3A enzyme. The percentage of total hydrophobic interactions are listed in the **Table 8**. ASN94:DA370 is complex specific to DNMT3A:CAC while ILE98 persistently contacts to DA362 in the complementary strand in DNMT3A:DNAs, except the DNMT3A:CAG. Major groove driven hydrophobic bonds further supports the stability of DNMT3A:CAG as different from DNMT3A:CpG/CAC complexes. In the DNMT3A:CAG, ASN272:DG356 and LYS278:DT367 are TRD loop mediated complex specific bonds while ARG326 stabilize the flipped cytosine in the DNMT3A:CAG. These results strengthen the role of ASN94, ILE98, LYS278 and ASN272 in the recognition of target DNA sequences for DNMT3A enzyme as in line with the hydrogen bond analysis.

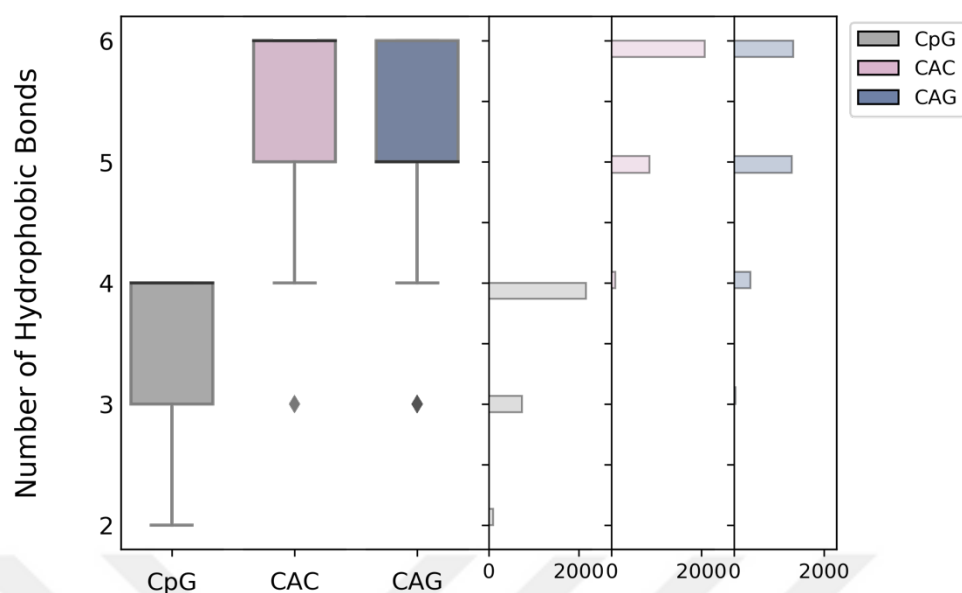


Figure 20: The distribution of intermolecular hydrophobic bonds in the DNMT3A:DNA-SAM complexes. DNMT3A shows similar profile of hydrophobic interaction network in the CpA-bound complexes.

Table 8: Time percentage of identified intermolecular hydrophobic bonds in DNMT3A:DNA complexes

Bond		DNMT3A					
		CpG	CpG-p	-CAC-	-CAC-p	-CAG-	-CAG-p
ASN94	DA370	-	-	94.06	56.91	-	-
ILE98	DA362	98.46	93.50	87.20	72.77	58.04	94.93
VAL99	DG368	90.13	96.53	98.66	99.70	98.43	98.00
ASN100	DG370/DA370	83.77	64.11	91.80	83.63	64.54	74.17
PRO101	DG360	-	-	86.37	71.17	55.04	58.68
	DG370	65.11	30.38	-	-	-	-
	DC371	-	-	56.11	48.91	-	-
ASN272	DG356	42.65	66.14	-	-	55.51	30.05
LYS278	DT367	-	-	-	-	47.88	27.82
ARG326	DC369	-	-	-	-	40.15	99.50

DNMT3B makes a greater total number of hydrophobic contacts compared to DNMT3A (please see Figure 20 and Figure 21). DNMT3B has the least number of contacts with the -CAG- sequence while the greatest number of contacts made with the -CAC- sequence. The minor groove interactions are balanced with the major groove interactions in the DNMT3B:DNA complexes, see **Table 9**. ASN98 and ARG317 makes complex-specific interaction with the CpG site. ASN272:DG356 and GLN281:DT359 are part of the complex-specific contacts in the DNMT3B:CAG. While these results support the role of ASN98 and ASN272 amino acids, it also indicates that its sequence specificity towards -CAG- sequence and CpG site is further enhanced by the GLN281 and ARG317, accordingly.

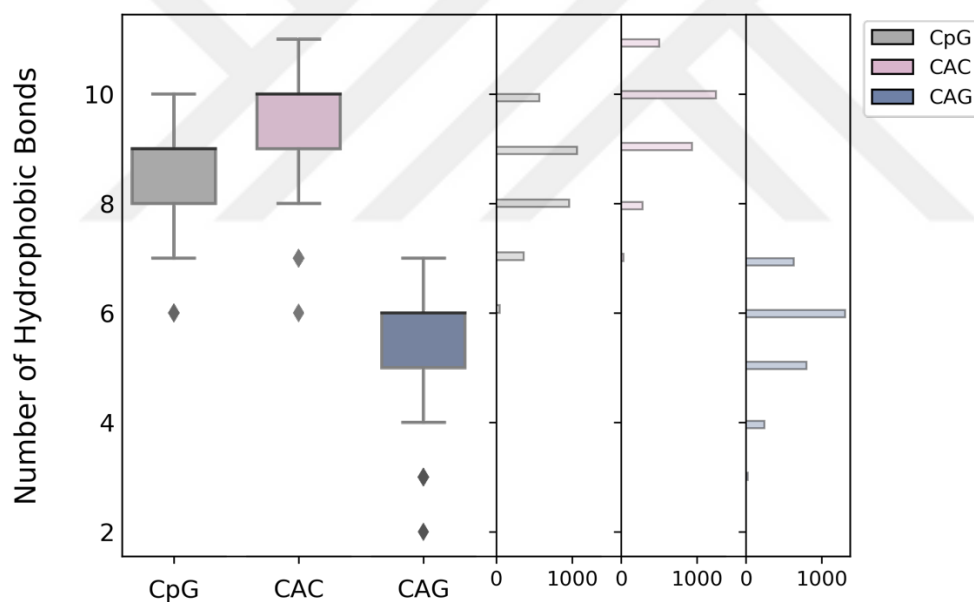


Figure 21: The distribution of intermolecular hydrophobic bonds in the DNMT3B:DNA-SAM complexes. DNMT3B makes a greater number of contacts with -CAC- sequence followed by CpG site and -CAG- sequence

Table 9: Time percentage of intermolecular hydrophobic bonds in DNMT3B:DNAs.

Bond		DNMT3B					
		CpG	CpG-p	-CAC-	-CAC-p	-CAG-	-CAG-p
ASN94	DA370	95.40	84.53	98.46	96.53	-	-
ASN98	DA362	41.18	46.55	-	-	-	-
VAL99	DG368	99.66	96.86	97.86	99.96	97.86	99.86
	DG370	29.62	41.05	34.45	71.64	34.45	51.44
ASN100	DG370/DA370	98.60	93.73	97.23	98.23	-	-
PRO101	DG360	92.50	91.00	84.10	95.86	-	-
	DA370	-	-	38.52	54.81	-	-
	DC371	-	-	-	-	-	-
THR266	DG368	63.31	87.87	95.03	60.41	54.94	53.51
THR268	DG368	63.31	69.14	74.37	92.90	-	-
	DC369	32.15	32.92	26.25	43.31	-	-
ASN272	DG356	-	-	-	-	32.38	40.8
LYS278	DA366	47.81	77.30	35.92	60.21	-	-
	DT367	64.94	68.31	77.84	79.40	69.81	60.07
GLN281	DT359	-	-	-	-	37.38	51.48
ARG317	DA355	30.92	76.87	-	-	-	-
ARG326	DC369	-	-	31.55	39.68	90.16	97.33

In line with the hydrogen bond analysis, hydrophobic bond analysis also revealed that DNMT3B enzyme regulates its sequence specificity through both the TRD-mediated and catalytic-loop-mediated interactions. On the other hand, DNMT3A enzyme makes hydrophobic contacts predominantly by its catalytic loop. How these loops mediate the recognition mechanism of DNMT3A/B enzymes will be explored in the following sections.

4.3.13. The evaluation of the loop dynamics via solvent accessible surface areas (SASA)

The enzymatic activity is tightly regulated by the solvation. To understand the impact of solvation on the working principle of DNMT3A/B enzymes, we have calculated the time evolution of solvent accessible surface area (SASA) of each complex. Since the

TRD loop and catalytic loop are the major contributors of DNMT3A/B:DNA interaction networks (see section 4.3.11), we have focused on the dynamic alteration of these two loops. Here, our assumption is that i) SASA of the DNMT3A's catalytic loop should assist us to discriminate the sequence specificity of the DNMT3A enzyme, ii) SASAs of both TRD and catalytic loop should help us to understand the sequence specificity of the DNMT3B enzyme.

In line with our first assumption, the catalytic loop of DNMT3A shows distinct SASA profiles across different DNA sequences (please see Figure 22): the catalytic loop of DNMT3A:CAC is the least solvent exposed, followed by the DNMT3A:CpG and DNMT3A:CAG. It's important to note that the catalytic loop of DNMT3A:CpG has a wider distribution. Lastly, TRD loops of DNMT3A:DNA complexes have comparable mean SASA. However, DNMT3A:CAC complex has a narrower distribution. These results also indicate, in line with the intermolecular interaction and the CSA analysis, that the dynamic alteration of the catalytic loop conformation of DNMT3A dictates the sequence specificity and stability of DNMT3A enzyme.

In line with our second assumption, DNMT3B enzyme utilizes both catalytic and TRD loop to govern its sequence specificity towards different methylation sites (please see section 4.3.11). DNMT3B:CpG and DNMT3B:CAC complexes have similar SASA profile while a more drastic change is measured in the DNMT3B:CAG complex. The catalytic loop of DNMT3B:CAG is the most solvent exposed while the TRD loop is the least solvent exposed one. This implies that the interplay of TRD and catalytic loops dictates the sequence specificity of DNMT3B enzymes.

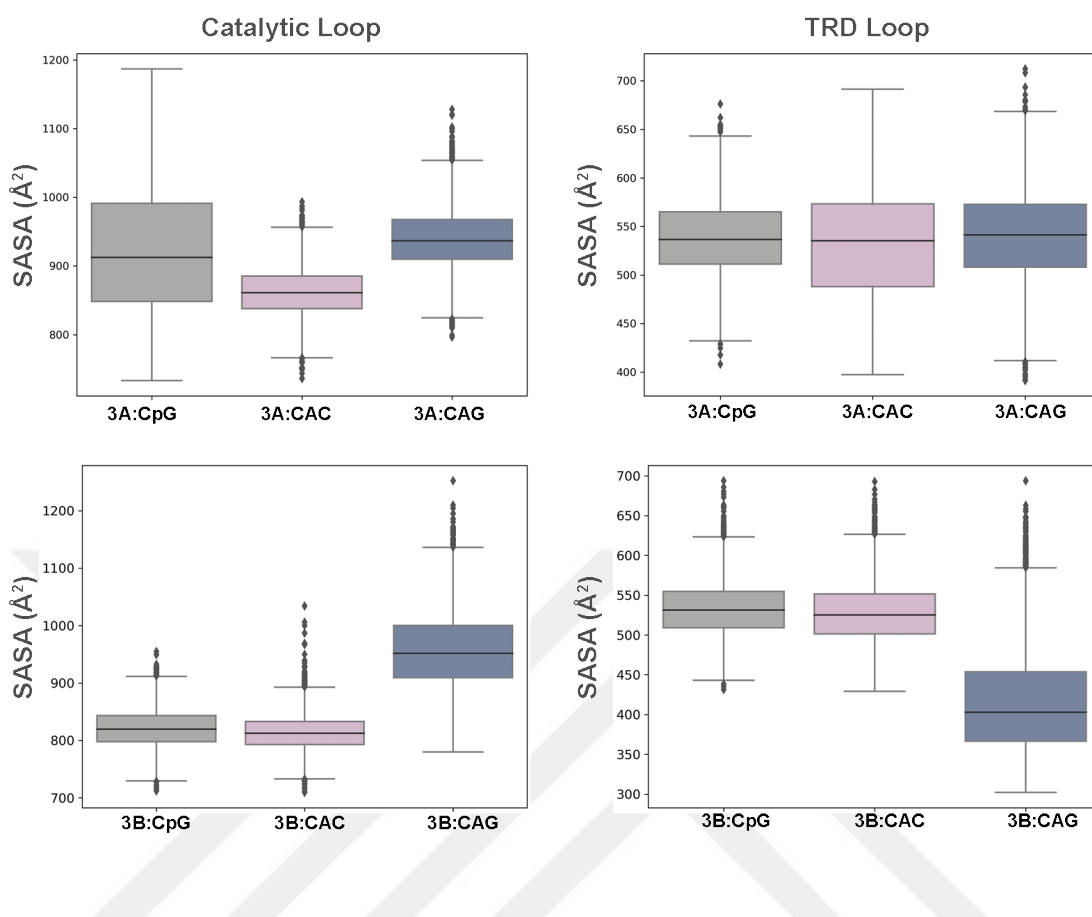


Figure 22: SASA-based Distribution of TRD and catalytic loops in DNMT3A/B:DNA complexes.

Consequently, these results imply that although the TRD and catalytic loops of DNMT3A/Bs do not make any contacts, they still dictate the sequence specificity of DNMT3A/B enzymes. In the following section, we will investigate how the distinct TRD and catalytic loop dynamics influence the dynamics of DNMT3A/Bs subdomains.

4.3.14. The investigation of internal dynamics of DNMT3A/B complexes

DNMT3A enzyme regulates the DNA methylation in a cooperative manner. However, the DNA methylation mechanism of DNMT3B enzyme is employed by a non-cooperative processive manner. Here, we have investigated the role of DNMT3A/B subdomains in their working style. For this, we have looked into the signal propensities of residue pairs for each complex. According to this analysis, the smallest signal propensity reflects the highest level of signal transduction between two residues. While comparing the interdomain signaling, we have

considered DNMT3s four subdomains: i) SAM-binding subdomain (residues 8-58), ii) catalytic subdomain (residues 83-199), iii) target recognition subdomain (residues 220-315), and (iv) SAM&DNA binding region (residues 325-351). The boundaries of the subdomains were approximately chosen, based on the clustered signal and their position on the structure.

In the **Figure 23**, the catalytic-loop-mediated signaling propensity of DNMT3A/B:DNA is shown. According to the analysis, The catalytic loop strongly communicates with DNMT3A:CAG and DNMT3B:CAC. As mentioned before, DNMT3A prefers -CAC- over -CAG- and the reverse is true for DNMT3B. Thereby, the catalytic loop might act as a negative regulator of methylation activity.

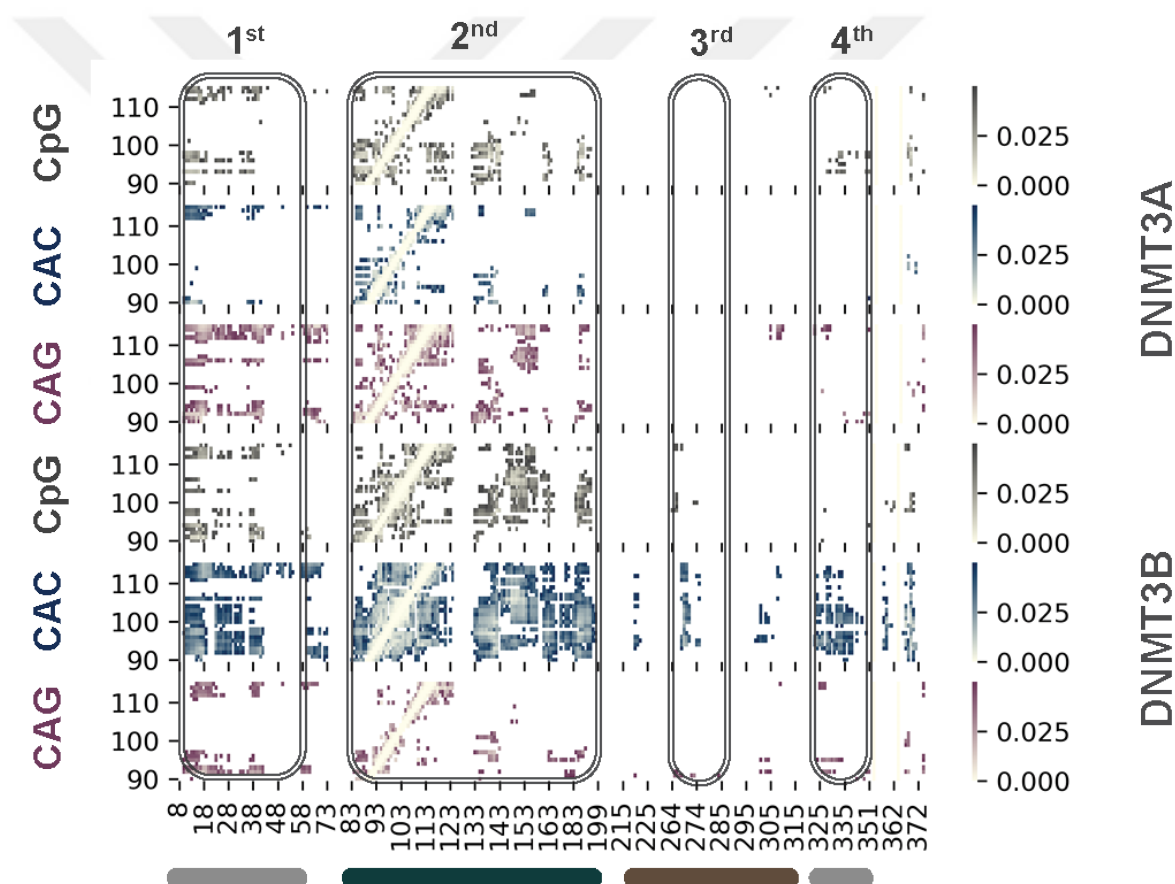


Figure 23: Signaling propensity of the catalytic loop with respect to the rest of DNMT3A/B:DNA complexes. The 1st region shows where SAM-binding motifs are located. The 2nd regions represents catalytic center, where most of the conserved motifs are located. The 3rd region shows TRD loop, and lastly the 4th region indicates SAM&DNA binding region.

In comparison to the catalytic loop, the difference of TRD loop signal propensity among DNMT3A/B:DNA complexes is less as seen in **Figure 24**. Among DNMT3A:DNA complexes, the most noticeable difference has been observed in the 4th region of DNMT3A:CAC complex. For the DNMT3B:DNA complexes, the greatest signal has been detected at the 1st and 4th region of DNMT3B:CAG complex. In line with this, we proposed that TRD loop might support the catalytic activity of DNMT3A/B enzymes not just altering DNA binding affinity but also by affecting SAM binding. Lastly, we could say that the TRD loop of DNMT3B is more effective since it also communicates with the 1st region unlike DNMT3A enzyme.

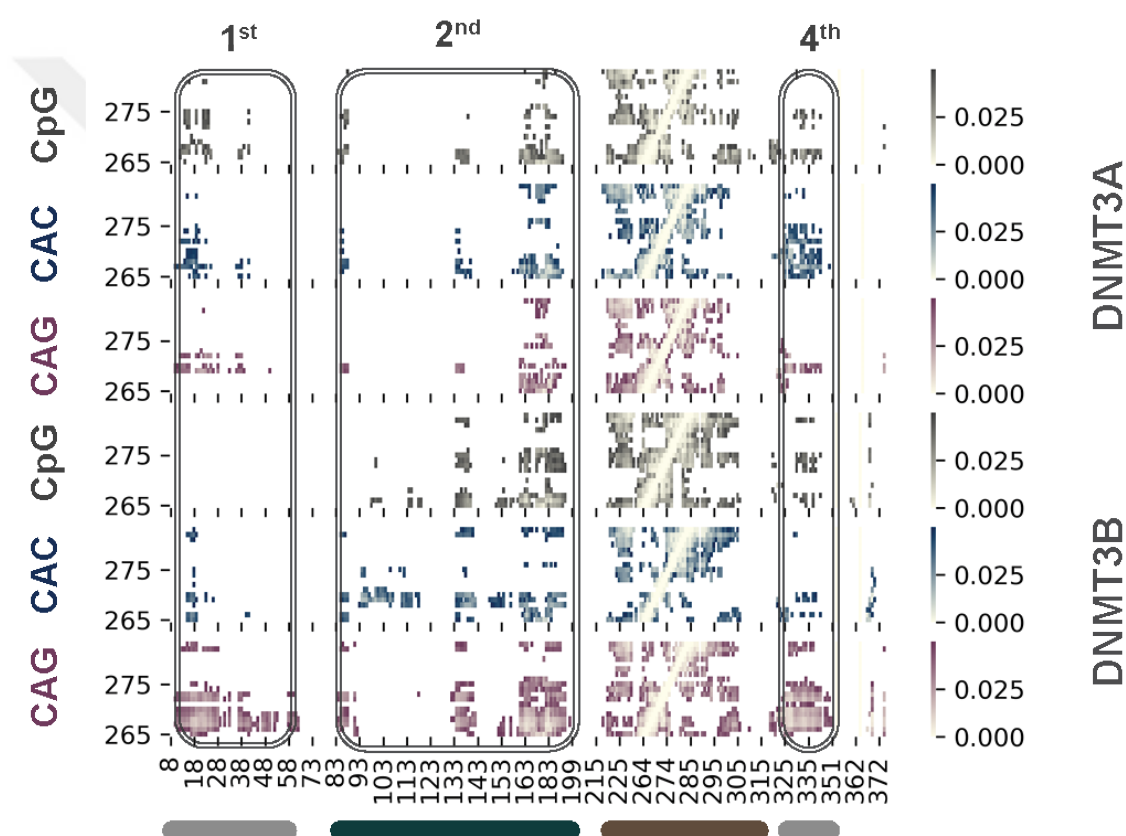


Figure 24: Signaling propensity of the TRD loop with respect to the rest of DNMT3A/B:DNA complexes. The 1st region shows where SAM-binding motifs are located. The 2nd regions represents catalytic center, where most of the conserved motifs are locate, and lastly the 4th region indicates SAM&DNA binding region.

4.3.15. Essential amino acids

Finally, overall essential amino acids are collected at the **Table 10**. These amino acids were chosen with regards to interaction analysis. Among them, ASN272 and GLN281 might contribute to the sequence specificity towards -CAC- and -CAG- sequences.

Table 10: Critical amino acids for the working principles of DNMT3A/B enzymes.

DNMT3A			DNMT3B		
Residue	Mutation	Targeting Site	Residue	Mutation	Targeting Site
S91	-	CpG	-	-	-
I98	-	CAG	N98	glioma	CAG
R103	-	CAC	-	-	-
T268	AML	CAG	T268		CpG
R270	-	CAG	K270		CpG,
N272	AML, West syndrome	CAC	N272		CAG
-	-	-	K278	cancer	CAG
Q281*	AML	CAC	Q281		CAG
-	-	-	K322		CpG

5. DISCUSSION

DNA methylation is an integral process for cellular homeostasis. Prokaryotes use the DNA methylation to protect their genomic content against the viral attack. On the other hand, eukaryotes utilize DNA methylation to regulate cellular differentiation, genomic stability and gene expression. Although it has been a widely studied topic for years, the working principles of the DNA methyltransferases is yet to be explored.

DNA methylation is carried out by the DNA methyltransferases in two steps: i) *de novo* methylation for the establishment of DNA methylation pattern, and ii) maintenance methylation to preserve already established DNA methylation pattern. The most common DNA methyltransferase is C5-methyltransferase, which transfers the methyl group of SAM to the C5 atom of the target cytosine.

Bacterial *HhaI* methyltransferase is the most well-studied simple model of C5-methyltransferases. It particularly methylates the second cytosine at the 5'-GCGC-3' DNA motifs. *HhaI* enzyme takes role in both *de novo* and maintenance methylation. For the sake of simplicity, we initially explored the working principles of *HhaI* enzyme. To understand the role of methyl donor (SAM) and sequence specificity of methyltransferase enzymes, we modelled four different *HhaI* enzymes (5MHT): i) 5MHT:CYT-SAM, ii) 5MHT:CYT-SAH, iii) 5MHT:5CM-SAM, and iv) 5MHT:5CM-SAH. CYT-involved complexes are used to investigate the structural mechanisms of *de novo* methylation while 5CM-included complexes assist to gain mechanistic insight into the maintenance methylation. After modeling, following the optimization of the *HhaI*:DNA interface, we had indicated that the type of ligand (SAM or SAH) do not have a significant effect on the interface energy scores. Despite this, we have shown that SAM and SAH induce different binding pattern, thus we have decided to use SAM in the further steps. To discriminate sequence specificity of *HhaI* enzyme, we further investigated the contribution of each interface residues to the total energy of models. This analysis has revealed that major groove mediated interactions are the key to distinguish DNA methylation states for the *HhaI* enzyme. ARG248 and LYS102 are crucial to target and stabilize hemi-methylated DNA. ARG185 is essential to differentiate SAM-bound complexes from the SAH-bound complexes. In line with the energy decomposition

analysis, intermolecular hydrogen bond analysis indicated that *HhaI* enzyme utilizes LYS89, ARG97, ARG165, ARG228, LYS234 and ARG240 amino acids to target its methylation motif. Furthermore, GLN237 of TRD loop makes base-specific hydrogen bonds with the complementary strand, and SER97 of catalytic loop in the 5MHT:5CM-SAM. In addition to GLN237, ARG260 also contributes to the sequence specificity of *HhaI* enzyme towards hemi-methylated DNA. Corresponding critical amino acids of the *HhaI* enzyme is listed in the **Table 5**. These amino acids further help to elucidate the critical amino acids in the mammalian *de novo* methyltransferase enzymes.

DNMT3A/B enzymes are intrinsic for proper embryonic development. The mutant DNMT3A/Bs prompt cancer or developmental defects because of altered sequence specificity or activity of DNMT3A/B enzymes. DNMT3A/B enzymes are predominantly responsible of *de novo* methylation, however, DNMT3B enzyme also takes role in the maintenance methylation. To understand the selectivity of DNMT3A/Bs towards hemi-methylated (5CM) and non-methylated (CYT) DNA, we modelled four different DNMT3A/B:DNA-SAM complexes: i) DNMT3A:CYT-SAM, ii) DNMT3A:5CM-SAM, iii) DNMT3B:CYT-SAM, and iv) DNMT3B:5CM-SAM.

We have seen that DNMT3B has no significant selectivity towards any of methylation states, yet, DNMT3A prefers non-methylated DNA over hemi-methylated DNA. On the other hand, because each model shows a wide range of standard deviations, it is not possible to jump into a certain conclusion. On the other hand, visual inspection of optimized conformations showed an artificial DNA topology. Therefore, we decided to investigate the working principles of DNMT3A/B enzymes by using MD simulations.

The *de novo* mechanism of DNMT3A/B enzymes is required for further investigation because of two reasons: i) methylation of non-CpG sites and ii) the controversial results regarding the working principles of DNMT3A enzyme. As different from *HhaI* enzyme, DNMT3A/B enzymes are not restricted to the single methylation site. They can methylate both CpG and non-CpG sites under the influence of flanking regions. The most methylated site is -CpG- sequence while it is followed by the -CAC- (for DNMT3A) and -CAG- sequences (for DNMT3B). Also, it has been shown that DNMT3B enzyme processively and non-cooperatively methylates DNA sequences, while DNMT3A follow either distributive or processive DNA methylation mechanism in a cooperative manner. Further

comparative studies revealed that DNMT3B enzyme efficiently methylates the nucleosomal DNA while DNMT3A loses its activity in presence of nucleosome.

To understand the working principles and sequence specificity of the DNMT3A/B enzymes, initially, we modelled six different complexes by using SAM as methyl donor: i) DNMT3A:CpG, ii) DNMT3A:CAC, iii) DNMT3A:CAG, iv) DNMT3B:CpG, v) DNMT3B:CAC, and vi) DNMT3B:CAG. Comparison of CpG and CpA sites would help to dissect the critical residues for the recognition of methylation sites. On the other hand, -CAC- and -CAG- sequences would help to dissect the influence of flanking regions.

To understand the stability of complexes, we compared the complex surface area of each case. This indicated that DNMT3B forms tighter complexes since the CSAs of DNMT3B:DNAs is predominantly smaller than DNMT3A:DNAs. On the other hand, the CSA might be used as a discriminative factor to understand the sequence specificity of DNMT3A enzyme, since DNMT3A shows significantly different distribution of CSAs across different DNA sequences. DNMT3A enzyme makes the most stable complex with the -CAC- sequence, followed by -CAG- sequence and CpG sites. This implies that the stability of DNMT3A:DNA complexes is dictated by the DNA sequences. On the other hand, DNMT3B is capable of forming stable complexes as not depending on the DNA sequences.

Next, to support the role of stability in the working principles of DNMT3A/B enzymes, we further compared the distribution of total number of intermolecular hydrogen bonds among the DNMT3A/B:DNA complexes as depending on the alterations in the DNA sequence. The results were in line with the CSA analysis since DNMT3B:DNA complexes have greater number of intermolecular hydrogen bonds compared to DNMT3A:DNA complexes. This striking difference supports the distinct structural mechanisms of DNMT3A/B enzymes. Since monomer DNMT3B enzyme displays more stable profile, it might carry out DNA methylation as a monomer while DNMT3A intrinsically requires the formation of higher-order complexes (please see section 4.2.1.). The higher-order complexes might further enhance DNA-binding capacity of monomer DNMT3A enzyme while increasing DNMT3A's activity. On the other hand, variations among the distribution of the intermolecular hydrogen bonds is consistently similar across DNMT3B:DNAs while they are significantly different among DNMT3A:DNAs.

DNMT3A:CAC has the greatest number of hydrogen bonds, followed by -CAG- sequence and CpG site. This implies that the sequence specificity of DNMT3A enzyme is primarily determined by DNA binding stability of DNMT3A enzyme. This might explain why the methylation activity of DNMT3A is significantly depends on its association with the DNMT3L enzyme.

To further elucidate the sequence specificity of DNMT3A/B enzymes towards distinct methylation sites, we filtered common hydrogen bonds out to gather complex-specific hydrogen bonds. The distributions of the complex-specific intermolecular hydrogen bonds are in line with the sequence specificity of DNMT3A/B enzymes. For instance, DNMT3A:CAC complex has the highest number of complex-specific interactions, followed by DNMT3A:CAG and DNMT3A:CpG. On the other hand, DNMT3B makes the greatest number of complex specific interactions with the -CAG- sequence, which is followed by the CpG site. Further categorization of hydrogen bonds, as depending on the DNA groove and strand, and base-specificity, revealed that DNMT3A:CAC involves minor groove driven interactions, which are made through the catalytic loop of DNMT3A enzyme. However, DNMT3B recognize -CAG- sequence through the major groove interactions, which are mediated by the TRD loop of DNMT3B enzyme. However, this difference might come from the type of sequence because DNMT3A utilizes the minor and major groove interactions to target -CAG- sequence. On the other hand, DNMT3B is not capable of making stable minor-groove interactions as DNMT3A. Thereby, it cannot stably bind to -CAC- sequence. This affects the activity of DNMT3B and the methylation level at the -CAC- sequence. Another point to add is that the greatest level of base-specific interactions is observed in the DNMT3A/B:CAG complex. On the other hand, DNMT3A/B enzymes recognize both target and complementary strand when they bound to their most preferred methylation site. However, DNMT3A predominantly contacts with the target strand of -CAG- sequence. These results imply that while complex-specific interactions enlighten the sequence specificity of DNMT3A/B enzymes, base-specific interactions and the type of most contacted DNA strand and groove also contribute to the sequence specificity of DNMT3A/B enzymes.

Another factor determining the stability and specificity of complex is hydrophobic bonds. The total number of hydrophobic bonds are much smaller than the hydrogen bonds. Since protein:DNA systems are predominantly driven by the electrostatic forces, this was

expected. On the other hand, DNMT3A makes the similar number of hydrophobic bonds with -CAC- and -CAG- sequences while DNMT3B has a distinct profile among the DNA sequences. DNMT3B makes the highest number of hydrophobic bonds with the -CAC- sequence while the least number of hydrophobic bonds are detected at the DNMT3B:CAG complex. Thereby, DNMT3B might utilize the hydrophobic bonds as a repulsive force to differentiate DNA sequences. The hydrophobic bonds in the DNMT3A:DNA complexes are dominated by the catalytic-loop-driven interactions while DNMT3B interacts through its catalytic and TRD loops.

The minor groove interactions are mediated by the catalytic loop of DNMT3A/B enzymes while target recognition loop of DNMT3A/Bs guides the major groove interactions. As we showed and stated in the literature, maintenance methyltransferases make a hydrogen bond between target recognition and catalytic loop while recognizing the hemi-methylated DNA. Since DNMT3s have shorter target recognition and catalytic loops, they cannot form such bond. However, interaction analysis revealed that although they cannot contact to each other, still interplay of TRD and catalytic loop shapes the methylation activity of DNMT3A/Bs.

Another important point is that DNMT3s use water molecules to carry out their function. The water molecules might contribute to the specificity of enzyme or the reaction mechanism of methylation. Because calculating the water-mediated interactions are computationally intensive, we had measured the solvent accessible surface areas (SASA) of target recognition and catalytic loops for each complex. The distribution of SASA revealed that target recognition loop of DNMT3A:DNA complexes are highly similar. On the other hand, the SASA profiles of catalytic loop are changes with the effect of DNA sequence. DNMT3A:CAC has the least solvent exposed catalytic loop among the DNMT3A:DNA complexes. In DNMT3B:DNA complexes, DNMT3B:CAG shows a unique profile: its target recognition loop is the least solvent exposed while the catalytic loop is the most solvent exposed. The SASA profiles of DNMT3B:CpG and DNMT3B:CAC are highly similar. The last point is that the distribution of SASA of DNMT3B:CAG and DNMT3A:CAC are wider than other DNA sequence. This shows that the target recognition loop requires flexibility to a certain extent. This flexibility also explains how same amino acid might contact to different nucleotides with high percentage. Overall, the SASA profiles shows that solvents are predominantly vital for

the catalytic activity of DNMT3A/Bs while they might also play an essential role in the DNA sequence specificity of DNMT3B enzyme especially towards -CAG- sequence. Finally, while catalytic loop is intrinsic to the activity of DNMT3A/B, target recognition loop is also prominent for the DNMT3B. On the other hand, the flexibility of target recognition loop is another factor to be considered while evaluating the sequence specificity of DNMT3A/Bs. To the best of our knowledge, this is the first time the interplay of target recognition loop and catalytic loop of the DNMT3 family has been shown.

To further dissect the intradomain dynamics of DNMT3A/B proteins, we evaluated the signaling propensity of DNMT3:DNA complexes. Signaling propensity measures the effective signaling between the residues as a function of fluctuations of distance between residue pairs. Since we have shown the impact of TRD and catalytic loops, we focused on their signal propensity throughout complex. According to results, catalytic loop might act as a negative modulator of non-CpG methylation activity since the smallest signal propensity is detected at the DNMT3A:CAC and DNMT3B:CAG complexes. For the CpG sites, the catalytic loop of DNMT3B communicates with residues 143-153 (at 2nd region) which is utilized to bind 3' end of complementary strand. However, the catalytic loop of DNMT3A do not behave similarly. Since residues 143-153 are variant among DNMT3A/B enzymes, this might imply that one of the factors governing processivity is communication of the catalytic loop with the DNA terminals (to be more specific, 3' end of complementary strand). On the other hand, TRD loop signals have smaller difference among complexes compared to the catalytic-loop-driven signals. However, DNMT3B:CAG sequence has the greatest signals at the SAM-binding regions. Mutations imposed at the TRD loop.

Overall, we found the crucial residues mediating sequence specificity of DNMT3A/B enzymes while also establishing critical factors for the structural mechanisms of DNMT3A/B enzymes through the computational structural biology approaches. These results would enlighten the controversial aspects of the working principles of DNMT3A/B enzymes while bringing a novel aspect regarding their regulatory mechanism.

6. CONCLUSION AND SUGGESTIONS

- We modelled and investigated the DNA methyltransferases bound to DNA and SAM/SAH. We have seen that such systems require more advanced computational tools, e.g. MD simulations, to be studied.
- The structural mechanism of DNA methyltransferase enzymes, governing sequence specificity towards distinct methylation sites, are conserved among species. It is dictated by the interplay of TRD loop and catalytic loop.
- The systematic comparison of intermolecular interaction networks would reveal the sequence specificity and the activity of DNA-binding enzymes. For this aim, we developed minnie python package.
- For the methylation activity of DNMT3A/B enzymes, not only specific interactions but also the stability of complexes shapes the sequence specificity.
- The collective communication of the TRD loop and catalytic loop, DNA sequence and SAM-binding regions defines the methylation activity of DNMT3A/B enzymes.
- To elucidate the impact of a regulatory protein, DNMT3L, on the working principles of DNMT3A enzyme, further analysis is required. Thereby, we designed another project (TUBITAK1002) to investigate the role of DNMT3A oligomerization on the working principles of DNMT3A enzyme. This is currently one of the ongoing research projects in our lab.

REFERENCES

- Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., . . . Mann, R. S. (2015). Deconvolving the recognition of DNA shape from sequence. *Cell*, 161(2), 307-318. doi:10.1016/j.cell.2015.02.008
- Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*, 44(W1), W344-350. doi:10.1093/nar/gkw408
- Bardwell, A. J., & Bardwell, L. (2015). Two hydrophobic residues can determine the specificity of mitogen-activated protein kinase docking interactions. *J Biol Chem*, 290(44), 26661-26674. doi:10.1074/jbc.M115.691436
- Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., . . . Zardecki, C. (2019). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res*, 47(D1), D464-D474. doi:10.1093/nar/gky1004
- Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., . . . Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, 66(Pt 1), 12-21. doi:10.1107/S0907444909042073
- Chennubhotla, C., & Bahar, I. (2007). Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput Biol*, 3(9), 1716-1726. doi:10.1371/journal.pcbi.0030172
- Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7), 1731-1737. doi:10.1021/ja026939x
- Dong, A., Zhou, L., Zhang, X., Stickel, S., Roberts, R. J., & Cheng, X. (2004). Structure of the Q237W mutant of HhaI DNA methyltransferase: an insight into protein-protein interactions. *Biol Chem*, 385(5), 373-379. doi:10.1515/BC.2004.041
- Emperle, M., Rajavelu, A., Reinhardt, R., Jurkowska, R. Z., & Jeltsch, A. (2014). Cooperative DNA binding and protein/DNA fiber formation increases the activity of the Dnmt3a DNA methyltransferase. *J Biol Chem*, 289(43), 29602-29613. doi:10.1074/jbc.M114.572032
- Faltejskova, K., Jakubec, D., & Vondrasek, J. (2020). Hydrophobic Amino Acids as Universal Elements of Protein-Induced DNA Structure Deformation. *Int J Mol Sci*, 21(11). doi:10.3390/ijms21113986
- Gao, L., Emperle, M., Guo, Y., Grimm, S. A., Ren, W., Adam, S., . . . Song, J. (2020). Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nat Commun*, 11(1), 3355. doi:10.1038/s41467-020-17109-4
- Golovenko, D., Brauning, B., Vyas, P., Haran, T. E., Rozenberg, H., & Shakked, Z. (2018). New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins. *Structure*, 26(9), 1237-1250 e1236. doi:10.1016/j.str.2018.06.006
- Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., . . . Song, H. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci*, 17(2), 215-222. doi:10.1038/nn.3607
- Handa, V., & Jeltsch, A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *J Mol Biol*, 348(5), 1103-1112. doi:10.1016/j.jmb.2005.02.044

- Hervouet, E., Peixoto, P., Delage-Mourroux, R., Boyer-Guittaut, M., & Cartron, P. F. (2018). Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma. *Clin Epigenetics*, 10, 17. doi:10.1186/s13148-018-0450-y
- Holz-Schietinger, C., Matje, D. M., & Reich, N. O. (2012). Mutations in DNA methyltransferase (DNMT3A) observed in acute myeloid leukemia patients disrupt processive methylation. *J Biol Chem*, 287(37), 30941-30951. doi:10.1074/jbc.M112.366625
- Ivani, I., Dans, P. D., Noy, A., Perez, A., Faustino, I., Hospital, A., . . . Orozco, M. (2016). Parmbsc1: a refined force field for DNA simulations. *Nat Methods*, 13(1), 55-58. doi:10.1038/nmeth.3658
- Ivanov, S. M., Cawley, A., Huber, R. G., Bond, P. J., & Warwicker, J. (2017). Protein-protein interactions in paralogues: Electrostatics modulates specificity on a conserved steric scaffold. *PLoS One*, 12(10), e0185928. doi:10.1371/journal.pone.0185928
- Jurkowska, R. Z., Anspach, N., Urbanke, C., Jia, D., Reinhardt, R., Nellen, W., . . . Jeltsch, A. (2008). Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L. *Nucleic Acids Res*, 36(21), 6656-6663. doi:10.1093/nar/gkn747
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10(6), 845-858. doi:10.1038/nprot.2015.053
- 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 Res*, 48(7), 3949-3961. doi:10.1093/nar/gkaa111
- Lin, M., & Guo, J. T. (2019). New insights into protein-DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Res*, 47(21), 11103-11113. doi:10.1093/nar/gkz963
- Lu, X. J., & Olson, W. K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, 31(17), 5108-5121. doi:10.1093/nar/gkg680
- Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res*, 5, 189. doi:10.12688/f1000research.7931.1
- Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1), 23-38. doi:10.1038/npp.2012.112
- Morra, G., Verkhivker, G., & Colombo, G. (2009). Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full-length dimer. *PLoS Comput Biol*, 5(3), e1000323. doi:10.1371/journal.pcbi.1000323
- Nguyen, B., Neidle, S., & Wilson, W. D. (2009). A role for water molecules in DNA-ligand minor groove recognition. *Acc Chem Res*, 42(1), 11-21. doi:10.1021/ar800016q
- Nguyen, T. V., Yao, S., Wang, Y., Rolfe, A., Selvaraj, A., Darman, R., . . . Bolduc, D. M. (2019). The R882H DNMT3A hot spot mutation stabilizes the formation of large DNMT3A oligomers with low DNA methyltransferase activity. *J Biol Chem*, 294(45), 16966-16977. doi:10.1074/jbc.RA119.010126
- Nimura, K., Ishida, C., Koriyama, H., Hata, K., Yamanaka, S., Li, E., . . . Kaneda, Y. (2006). Dnmt3a2 targets endogenous Dnmt3L to ES cell chromatin and induces regional DNA methylation. *Genes Cells*, 11(10), 1225-1237. doi:10.1111/j.1365-2443.2006.01012.x
- Oka, M., Rodic, N., Graddy, J., Chang, L. J., & Terada, N. (2006). CpG sites preferentially methylated by Dnmt3a in vivo. *J Biol Chem*, 281(15), 9901-9908. doi:10.1074/jbc.M511100200

- Rajavelu, A., Jurkowska, R. Z., Fritz, J., & Jeltsch, A. (2012). Function and disruption of DNA methyltransferase 3a cooperative DNA binding and nucleoprotein filament formation. *Nucleic Acids Res*, 40(2), 569-580. doi:10.1093/nar/gkr753
- Ribeiro, J., Rios-Vera, C., Melo, F., & Schuller, A. (2019). Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions. *Bioinformatics*, 35(18), 3499-3501. doi:10.1093/bioinformatics/btz062
- Sankpal, U. T., & Rao, D. N. (2002). Structure, function, and mechanism of HhaI DNA methyltransferases. *Crit Rev Biochem Mol Biol*, 37(3), 167-197. doi:10.1080/10409230290771492
- Sinha, N., & Smith-Gill, S. J. (2002). Electrostatics in protein binding and function. *Curr Protein Pept Sci*, 3(6), 601-614. doi:10.2174/1389203023380431
- Suetake, I., Miyazaki, J., Murakami, C., Takeshima, H., & Tajima, S. (2003). Distinct enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b. *J Biochem*, 133(6), 737-744. doi:10.1093/jb/mvg095
- UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47(D1), D506-D515. doi:10.1093/nar/gky1049
- van Dijk, M., van Dijk, A. D., Hsu, V., Boelens, R., & Bonvin, A. M. (2006). Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res*, 34(11), 3317-3325. doi:10.1093/nar/gkl412
- Vasu, K., & Nagaraja, V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev*, 77(1), 53-72. doi:10.1128/MMBR.00044-12
- Veland, N., Lu, Y., Hardikar, S., Gaddis, S., Zeng, Y., Liu, B., . . . Chen, T. (2019). DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells. *Nucleic Acids Res*, 47(1), 152-167. doi:10.1093/nar/gky947
- Zhang, Z. M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., . . . Song, J. (2018). Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature*, 554(7692), 387-391. doi:10.1038/nature25477

APPENDIX A.

Slurm file to run MD simulations.

```
#!/bin/bash
#SBATCH -p akya-cuda
#SBATCH -A ddogan
#SBATCH -J cac
#SBATCH -N 1
#SBATCH -n 40
#SBATCH --gres=gpu:1
#SBATCH --time=7-00:00:00
#SBATCH --mail-type=ALL
#SBATCH --mail-user=deenizzdogan@gmail.com
#SBATCH --output=cac-%j.out
#SBATCH --error=cac-%j.err
source /etc/profile.d/modules.sh

source /truba/sw/centos7.3/comp/intel/PS2018-update2/bin/compilervars.sh intel64

module load centos7.3/comp/gcc/6.4
module load centos7.3/comp/cmake/3.10.1
module load centos7.3/app/gromacs/2020-imp1-mkl-PS2018-GOLD-CUDA
module load centos7.3/lib/cuda/10.0
module load centos7.3/lib/openmpi/1.8.8-gcc-4.8.5

export OMP_NUM_THREADS=20
echo "SLURM_NODELIST $SLURM_NODELIST"
echo "NUMBER OF CORES $SLURM_NTASKS"

set -e

# echo -e "1\n1\n0\n1\n0\n0\n0\n0\n0\n0\n" | gmx pdb2gmx -f MD1.pdb -o conf.gro -p
# topol.top -ter -his -ignh

# sed -i -e 's/1000 1000 1000/2500 2500 2500/g' posre*.itp

# gmx editconf -f conf.gro -o newbox.gro -bt dodecahedron -d 1.4 > log.editconf 2>&1
# gmx solvate -cp newbox.gro -cs spc216.gro -p topol.top -o solv.gro > log.solvation 2>&1
# gmx grompp -f 01_em_SD_WI_PME.mdp -c solv.gro -r solv.gro -p topol.top -o ion.tpr -
# maxwarn 1 > log.grompp.ion 2>&1
# echo -e "17\n" | gmx genion -s ion.tpr -o solv_ions.gro -p topol.top -pname K -nname CL -
# conc 0.15 -neutral > log.genion 2>&1

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 01_em_SD_WI_PME.mdp -c
solv_ions.gro -r solv_ions.gro -p topol.top -o ion_SD.tpr > log.grompp.ionsd 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm ion_SD > log.mdrun.ionsd 2>&1
```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 02_em_CG_WI_PME.mdp -c
ion_SD.gro -r ion_SD.gro -p topol.top -o ion_CG.tpr > log.grompp.ioncg 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm ion_CG > log.mdrun.ioncg 2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 03_em_SD_SYS_PME.mdp -c
ion_CG.gro -r ion_CG.gro -p topol.top -o SYS_SD.tpr > log.grompp.syssd 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm SYS_SD > log.mdrun.syssd 2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 04_em_CG_SYS_PME.mdp -c
SYS_SD.gro -r SYS_SD.gro -p topol.top -o SYS_CG.tpr > log.grompp.syscg 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm SYS_CG > log.mdrun.syscg 2>&1

```

```

gmx make_ndx -f SYS_CG.gro -o index.ndx<<EOD

```

```

1|12|13

```

```

q

```

```

EOD

```

```

$GROMACS_DIR/bin/gmx_mpi grompp -f 05_nvt_pr2500_PME.mdp -c SYS_CG.gro -p
topol.top -n index.ndx -o nvt_pr2500.tpr -r SYS_CG.gro > log.grompp.nvt2500 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm nvt_pr2500 -pin on > log.mdrun.nvt2500
2>&1

```

```

sed -i -e 's/2500 2500 2500/ 500 500 500/g' posre*.itp

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 01_em_SD_WI_PME.mdp -c
nvt_pr2500.gro -r nvt_pr2500.gro -p topol.top -o nvt_pr500.tpr > log.grompp.nvt500 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm nvt_pr500 -pin on > log.mdrun.nvt500
2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 02_em_CG_WI_PME.mdp -c
nvt_pr500.gro -p topol.top -r nvt_pr500.gro -o nvt_pr500_CG.tpr > log.grompp.nvt500cg
2>&1

```

```

$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr500_CG -v -pin on >
log.mdrun.nvt500cg 2>&1

```

```

sed -i -e 's/ 500 500 500/ 400 400 400/g' posre*.itp

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 01_em_SD_WI_PME.mdp -c
nvt_pr500_CG.gro -p topol.top -r nvt_pr500_CG.gro -o nvt_pr400_SD.tpr >
log.grompp.nvt400sd 2>&1

```

```

$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr400_SD -v -pin on >
log.mdrun.nvt400sd 2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 02_em_CG_WI_PME.mdp -c
nvt_pr400_SD.gro -p topol.top -o nvt_pr400_CG.tpr -r nvt_pr400_SD.gro >
log.grompp.nvt400cg 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr400_CG -v -pin on >
log.mdrun.nvt400cg 2>&1

```

```

sed -i -e 's/ 400 400 400/ 300 300 300/g' posre*.itp
srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 01_em_SD_WI_PME.mdp -c
nvt_pr400_CG.gro -p topol.top -o nvt_pr300_SD.tpr -r nvt_pr400_CG.gro >
log.grompp.nvt300sd 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr300_SD -v -pin on >
log.mdrun.nvt300sd 2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 02_em_CG_WI_PME.mdp -c
nvt_pr300_SD.gro -p topol.top -o nvt_pr300_CG.tpr -r nvt_pr300_SD.gro >
log.grompp.nvt300cg 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr300_CG -v -pin on >
log.mdrun.nvt300cg 2>&1

```

```

sed -i -e 's/ 300 300 300/ 200 200 200/g' posre*.itp
srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 01_em_SD_WI_PME.mdp -c
nvt_pr300_CG.gro -p topol.top -o nvt_pr200_SD.tpr -r nvt_pr300_CG.gro >
log.grompp.nvt200sd 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr200_SD -v -pin on >
log.mdrun.nvt200sd 2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 02_em_CG_WI_PME.mdp -c
nvt_pr200_SD.gro -p topol.top -o nvt_pr200_CG.tpr -r nvt_pr200_SD.gro >
log.grompp.nvt200cg 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr200_CG -v -pin on >
log.mdrun.nvt200cg 2>&1

```

```

sed -i -e 's/ 200 200 200/ 100 100 100/g' posre*.itp
srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 01_em_SD_WI_PME.mdp -c
nvt_pr200_CG.gro -p topol.top -o nvt_pr100_SD.tpr -r nvt_pr200_CG.gro >
log.grompp.nvt100sd 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr100_SD -v -pin on >
log.mdrun.nvt100sd 2>&1

```

```

srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 02_em_CG_WI_PME.mdp -c
nvt_pr100_SD.gro -p topol.top -o nvt_pr100_CG.tpr -r nvt_pr100_SD.gro >
log.grompp.nvt100cg 2>&1
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm nvt_pr100_CG -v -pin on >
log.mdrun.nvt100cg 2>&1

```

```

sed -i -e 's/ 100 100 100/ 50 50 50/g' posre*.itp

```

```
gmx make_ndx -f nvt_pr100_CG.gro -o index-nvt.ndx <<EOD
```

```
1|12|13
```

```
q
```

```
EOD
```

```
srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 06_npt_pr50_PME.mdp -c  
nvt_pr100_CG.gro -p topol.top -n index-nvt.ndx -o npt.tpr -r nvt_pr100_CG.gro >  
log.grompp.npt 2>&1
```

```
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm npt -pin on > log.mdrun.npt 2>&1
```

```
srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 07_npt_NOpr_PME.mdp -c npt.gro -p  
topol.top -n index-nvt.ndx -o npt_NOpr.tpr -r npt.gro > log.grompp.nptnopr 2>&1
```

```
$GROMACS_DIR/bin/gmx_mpi mdrun -v -deffnm npt_NOpr -pin on > log.mdrun.nptnopr  
2>&1
```

```
srun -n 1 $GROMACS_DIR/bin/gmx_mpi grompp -f 08_md_PME.mdp -c npt_NOpr.gro -p  
topol.top -n index-nvt.ndx -o md_0_1.tpr -r npt_NOpr.gro > log.grompp.prod 2>&1
```

```
$GROMACS_DIR/bin/gmx_mpi mdrun -deffnm md_0_1 -pin on > log.grompp.prodrun  
2>&1
```

```
exit
```

APPENDIX B

DNMT3A shows altered intradomain communications with respect to the DNA sequence. In the DNMT3A:CpG complex, signals are predominantly clustered around the SAM-binding subdomain, catalytic subdomain and target recognition subdomain. The greatest signaling intensity of DNMT3A:CpG is measured in the target recognition subdomain. The evaluation of cross communications is further revealed that the SAM-binding subdomain (residues 8-58) communicates with the rest of all subdomains in all cases. On the other hand, the signaling of DNMT3A:CAC/CAG are predominantly clustered into two subregions (SAM-binding domain and target recognition subdomain). The catalytic subdomain either move as in harmony with the target recognition subdomain or SAM binding subdomain. The signaling of DNMT3A:CAG is dominant to DNMT3A:CAC due to the altered cross-communication among residues. DNMT3A:CAG/CAC show similar communication propensity at the target recognition subdomain while decreased signaling intensity towards SAM binding domain is recorded in DNMT3A:CAC. Other differences are detected at the residues 90-115, 111-124, 143-153, 265-285, 295-315 and 315-325. Residues 90-115 defines the catalytic loop while residues 265-285 indicates the target recognition domain loop. Residues 143-153 are utilized to the 3' end of complementary strand while residues 315-325 contacts to both SAM and 5' end of complementary strand. Residues 111-124 are placed at the hydrophobic interface while residues 295-315 is in the hydrophilic interface. The signaling intensity of the catalytic loop is lowest in the DNMT3A:CAC. The signaling pattern of target recognition loop changes as depending on the DNA sequence. The signaling of residues 143-153 and residues 315-325 with the SAM-binding subdomain reach the greatest intensity in the DNMT3A:CAG. Residues 143-153 further communicates with the residue 103-123 in DNMT3A:CAG. The region of residues 143-153 is in close proximity to the 3' end of complementary strand while residues 315-325 is positioned towards the 3' end of target strand and residues 103-123 encompasses the DNMT3L binding region. The residues 93-113, catalytic loop, show the lowest communication level with the rest of DNMT3A:CAC complex.

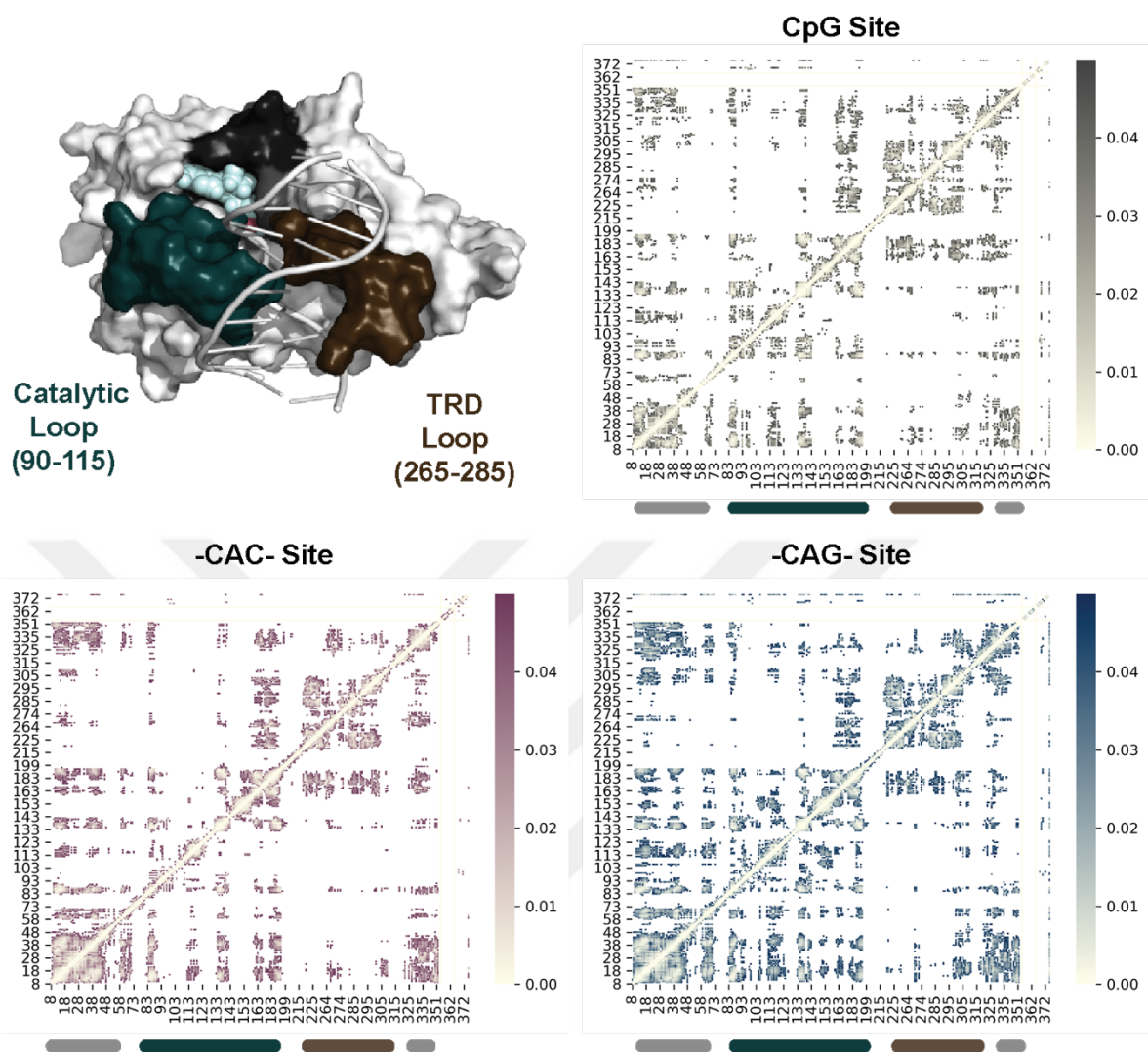


Figure 25: Signalling propensity of DNMT3A:DNA complexes.

APPENDIX C.

In the DNMT3B:DNAs, the intensity of signaling propensity are differentially clustered around the predefined four subdomains. DNMT3B:CpG complex shows effective communication within catalytic and target recognition subdomains. Compared to CpA sites, SAM binding regions have low signals in the DNMT3B:CpG. The SAM&DNA binding region and the catalytic subdomain communicate with the SAM-binding subdomain in DNMT3B:CpG. Among the CpA sites, catalytic subdomain might be further divided into the two parts: residues 90-115 and residues 133-199. The second part of catalytic subdomain is in harmony either with the SAM-binding region or with the target recognition subdomain. In the DNMT3B:CAC complex, the SAM-binding subdomain and catalytic subdomain acts as a united group and dominates the signaling map. On the other hand, the signaling intensity of target recognition subdomain is greater than the SAM-binding region. The second part of catalytic subdomain joins to the target recognition subdomain. The majority of difference between -CAC- and -CAG- sequence is acquired from the first part of catalytic subdomain (residues 90-115). The low signaling intensity is detected in the first half of catalytic subdomain in DNMT3B:CAG. On the other hand, DNMT3B:CAC complex has a greater intensity in the first half of catalytic subdomain. The SAM-binding domain of the DNMT3B:CAG further communicate with the second half of catalytic subdomain, target recognition subdomain and SAM&DNA binding region.

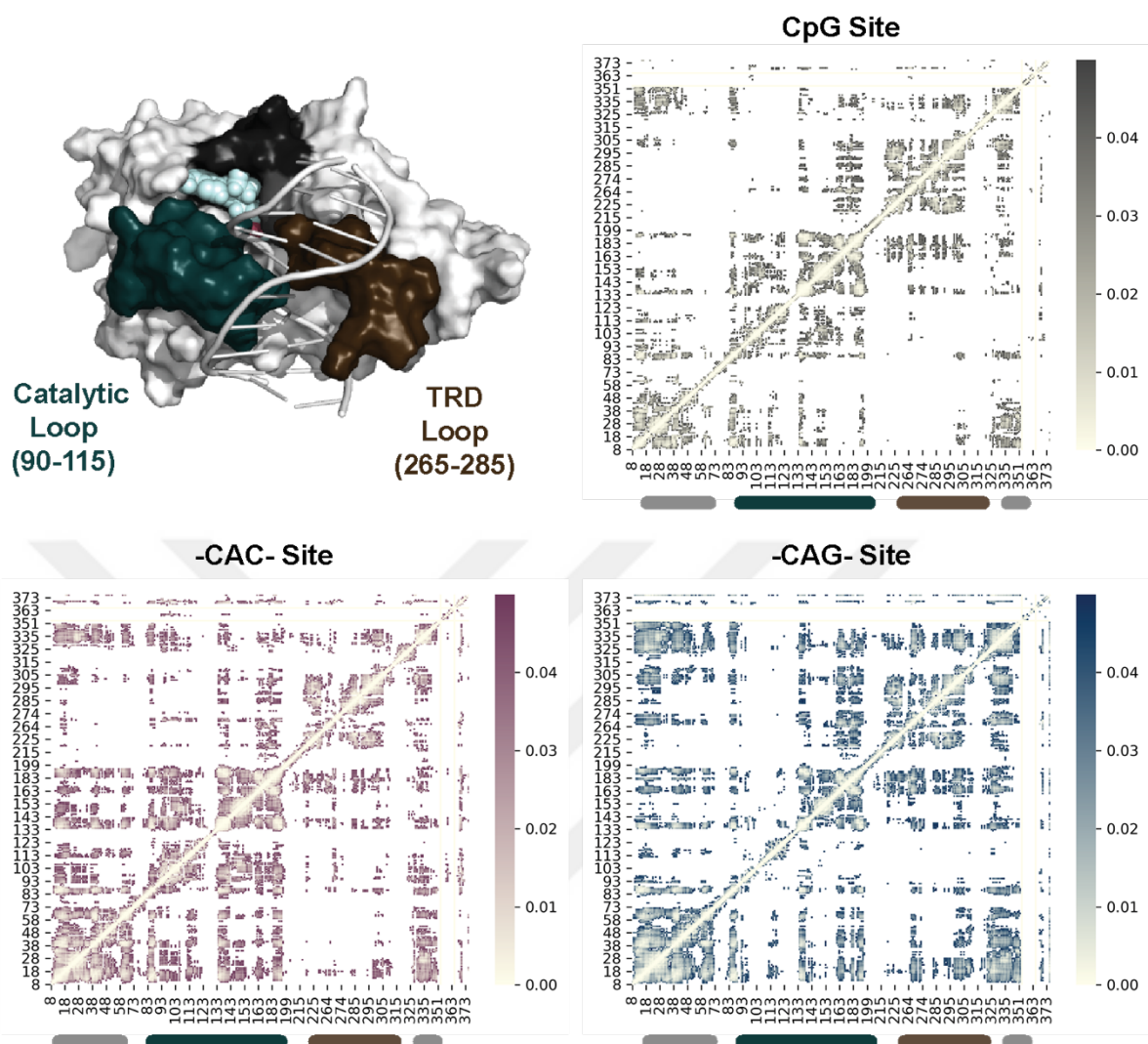


Figure 26: Signalling propensity of DNMT3B:DNA complexes.

Overall, we showed that the internal dynamics of DNMT3A/B enzymes are reshaped with the effect of DNA sequences. The signal propensities of DNMT3A/B:DNA complexes are clustered around either two (in the CpA sites) or three centers (in the CpG site). The lower signal propensities mediated by the catalytic loop correlates with the higher sequence specificity of DNMT3A/B enzymes. On the other hand, TRD-mediated signal propensity shows the drastic change only in DNMT3B:DNA complexes. In line with the previous results, the signal propensity analysis also supports the role of catalytic and TRD loops on the working principles of DNMT3A/B enzymes. However, these

results are not enough to explain the processivity of DNMT3A/B enzymes. On the other hand, the comparison of DNMT3A:CAC and DNMT3B:CAG complexes indicates that all subdomains of DNMT3B enzyme efficiently communicates with the SAM- and SAM&DNA binding subdomain unlike DNMT3A enzyme. These results imply that i) DNA sequence are crucial for the internal dynamics of DNMT3A/Bs, ii) the regulation of SAM is more crucial for the DNMT3B-mediated DNA methylation.



APPENDIX D.

Residue number at the PDB 6F57	Residue number at the MD	DNMT3A Residue name	DNMT3B Residue name
630	8	K	R
631	9	R	R
632	10	K	R
633	11	P	P
634	12	I	I
635	13	R	R
636	14	V	V
637	15	L	L
638	16	S	S
639	17	L	L
640	18	F	F
641	19	D	D
642	20	G	G
643	21	I	I
644	22	A	A
645	23	T	T
646	24	G	G
647	25	L	Y
648	26	L	L
649	27	V	V
650	28	L	L
651	29	K	K
652	30	D	E
653	31	L	L
654	32	G	G
655	33	I	I
656	34	Q	K
657	35	V	V
658	36	D	G
659	37	R	K
660	38	Y	Y
661	39	I	V
662	40	A	A

663	41	S	S
664	42	E	E
665	43	V	V
666	44	C	C
667	45	E	E
668	46	D	E
669	47	S	S
670	48	I	I
671	49	T	A
672	50	V	V
673	51	G	G
674	52	M	T
659	37	R	K
660	38	Y	Y
661	39	I	V
662	40	A	A
663	41	S	S
664	42	E	E
665	43	V	V
666	44	C	C
667	45	E	E
668	46	D	E
669	47	S	S
670	48	I	I
671	49	T	A
672	50	V	V
673	51	G	G
674	52	M	T
675	53	V	V
676	54	R	K
677	55	H	H
678	56	Q	E
679	57	G	G
680	58	K	N
681	59	I	I
682	65	M	K
683	66	Y	Y
684	67	V	V

685	68	G	N
686	69	D	D
687	70	V	V
688	71	R	R
689	72	S	N
690	73	V	I
691	74	T	T
692	75	Q	K
693	76	K	K
694	77	H	N
695	78	I	I
696	79	Q	E
697	80	E	E
698	81	W	W
699	82	G	G
700	83	P	P
701	84	F	F
702	85	D	D
703	86	L	L
704	87	V	V
705	88	I	I
706	89	G	G
707	90	G	G
708	91	S	S
709	92	P	P
710	93	C	C
711	94	N	N
712	95	D	D
713	96	L	L
714	97	S	S
715	98	I	N
716	99	V	V
717	100	N	N
718	101	P	P
719	102	A	A
720	103	R	R
721	104	K	K
722	105	G	G

723	106	L	L
724	107	Y	Y
725	108	E	E
726	109	G	G
727	110	T	T
728	111	G	G
729	112	R	R
730	113	L	L
731	114	F	F
732	115	F	F
733	116	E	E
734	117	F	F
735	118	Y	Y
736	119	R	H
737	120	L	L
738	121	L	L
739	122	H	N
740	123	D	Y
741	124	A	S
742	125	R	R
743	126	P	P
744	127	K	K
745	128	E	E
746	129	G	G
747	130	D	D
748	131	D	D
749	132	R	R
750	133	P	P
751	134	F	F
752	135	F	F
753	136	W	W
754	137	L	M
755	138	F	F
756	139	E	E
757	140	N	N
758	141	V	V
759	142	V	V
760	143	A	A

761	144	M	M
762	145	G	K
763	146	V	V
764	147	S	G
765	148	D	D
766	149	K	K
767	150	R	R
768	151	D	D
769	152	I	I
770	153	S	S
771	154	R	R
772	155	F	F
773	156	L	L
774	157	E	E
775	158	S	C
776	159	N	N
777	160	P	P
778	161	V	V
779	171	M	M
780	172	I	I
781	173	D	D
782	174	A	A
783	175	K	I
784	176	E	K
785	177	V	V
786	178	S	S
787	179	A	A
788	181	A	A
789	182	H	H
790	183	R	R
791	184	A	A
792	185	R	R
793	186	Y	Y
794	187	F	F
795	192	W	W
796	193	G	G
797	194	N	N
798	195	L	L

799	196	P	P
800	199	G	G
801	200	M	M
802	201	N	N
803	204	R	R
804	205	P	P
805	206	L	V
806	211	A	I
807	212	S	A
808	213	T	S
809	214	V	K
810	215	N	N
811	216	D	D
812	217	K	K
813	218	L	L
814	219	E	E
815	220	L	L
816	221	Q	Q
817	222	E	D
818	223	C	C
819	224	L	L
820	225	E	E
821	226	H	Y
822	227	G	N
823	228	R	R
824	229	I	I
825	230	A	A
826	231	K	K
827	261	F	L
828	262	S	K
829	263	K	K
830	264	V	V
831	265	R	Q
832	266	T	T
833	267	I	I
834	268	T	T
835	269	T	T
836	270	R	K

837	271	S	S
838	272	N	N
839	273	S	S
840	274	I	I
841	275	K	K
842	276	Q	Q
843	277	G	G
844	278	K	K
845	280	D	N
846	281	Q	Q
847	282	H	L
848	283	F	F
849	284	P	P
850	285	V	V
851	286	F	V
852	287	M	M
853	288	N	N
854	289	E	G
855	290	K	K
856	291	E	E
857	292	D	D
858	293	I	V
859	294	L	L
860	295	W	W
861	296	C	C
862	297	T	T
863	298	E	E
864	299	M	L
865	300	E	E
866	301	R	R
867	302	V	I
868	303	F	F
869	304	G	G
870	305	F	F
871	306	P	P
872	307	V	V
873	308	H	H
874	309	Y	Y

875	310	T	T
876	311	D	D
877	312	V	V
878	313	S	S
879	314	N	N
880	315	M	M
881	316	S	G
882	317	R	R
883	318	L	G
884	319	A	A
885	320	R	R
886	321	Q	Q
887	322	R	K
888	323	L	L
889	324	L	L
890	325	G	G
891	326	R	R
892	327	S	S
893	328	W	W
894	329	S	S
895	330	V	V
896	331	P	P
897	332	V	V
898	333	I	I
899	334	R	R
900	335	H	H
901	336	L	L
902	337	F	F
903	338	A	A
904	345	P	P
905	346	L	L
906	347	K	K
907	348	E	D
908	349	Y	Y
909	350	F	F
910	351	A	A
911	352	C	C

Deniz Doğan

Adress	Dokuz Eylul Health Campus, Balcova, Izmir, Turkey		+905438172803
Date of Birth	6 th February 1993		deenizzdogan@gmail.com
Nationality	Turkish		denizdgn

Education

2017 - Present	MSc in Molecular Biology and Genetics - Izmir Biomedicine and Genome Center Master level education touching on computational structural biology, molecular phylogenetics and programming. On track to graduate cum laude
2016 - 2017	MSc in Molecular Biology and Genetics - Izmir Institute of Technology Only lectures were taken.
2012 - 2016	BSc in Molecular Biology and Genetics - Izmir Institute of Technology
2011 - 2012	English Preparatory Class

Computer Skills

Programming Languages	Python (Pandas, Pathos, MDAnalysis, MDTraj, Interfacea, Networkx, Seaborn, Matplotlib, vmd-python), BASH, R
Software Tools	GROMACS, HADDOCK, git, github, HPC, Linux systems, Mac OS X, PyMOL, Adobe Photoshop, Adobe Illustrator, Adobe Indesign

Research Experience

Jan 2020 - Aug 2020	Izmir Biomedicine and Genome Center, Izmir, Turkey <i>Researcher, Computational Biophysics Lab</i> Investigating regulatory mechanism of chromosome segregation during cell division • Modelling, simulating and analyzing thirteen different combinations of nucleosomes • Writing a python-based commandline tool: NUCLEOSOME » <i>Programs used in this project: Gromacs, Python, Pymol, VMD</i>
Jan 2019 - Feb 2019	University of Groningen, Groningen, The Netherlands <i>HPC-Europa3 visitor, Molecular Dynamics Lab</i> Exploring sequence specificity of mammalian de novo DNA methyltransferases (DNMT3s) towards different DNA sequences. • Modelled and simulated six different complexes of DNMT3A/B bound to different methylation sites » <i>Programs used in this project: Haddock, Gromacs, Pymol, Python</i>

Research Experience

Sep 2017 - Present **Izmir Biomedicine and Genome Center, Izmir, Turkey**
MSc Student, Computational Structural Biology Lab

A structural insight into maintenance methylation

- Looked for sequence and structure conservation of DNMTs. Comparative analysis of modelled bacterial and mammalian DNMT1 complexes bound to CpG or hemi-methylated CpG sites after defining parameters of SAM (methyl donor ligand) and methylated cytosine.
- » Got familiarized with protein-DNA systems and computational approaches.

Investigating influence of a newly detected missense mutation on KIF21A's activity.

- Comparing the dynamic behavior of both mutant and WT complexes
- » Indicated how this mutation alters mechanistic principles of KIF21A's activity. This result might contribute to develop new therapeutic approaches.
- » Programs used in these projects: Haddock, Gromacs, Pymol, Python, Bash, R

Scripts are organized to provide a python-based commandline tool (MINNIE). MINNIE aids to yield comprehensive interaction profiles of MD trajectories. It also performs comparative analysis to acquire critical interactions.

Publications

Doğan, D., Arslan, M., Uluçay, T., Uzunlar, S.K., Dimitrov, S., and Kale, S. (2020) CENP-A nucleosome is a sensitive allosteric scaffold for DNA and chromatin factors. *Journal of Molecular Biology* (submitted)

Awarded Grant Proposals as Team Member

2019 **The Scientific and Technological Research Council of Turkey (TUBITAK 1002):** Structure-Based Computational Approach to Identify Molecular Principles of Human DNMT3A Enzyme

2019 **HPC-Europa3 Transnational Access programme:** Structural Investigation of *de novo* DNA Methyltransferase Sequence Specificity

Languages

Turkish (Native Language), English (Fluent), French (Beginner)

Interests

Travelling & Camping, Cooking, Yoga, Swimming, Cycling

Referees

Name	Ezgi Karaca	Name	Ignacio Faustino
Company	IBG	Company	Colotech Pharma
Position	Assistant Professor	Position	Managing Director
Contact	ezgi.karaca@ibg.edu.tr	Contact	ignacio.faustino@gmail.com