KADİR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
PROGRAM OF COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

IDENTIFYING THE WORKING PRINCIPLES OF HUMAN DNA METHYLTRANSFERASE 3A ENZYME BY COMPUTATIONAL METHODS

BÜŞRA SAVAŞ

MASTER’S THESIS

İSTANBUL, FEBRUARY, 2021

Büşra Savaş M.S. Thesis 2021

Identifying the Working Principles of Human DNA Methyltransferase 3A Enzyme by Computational Methods

Büşra Savaş

MASTER'S THESIS

Submitted to the School of Graduate Studies of
Kadir Has University in partial fulfillment of the requirements for the degree of
Master of Science in Computational Biology and Bioinformatics

İSTANBUL, February, 2021

DECLARATION OF RESEARCH ETHICS / METHODS OF DISSEMINATION

I, Büşra Savaş, hereby declare that;

- this master's thesis is my own original work and that due references have been appropriately provided on all supporting literature and resources;
- this master's thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- I have followed *Kadir Has University Academic Ethics Principles prepared in accordance with The Council of Higher Education's Ethical Conduct Principles.*

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

Furthermore, both printed and electronic copies of my work will be kept in Kadir Has Information Center under the following condition as indicated below (SELECT ONLY ONE, DELETE THE OTHER TWO):

□ The full content of my thesis will be accessible from everywhere by all means.

□ The full content of my thesis will be accessible only within the campus of Kadir Has University.

☒ The full content of my thesis will not be accessible for __1__ years. If no extension is required by the end of this period, the full content of my thesis will be automatically accessible from everywhere by all means.

Büşra Savaş
10.02.2021

KADİR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES

ACCEPTANCE AND APPROVAL

This work entitled Identifying the Working Principles of Human DNA Methyltransferase 3A Enzyme by Computational Methods prepared by Büşra Savaş has been judged to be successful at the defense exam on 10.02.2021 and accepted by our jury as master's thesis.

APPROVED BY:

Asst. Prof. Şebnem Eşsiz (Advisor) ....
Kadir Has University

Asst. Prof. Ezgi Karaca Erek (Co-advisor) ....
Dokuz Eylul University,
Izmir Biomedicine and Genom Center

Prof. Ebru Demet Akten ....
Kadir Has University

Prof. Kemal Yelekçi ....
Kadir Has University

Prof. Türkan Haliloğlu ....
Bogazici University

I certify that the above signatures belong to the faculty members named above.

....
Prof. Emine Füsun Alioğlu
Dean of School of Graduate Studies
DATE OF APPROVAL: 10.02.2021

TABLE OF CONTENTS

Identifying the Working Principles of Human DNA Methyltransferase 3A Enzyme by Computational Methods

ABSTRACT

DNA methylation is one of the epigenetic mechanisms in mammalians, responsible from maintenance and establishment of methylation pattern. DNA methyltransferase 3 (DNMT3) enzyme family modulates methylation from scratch and named as de novo methyltransferases. Two member of this family, DNMT3A and DNMT3B catalyze the reaction in between DNA and S-Adenosylmethionine (SAM). Although other member, DNMT3L, does not interact with DNA directly and catalytically inactive, stimulates of the reaction allosterically by binding to DNMT3A and DNMT3B. Additionally, efficiency of the reaction increases more with formation of DNMT3A:DNMT3L complex.

DNMT3A:DNMT3L complex is found as heterodimer, where hydrophobic interface is formed in between proteins, and heterotetramer including both hydrophobic and hydrophilic interfaces. In this study we studied the effect of these interfaces and DNMT3L on working principles of DNMT3A. To that end, PCA analysis is applied to detect allosteric effect of DNM3L. Additionally, protein-DNA interactions that occur during MD simulation are examined for both heterodimer and heterotetramer complexes.

Keywords: DNMT3A, DNMT3L, Molecular Dynamics, PCA, Protein-DNA Interactions

İnsan DNA Metiltransferaz Enziminin Çalışma Mekanizmasının Hesaplamalı Yöntemlerle Belirlenmesi

ÖZET

DNA metilasyonu, metilasyon deseninin sürdürülmesi ve oluşturulmasından sorumlu epigenetik mekanizmalardan biridir. DNA metiltransferaz 3 (DNMT3) enzim ailesi DNA'nın sıfırdan metilasyonunu düzenler bu nedenle de novo metiltransferazlar olarak adlandırılır. İki üyesi, DNMT3A ve DNMT3B, DNA ve S-Adenosilmetiyonin (SAM) arasındaki reaksiyonu katalize eder. Diğer üyesi olan DNMT3L ise DNA ile doğrudan etkileşime girmese de, DNMT3A ve DNMT3B'ye bağlanarak reaksiyonu allosterik olarak hızlandırır. Ek olarak, reaksiyonun etkinliğinin DNMT3A: DNMT3L kompleksinin oluşumu ile daha da arttığı görülmüştür.

DNMT3A: DNMT3L kompleksi, proteinler arasında hidrofobik arayüzün oluştuğu heterodimer ve hem hidrofobik hem de hidrofilik arayüzler içeren heterotetramer formunda olarak bulunur. Bu çalışmada, bu arayüzlerin ve DNMT3L'nin DNMT3A'nın çalışma prensipleri üzerindeki etkisini bulmayı hedefledik. Bu amaç doğrultusunda, DNM3L'nin allosterik etkisini tespit etmek için PCA analizini kullandık. Ek olarak, MD simülasyonu sırasında meydana gelen protein-DNA etkileşimleri hem heterodimer hem de heterotetramer kompleksleri için incelendi.

Anahtar Sözcükler: DNMT3A, DNMT3L, Moleküler Dinamik, PCA, Protein-DNA etkileşimleri

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Asst. Prof. Şebnem Eşsiz and co advisor Asst. Prof. Ezgi Karaca for their endless support and for their patience and motivation throughout this thesis. Their expertise in the field is inspired and motivated me to continue my academic career.

This work is arised as a result of findings in project supported by BIDEB2232 (1109B321700106) and HPC-EUROPA3 (INFRAIA-2016-1-730897) and supported by TUBITAK under the 1002 support program with the grant number 119Z828.

I would like to thank to TRUBA for offering opportunity to run molecular dynamics simulations that analyzed in this thesis. Additionally, I would like to thank Deniz Doğan who initiated this study.

I would also like to thank my close friends from high school Beyza Sağlam Özbaş and Dilara Uzuner for their endless patience, love and beautiful friendship.

Last but not least, I would like to express my gratitude to my father Osman Savaş, my mother Zülbiye Savaş, my beloved sister Neslihan Savaş Aksu and my little brother Ömer Faruk Savaş for their unconditional love. Their support helped me to succeed, overcome my stress and filled me with the confidence that I am capable of doing anything I put my mind to. Thank you for everything.

To those who will not read this

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Å	Angstrom
K	Kelvin
5-mC	5-methylcytosine
A	Adenine
A	Alanine, ALA
Acepype	AnteChamber PYthon Parser interfacE
C	Cytosine
Cα	Carbon Alpha Atom
CCor	Cross Correlation
Cl$^-$	Chloride ion
CpG	Cytosine-phosphate-Guanine
DNMT	DNA Methyltransferase
DNMT3A	DNA Methyltransferase 3A Enzyme
DNMT3L	DNA Methyltransferase 3-Like
E	GLU, Glutamic acid
G	Guanine
GROMACS	Groningen Machine for Chemical Simulation
H	Histidine, HIS
H	Hydrogen
HADDOCK	High Ambiguity Driven Biomolecular DOCKing
HPC	High Performance Computing
it0	Rigid Body Docking
it1	Semi Flexible Simulated Annealing
K$^+$	Potassium ion
kcal	Kilo calorie
MD	Molecular Dynamics
Mindist	Minimum Distance Between Periodic Images
minnie	Molecular INteractioN fIngErprints

nm	Nanometer
ns	Nanosecond
NPT	Constant Number of Particles, Pressure and Temperature
NVT	Constant number of particles, volume and temperature
PCA	Principal Component Analysis
PCs	Principal Components
PDB	Protein Data Bank
PDS	Protein-DNA-SAM
ps	Picosecond
R	Arginine, ARG
R_g	Radius of Gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SAH	S-Adenosyl-L-homocysteine
SAM	S-Adenosylmethionine
T	Thymine
TRUBA	Turkish National Science e-Infrastructure

1. INTRODUCTION

1.1 Epigenetic: On Top of Genetics

In biology, epigenetic explains the heritable changes in gene expression that cannot be explained by the changes in DNA sequence. This term was firstly introduced by Waddington during early 1940's (Waddington, 2011). Alterations in gene expression determines when, where and how long a gene will work. Additionally, epigenetic mechanisms can also dictate cell differentiation to determine which cells will end up as muscle cells, pancreatic cells, and blood cells etc. Therefore, significant alterations on epigenetic mechanisms can lead to detrimental outcomes, such as over expression of a gene, leading to cancer formation. In addition to cancer, aberrant epigenetic mechanisms are associated to several diseases such as intellectual disability, immunodeficiency, anemia, obesity and organ overgrowth (Au, Eaton and Dyment, 2020; Zeng et al., 2020; Dalfrà et al., 2020). The change in epigenetic mechanisms can be induced by various factors such as environment, development process, oxidative stress, drugs, aging, lifestyle and diet. DNA methylation is one of the most studied epigenetic mechanisms in mammalians, besides histone modifications.

1.2 DNA Methylation

In the bacterial world, methylation of DNA affects survival processes such as protection of the genome from bacteriophages, phase variation and antibiotic resistance (Casadesús and Low, 2006; Zhu 2009). Similar to the bacteria, DNA methylation is used as a defense mechanism to suppress the genes belong to retroviruses and transposons in eukaryotes (Sánchez-Romero and Casadesús, 2020). In mammalians DNA methylation plays a crucial role in biological processes, such as gene and transposon silencing, cell differentiation, gametogenesis, genomic imprinting and X chromosome silencing (Law and Jacobsen 2010; Kulis and Esteller 2010). Dysregulation of DNA methylation could result in cancer, developmental defects, and eventually death (Robertson 2005). Since DNA methylation serves as a regulator of transcription and contributes diseases such as cancer, it still attracts a lot of attention from academical researchers.

DNA methylation is a chemical reaction leading to the formation of a covalent bond between DNA and a methyl group. Specifically, the methyl group binds to 5' carbon of the target cytosine ring which is converted to 5-methylcytosine (5-mC). Methyl donor of methylation process, S-Adenosylmethionine (SAM), turns into S-Adenosyl-L-homocysteine (SAH) after the reaction (Figure 1.1).

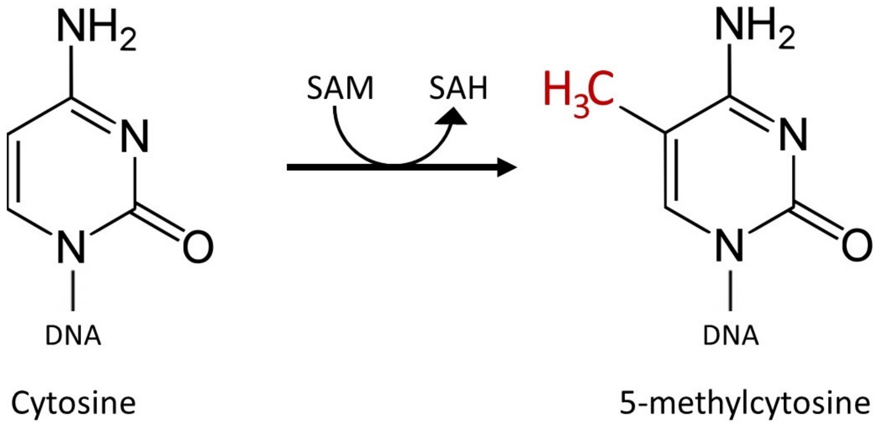


Figure 1.1 Representation of methylation reaction

DNA methylation is categorized by the sequence difference after the methylated cytosine. The most abundant methylation mark is CpG (5'-Cytosine-phosphate-Guanine-3'). The other methylation sequences are classified as non-CpG, where the neighboring nucleotide can be A, T or C. CpG dinucleotides are methylated in the human genome up to %80 (Tost, 2009).

1.3 Human DNA Methyltransferases

DNA methylation reaction is catalyzed by DNA methyltransferase (DNMT) enzymes. Until today, five DNMTs have been found in mammalians; DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L. All DNMTs have highly conserved catalytic domain in their C terminal, which directs the reaction (Figure 1.2)(Chen and Riggs, 2011).

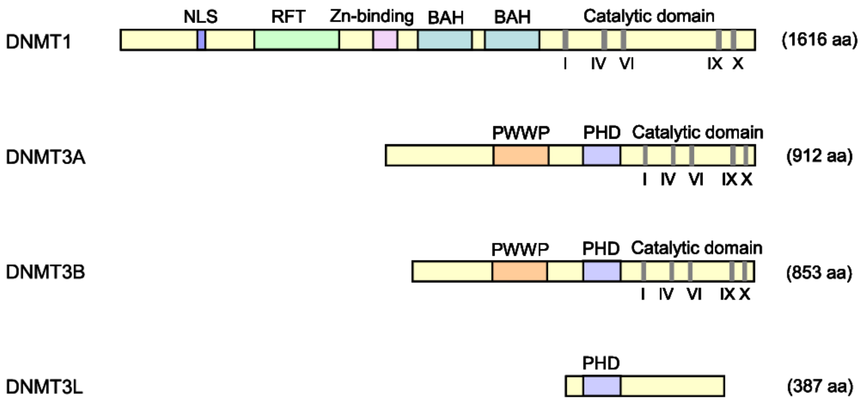


Figure 1.2 Domains in human DNMTs (Chen and Riggs, 2011)

Although DNMT2 was thought to be a DNA methyltransferase enzyme, it was found that it acts as RNA methyltransferase (Ashapkin, Kutueva and Vanyushin, 2016). Among the others, DNMT1 is associated with maintenance of methylation pattern, which means it only methylates the hemi-methylated (half-methylated) DNA after replication (Jeltsch, 2006).

DNMT3A, DNMT3B and DNMT3L enzymes belong to DNMT3 enzyme family and known as de novo methyltransferases, since they methylate DNA from scratch (ab initio). DNMT3s are responsible for establishing methylation pattern during early stages of development (Chédin, 2011). Additionally, they have been associated with maintaining of methylation pattern (Chen and Riggs, 2011).

Although DNMT3A and DNMT3B enzymes share high sequence and structural similarity, they tend to methylate different regions in the human genome (Chen and Riggs, 2011). Last member of DNMT3s, DNA methyltransferase-like protein

(DNMT3L), is seen mostly in the germ cells. DNM3L is catalytically inactive due to lack of essential catalytic motifs (Figure 1.2). Although DNMT3L does not interact with DNA and SAM directly, it still accelerates the methylation reaction upon binding to DNMT3A and DNMT3B. It has been shown that the effect of DNMT3L on DNMT3A is much higher compared to DNMT3B (Holz-Schietinger et al., 2011;Gowher and Jeltsch, 2018). Therefore, we focused on understanding the effect of DNMT3L on DNMT3A at molecular level.

1.4 Structures of DNMT3A and DNMT3L

The first DNMT3A:DNMT3L complex, will be referred as 3A:3L for simplicity, structure was resolved in 2007 with 2.89 Å resolution with X-ray crystallography (PDB ID:2QRV)(Jia et al., 2007). 2QRV complex is in heterotetramer form, where two DNM3A monomers face each other, while interacting with DNMT3L. The complex is in apo state, meaning that it does not contain the substrate DNA. In this complex, DNMT3L structure is bound unmethylated to a histone tail, illuminating the link between different epigenetic mechanisms. After 7 years, two 3A:3L structures were published by Guo and colleagues with 3.82 Å and 2.90 Å resolutions. (PDB IDs: 4U7P,4U7T) (Guo et al., 2015). Both of the complexes are in apo forms and found in heterodimeric states, as they contain two different complex (DNMT3A and DNMT3L). Additional to catalytic domain, both structures have also the ADD domains, where in 4U7T ADD is bound to , Histone 3 tail.

First 3A:3L, structure bound to DNA is resolved at 2.65 Å resolution in the heterotetramer form (PDB ID:5YX2)(Zhang et al., 2018). Here, the DNA structure contains two CpG sites, which are occupied by two different DNMT3A enzymes. Later on, the same group also published 6BRR and 6F57 structures with 2.97 and 3.10 Å resolution respectively. 6BRR structure shares the same stoichiometric form with 5YX2. Also, R836A DNMT3A mutation is present across the 3A:3A hydrophilic interface which forms by binding of two DNMT3A structures (Figure 1.3C). 6F57 complex is found in heterodimer form with a CpG containing DNA. In 2020, four structures were published in the heterotetramer form with both CpG and non-CpG DNAs (PDB IDs: 6W8J, 6W8B, 6W8D and 6W89)(Anteneh, Fang and Song, 2020). Except 6W8B structure, all three complexes have R882H mutation across their hydrophilic interfaces. All mentioned structures contain C terminal of the proteins as methylation reaction takes place in catalytic domain of DNMT3A.

Our goal in this thesis is to study the effect of DNMT3L and both hydrophilic and hydrophobic interfaces on the working mechanism of DNMT3A at molecular level by using computational methods. Therefore, we chose two different structures to investigate. First one is 6F57 structure to understand the effect of hydrophobic interface in between DNMT3A and DNMT3L (Figure 1.3A). The other one is 5YX2 structure to illuminate the effect of hydrophilic interface by comparing these two complexes (Figure 1.3C). Each complex is simulated for three times. After quality control of each replica simulation, we applied principal component analysis to understand allosteric effect of DNMT3L. Moreover, interfacea package is used for tracing interaction profile of complexes.

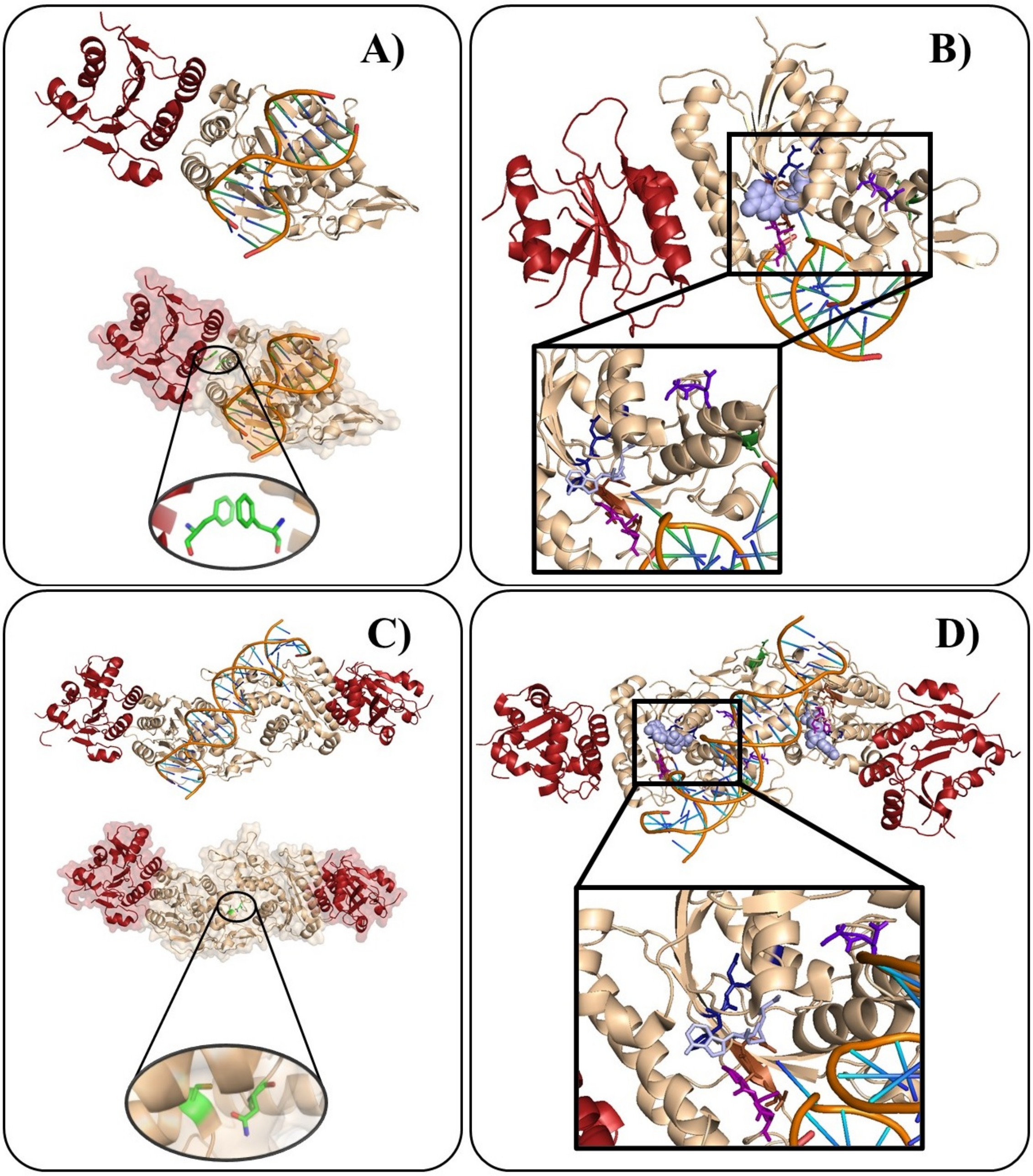


Figure 1.3 Formation of different oligomeric states in 3A:3L. A) Formation of heterodimer by 3A:3L binding with cartoon representation. DNMT3A is shown in wheat and dark red represents DNMT3L. The hydrophobic interface between 3A:3L is shown in surface representation. B) Essential catalytic motifs (I, IV, VI, XI and X) are shown in sticks on dimer structure with dark blue, magenta, brown, green and purple, respectively. SAM molecule is represented with both stick and ball representations in light blue color. C) The depiction of hydrophilic interface in between DNMT3A:DNMT3A and two hydrophobic interfaces in between 3A:3L both in cartoon and surface representations. D) Essential motifs on tetramer structure.

2. METHODS

2.1 Structure Preparation

Each component is examined to check the full length of protein is present in PDB structures. Then missing atoms in each complexes are completed as we explained in detail in the next sections. Obtained structures are represented in Figure 2.1.

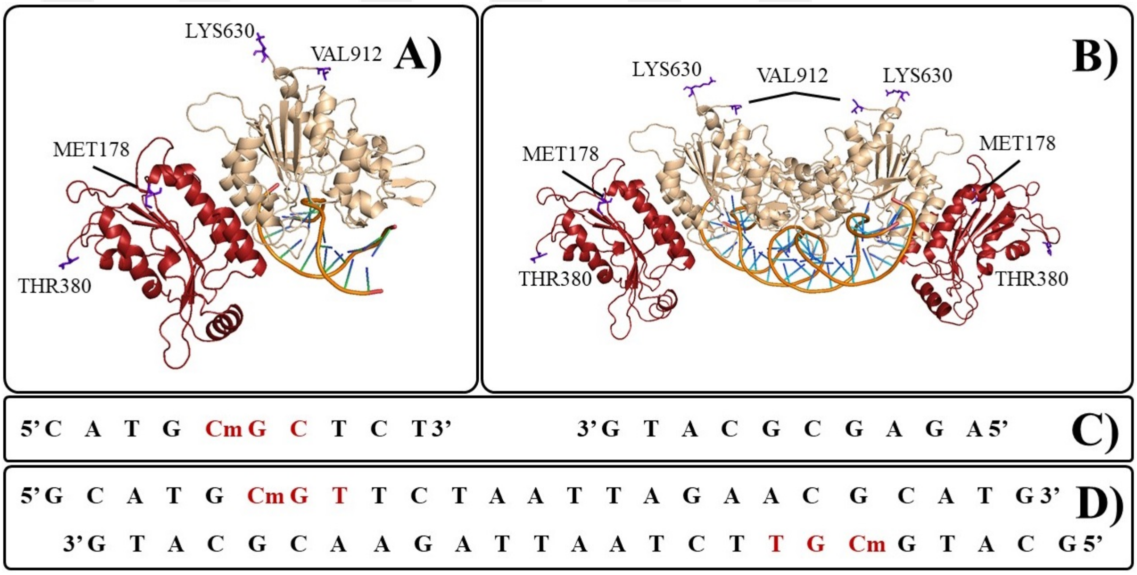
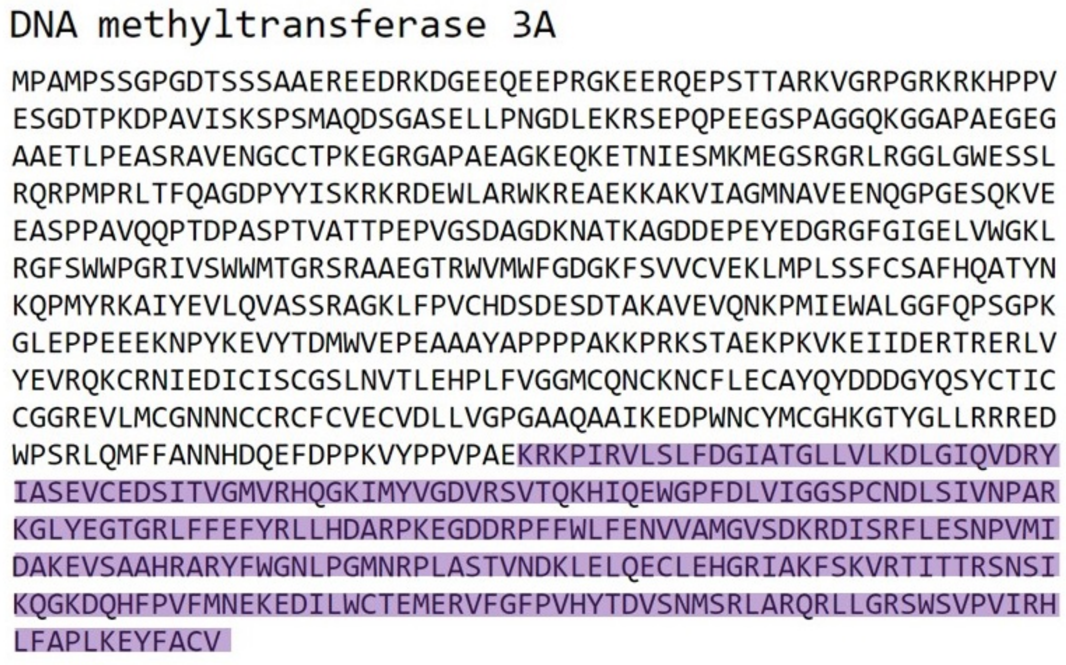


Figure 2.1 Whole structure of dimer and tetramer structures. A) N and C terminals of each protein for dimer used for simulations is shown in purple sticks. B) N and C terminals on tetramer structure. C) DNA sequence in dimer where CpGpX (X=C, T) is colored with red and Cm indicates flipped cytosine. D) Two strands of DNA in tetramer structure.

2.1.1 Preparation of DNMT3A

The heterodimer 6F57 complex is complexed with a 10 base pairs long CpG DNA, where the target cytosine is found in a pre-reaction flipped form (PDB ID: 6F57). The heterotetramer 5YX2 complex contains 25 base pairs long DNA, spanning two CpG sites with two flipped cytosines (PDB ID: 5YX2). Amino acids that does not interact with any amino acid within the protein and could move more and suppress the movement of important parts are discarded, A (ALA) and E (GLU) residues in N terminal. Whole sequence of DNMT3A probed in this thesis is shown in Figure 2.2 and the DNA sequences used in this work are provided in Figure 2.1.



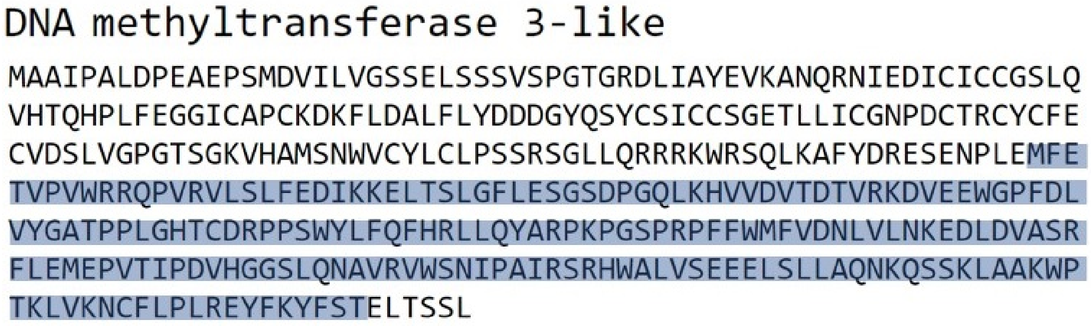
DNA methyltransferase 3A

```
MPAMPSSGPGDTSSSAAEREEDRKDGEEQEEPRGKEERQEPSTTARKVGRPGRKRKHPPV
ESGDTPKDPAVISKSPSMAQDSGASELLPNGDLEKRSEPQPEEGSPAGGQKGGAPAEGEG
AAETLPEASRAVENGCCTPKEGRGAPAEAGKEQKETNIESMKMEGSRGRLRGGLGWESSL
RQRPMPRLTFQAGDPYYISKRKRDEWLARWKREAEKKAKVIAGMNAVEENQGPGESQKVE
EASPPAVQQPTDPASPTVATTPEPVGSDAGDKNATKAGDDEPEYEDGRGFGIGELVWGKL
RGFSWWPGRIVSWWMTGRSRAAEGTRWVMWFGDGKFSVVCVEKLMPLSSFCSAFHQATYN
KQPMYRKAIYEVLQVASSRAGKLFPVCHDSDESDTAKAVEVQNKPMIEWALGGFQPSGPK
GLEPPEEEKNPYKEVYTDMWVEPEAAAYAPPPPAKKPRKSTAEKPKVKEIIDERTRERLV
YEVRQKCRNIEDICISCGSLNVTLEHPLFVGGMCQNCKNCFLECAYQYDDDGYQSYCTIC
CGGREVLMCGNNNCCRCFCVECVDLLVGPGAAQAAIKEDPWNCYMCGHKGTYGLLRRRED
WPSRLQMFFANNHDQEFDPPKVYPPVPAEKRKPIRVLSLFDGIATGLLVLKDLGIQVDRY
IASEVCEDSITVGMVRHQGKIMYVGDVRSVTQKHIQEWGPFDLVIGGSPCNDLSIVNPAR
KGLYEGTGRLFFEFYRLLHDARPKEGDDRPFFWLFENVVAMGVSDKRDISRFLESNPVMI
DAKEVSAAHRARYFWGNLPGMNRPLASTVNDKLELQECLEHGRIAKFSKVRTITTRSNSI
KQGKDQHFPVFMNEKEDILWCTEMERVFGFPVHYTDVSNMSRLARQRLLGRSWSVPVIRH
LFAPLKEYFACV
```

Figure 2.2 Whole sequence of DNMT3A. The sequence used in simulations emphasized with purple color.

2.1.2 Preparation of DNMT3L

As there were unresolved structural parts of DNMTL in both 6F57 and 5YX2 complexes, we replaced these DNMT3L coordinates with their full DNMT3L versions, resolved by Ooi and colleagues (PDB ID: 2PV0) (Ooi et al., 2007). In detail, 2PV0 structure consists three chains of DNMT3L with a longer sequence range than DNMT3Ls in 6F57 and 5YX2. The full DNMT3L sequence used in this work is provided in Figure 2.3.



DNA methyltransferase 3-like

```
MAAIPALDPEAEPSMDVILVGSSELSSSVSPGTGRDLIAYEVKANQRNIEDICICCGSLQ
VHTQHPLFEGGICAPCKDKFLDALFLYDDDGYQSYCSICCSGETLLICGNPDCTRCYCFE
CVDSLVGPGTSGKVHAMSNWVCYLCLPSSRSGLLQRRRKWRSQLKAFYDRESENPLEMFE
TVPVWRRQPVRVLSLFEDIKKELTSLGFLESGSDPGQLKHVVDVTDTVRKDVEEWGPFDL
VYGATPPLGHTCDRPPSWYLFQFHRLLQYARPKPGSPRPFFWMFVDNLVLNKEDLDVASR
FLEMEPVTIPDVHGGSLQNAVRVWSNIPAIRSRHWALVSEEELSLLAQNKQSSKLAAKWP
TKLVKNCFLPLREYFKYFSTELTSSL
```

Figure 2.3 Whole sequence of DNMT3L. The sequence used in simulations emphasized with blue color.

ELTSSL sequence in C terminal is discarded to reduce motion. The missing atoms and mismatch in DNTM3L at 278th residue are completed and corrected with HADDOCK refinement, which will be introduced in section 2.2.

2.1.3 Preparation of SAM

6F57 and 5YX2 structures contain a SAH molecule which is the processed SAM molecule (missing one CH3 group). To create a realistic scenario, we replaced SAH molecule with SAM, obtained from a X-ray crystal structure of mouse DNMT1 published by Takeshita and colleagues (PDB ID:3AV6)(Takeshita et al., 2011).

2.2 Structure refinement with HADDOCK server

To complete missing atoms, the heterodimer and heterotetramer complexes were refined with HADDOCK 2.2 (High Ambiguity Driven Biomolecular DOCKing). HADDOCK is a free web service designed for flexible docking of molecules (Zundert et al., 2016). It can take proteins, nucleic acids, and small molecules as an input. HADDOCK is composed of three steps, i.e., it0, it1 and water refinement (Vries, van Dijk and Bonvin, 2010). In this work, we used the water refinement option of HADDOCK.

- Before refinement, the missing atoms of the initial structure are completed by using a conformer optimization protocol. The processed structures are then resolved in TIP3P water model, where it is subjected to a series of very short molecular dynamic (MD) simulations. As a result of this refinement process, top four scoring structures are provided to the user. The water refinement process enables to improve the interaction energy of the inter-molecular interfaces (Dominguez, Boelens and Bonvin, 2003).

Both dimer and tetramer complexes were submitted as multi-bodies, containing four and six monomers, respectively. The number of MD simulations (sampling size) was set to 200. To preserve the symmetric nature of the complexes, C2 and non-crystallographic symmetry restraints were applied during refinement (Karaca et al., 2010). After the HADDOCK refinement each complex is renumbered to match a previous study's number system (Dogan, 2020). The numbering scheme is provided in Appendix A and B.

The first realistic molecular dynamics simulation was carried out for liquid argon in 1964 (Rahman, 1964). The MD simulations allow us to examine the movements of biological molecules at a given temperature, pressure, and ion concentration. There are a number of simulation packages available for running MD simulations. GROMACS, NAMD, CHARMM and AMBER packages are widely used for this purpose (Páll et al., 2015; Abraham et al., 2015; Phillips et al., 2005; Pearlman et al., 1995). Among these, GROMACS (Groningen Machine for Chemical Simulation) is an open-source package, offering the use of various force fields, such as AMBER, CHARMM, GROMOS and OPLS (Abraham et al., 2015).

2.3 Molecular Dynamics Simulation

Molecular dynamics (MD) algorithms were developed towards the end of the 19th century, began to be widely used due to the rapid development in computer technologies (Figure 2.4). MD simulations use the Newtonian equation of motion, therefore primarily used in the field of theoretical physics, then in material science and biology (Alder and Wainwright, 1959).

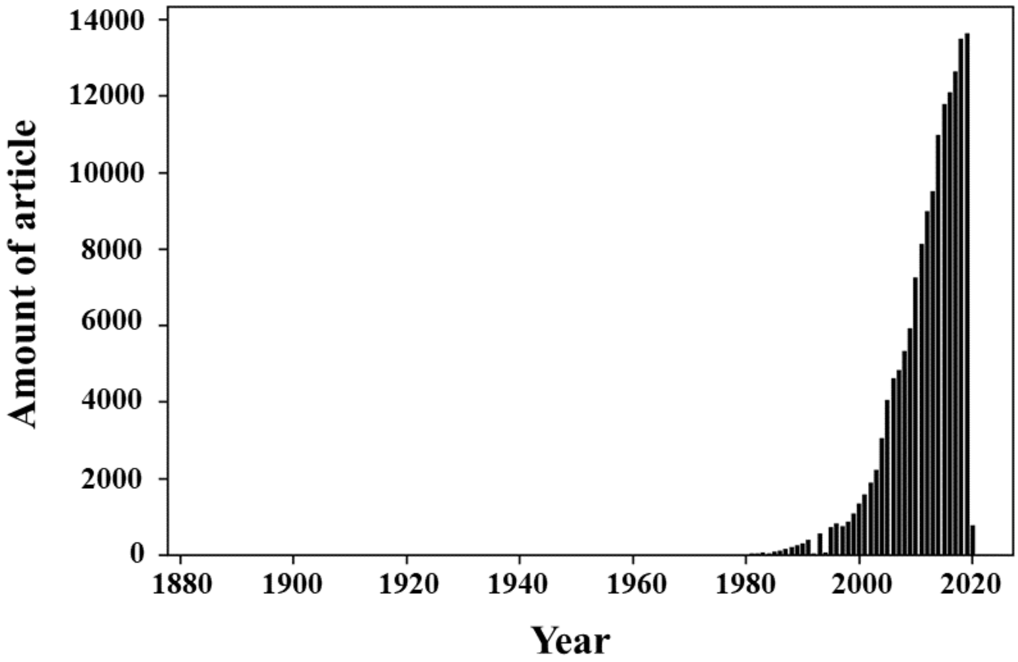


Figure 2.4 Number of articles containing molecular dynamics keyword in PubMed database by years.

The MD calculations are often applied at the atomic level, therefore they computationally challenge the current informatics technology. Thus, using high performance computing (HPC) systems are required for faster MD calculations. Before simulating the DNMT systems, we focused on the performance of MD simulations in TRUBA.

2.3.1 Performance optimization on TRUBA

TRUBA, Turkish National e-Science e-Infrastructure, formerly named as TrGrid is used for performing MD calculations. TRUBA is provided by TUBITAK ULAKBIM and serves more than 1500 researchers with 19000 CPU and 36 GPU. For performance optimization, we chose 3 different clusters in TRUBA which contain a GPU card, namely levrekv2-cuda, barbun-cuda and akya-cuda. Amount of CPU and GPU used can be found in Table 2.1 and in the Wiki page of TRUBA (http://wiki.truba.gov.tr)

Table 2.1 Detailed information about TRUBA clusters

Cluster	# of CPU	CPU Model	# of GPU	GPU Model
levrekv2-cuda	24	Intel Xeon E5- 2680 v3	2	M2090
barbun-cuda	40	Intel Xeon Scalable 6148	2	P100
akya-cuda	40	Intel Xeon Scalable 6148	4	V100

In our benchmarking efforts, two different versions of GROMACS were examined, namely GROMACS 5.1.4 and GROMACS 2020. Additionally, we aimed to understand the effect of using different energy groups such as Protein-DNA-SAM (PDS) and the whole system. A set of 12 simulation were designed to understand the impact of different system and infrastructure variables (Table 2.2). For this, each benchmarking simulation was run up to one nanosecond (ns).

Table 2.2 Simulation set for performance optimization on TRUBA

Version	GROMACS version	Cluster	CPU/GPU	Energy Group
1	5.1.4	levrekv2-cuda	24/1	PDS
2	5.1.4	levrekv2-cuda	24/1	system
3	5.1.4	akya-cuda	40/1	PDS
4	5.1.4	akya-cuda	40/1	system
5	2020	akya-cuda	40/1	PDS
6	2020	akya-cuda	40/1	system
7	2020	akya-cuda	40/2	system
8	2020	akya-cuda	40/3	system
9	2020	akya-cuda	40/4	system
10	2020	barbun-cuda	40/1	PDS
11	2020	barbun-cuda	40/1	system
12	2020	barbun-cuda	40/2	system

According to performance results, we decided to use akya-cuda cluster with GROMACS 2020. CPU/GPU ratio was chosen to be 40/1. The results of performance optimization is shared over Github and will be discussed in section 3.1.

2.3.2 Simulation parameters

The first stage of MD preparation is choosing force field and water type to be used. In this thesis amber PARMBSC1 force field was used, since considered as the best force field to sample protein-DNA dynamics (Ivani et al., 2016).

After generation of topology files for protein DNA structures, 3’ and 5’ of DNA chains were specified by adding 3 or 5 to residue name. As cofactor of the reaction, SAM, was not recognizable by the force field. We used topology files created with acepype for SAM by Deniz Doğan. Acepype, AnteChamber PYthon Parser interfacE, is a python package for generating topology files (Silva and Vranken, 2012; Batista et al., 2006). The topology information for SAM was amended to the topology file of protein and DNA manually.

The whole system was then placed in a dodecahedron periodic box with 1.4 nm diameter. The energy of the system was minimized in vacuum. The system was then solvated in TIP3P water molecules, by adding periodic boundary conditions. The system was then energy minimized with steepest descent minimization algorithm by 5000 steps. To neutralize the system K$^+$ and Cl$^-$ ions were added to solution at a concentration of 0.15 mol/L. At the end, the dimer simulations contained 141 K$^+$ and 131 Cl$^-$. The tetramer simulations contained 415 K$^+$ and 383 Cl$^-$. Total numbers of atoms were 142963 and 416472 for dimer and tetramer simulations, respectively.

The system was simulated by gradually decreasing harmonic restraints (25, 5, 4, 3, 2 and 1 kcal/mol/ Å). For each restraint 2 ns simulation time was set, where step size was defined as 1000 (2 ps) at 310K and constant volume, NVT. Then, system was simulated for 2 ns with 1000 steps (2 ps) at 310K and 1 atm, NPT, with (1, 0.50 and 0.10 kcal/mol/ Å, respectively). After that, all restraints were removed. and the system was simulated for 20 ns with 10000 steps (2 ps) before the production run. Each simulation complex was replicated with three different random seeds and were run for 500 ns (Table 2.3).

Table 2.3 Dimer and tetramer simulations in detail

Version	Random Seed	Ions	Temperature
dimer_v1	1716	KCl	310K
dimer_v2	2576	KCl	310K
dimer_v3	1257	KCl	310K
tetramer_v1	1716	KCl	310K
tetramer_v2	2576	KCl	310K
tetramer_v3	1257	KCl	310K

2.4 Quality control with GROMACS tools (ensures integrity of simulations)

The quality control of the simulations was checked with GROMACS tools. Here, minimum distance between periodic images (Mindist), root mean square deviation (RMSD), radius of gyration (R_g), and root mean square fluctuation (RMSF) analyses were used.

2.4.1 Minimum Distance Between Periodic Images to check if structures conflict themselves

During the simulation, structures move inside the periodic box. If the molecules move to the neighboring periodic box, a virtual jump is observed in the trajectory. Therefore, before performing any analysis, the molecules should be gathered in the centered box. To checked whether molecules structures see their periodic images, the minimum distance between periodic images (mindist) should be calculated. Mindist (p, r) is defined as the shortest distance from the point p to the box side r. The example of the mindist command is given below.

```
$gmx mindist -f dimer.xtc -s dimer.tpr -od mindist.xvg -pi
```

2.4.2 Radius of gyration to check structure maintains its form

R_g is defined as the distance from the rotation axis of the structure to the center of mass. It describes the equilibrium conformation of a whole system and is calculated in two steps. In the first step, central coordinates of R_c is determined without considering H atoms (only heavy atoms).

$$\sum m_i(r_i - R_c) = 0 \tag{2.1}$$

In equation 2.1, m_i describes the mass of i^{th} atom and r_i gives the coordinates of i^{th} atom. In 3D the equation becomes 2.2.

$$R_g^2 = \sum m_i(r_i - R_c)^2/M \tag{2.2}$$

M stands for total mass of atoms in proteins. For an equal mass system, we obtain equation 2.3.

$$R_g^2 = \sum_{i=1}^{N}(r_i - R_c)^2/N \tag{2.3}$$

where N is the total number of atoms in protein except H atoms. R_g gives us information about compactness of structures. For example, in the case of unfolding of protein R_g value gets higher. Radius of gyration can be calculated with gmx gyrate command:

```
$gmx gyrate −f dimer.xtc −s dimer.tpr −o gyration.xvg
```

2.4.3 Root Mean Square Deviation

Root mean square deviation enables to compare a conformation to a reference state. In this thesis, initial or average structures were used as the reference state. With gmx rms tool, a group of desired atoms is chosen, such as backbone atoms or all atoms. RMSD can be calculated by equation 2.4.

$$RMSD(t) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i|r_i(t) - r_i^{ref}|^2\right]^{1/2} \tag{2.4}$$

M is total mass, $\mathrm{r_i(t)}$ is coordinates of $\mathrm{i^{th}}$ atom at t time. $\mathrm{r_i{}^{ref}}$ is coordinates of $\mathrm{i^{th}}$ atom belonging to the reference structure. An example of command for calculating RMSD from initial and average structures:

```
$gmx rms -f dimer.xtc -s dimer.tpr -o rmsd_inital.xvg
$gmx rms -f dimer.xtc -s dimer_avg.pdb -rmsd_average.xvg
```

2.4.4 Root Mean Square Fluctuations to Detect Flexiblity

Throughout the simulation, structure of complexes changes constantly and some regions tend to be more flexible. Root mean square fluctuation (RMSF) measures atomic fluctuations compared to average structure within a given a time period. RMSF can be calculated with Equation 2.5 , where T is time and $\mathrm{r_i(t_j)}$ is coordinates of $\mathrm{i^{th}}$ atom in time $\mathrm{t_j}$.

$$RMSF = \left[\frac{1}{T}\sum_{t_j=1}^{T}|r_i(t_j) - r_i^{ref}|^2\right]^{1/2} \tag{2.5}$$

RMSF corresponds to crystallographic temperature or b factor. The higher the temperature factor, the more mobile the atom will be. In GROMACS, RMSF can be measured with gmx rmsf tool. Another advantage of calculating RMSF is, an average structure is produced during the analysis, which can be used for further analysis.

2.5 Principal component analysis to identify repeating patterns

The principal component analysis (PCA), also known as essential dynamics analysis or covariance analysis, is commonly used for interpreting big datasets. PCA enables to find correlations and detect specific patterns by decreasing dimensions. It is important to minimize information loss while reducing dimensionality. PCA analysis finds new variables to reduce information loss. These new variables are specific to dataset and called as principal components (PCs).

With the progresses in technology, we are able to produce longer MD simulations for more complex systems. As time and complexity increase, it becomes harder to analyze such systems. PCA analysis is a popular tool to analyze MD simulations. PCA analysis can reveal major conformational changes by separating amplitude motions. Simply, PCA analysis uses a covariance matrix between carbon alpha (Cα) atoms of $\mathrm{i^{th}}$ and $\mathrm{j^{th}}$ amino acids (Equation 2.6)

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle (i, j = 1, 2, 3, ..., 3N) \tag{2.6}$$

$\mathrm{x_i}$ and $\mathrm{x_j}$ are cartesian coordinates of Cα atoms of $\mathrm{i^{th}}$ and $\mathrm{j^{th}}$ amino acids. N is total number Cα atoms in system where $\mathrm{x_i}$ and $\mathrm{x_j}$ in brackets are coordinates of Cα atoms of $\mathrm{i^{th}}$ and $\mathrm{j^{th}}$ in average structure obtained from MD simulation.

After diagonalizing covariance matrix, eigenvectors and eigenvalues are calculated. Eigenvectors are derived from eigenvalues and also called as PCs. It has been showed that, after ordering eigenvalues, large motions of protein can be obtained from the first PCs (i.e., PC1, PC2 and PC3) (Sittel, Jain and Stock, 2014; Amadei, Linssen and Berendsen, 1993; Groot et al., 2001; Isaak et al., 2018). PCA analysis is performed with ProDy which is an open-source Python package to analyze protein dynamics (Bakan, Meireles and Bahar, 2011; Bakan et al., 2014).

Projection of the PC1 and PC2 on simulation trajectories is performed with GROMACS covar and anaeig tools. GROMACS anaeig calculates the projections with eigenvectors which is created by GROMACS covar.

For PCA analysis, we split each frames in last 300 ns of simulations by 100 ps time step. 3001 frames are investigated at total.

2.6 Interaction Analysis with Interfacea package

Interfacea is an open access Python package written by João Rodrigues and can be installed over GitHub (https://github.com/JoaoRodrigues/interfacea). It calculates intra monomer and inter monomer interactions such as electrostatic, hydrophobic and pi-pi stacking. In this thesis, we used interfacea package over minnie (Molecular INteractioN fIngErprints) developed by Deniz Doğan and Ezgi Karaca. Minnie uses interfacea package to analyze trajectory files, see (https://github.com/CSB-KaracaLab/minnie).

To apply interfacea analysis, trajectory files should be separated into frames. Minnie splitpdbs option creates a directory with the given project name and splits each frame in given ensemble file. Findbonds option uses interfacea package to find interactions within each frame.

As in methodology of PCA analysis, we gathered 3001 frames through the last 300 ns of simulations with 100 ps time step for interaction analysis.

3. RESULTS and DISCUSSIONS

3.1 Performance Optimization on Dimer Complex

In this section, the effect of using different GROMACS versions and computing infrastructure is examined. As can be seen in Table 3.1, the worst performances were produced on the levrekv2-cuda machine. This is expected, since levrekv2-cuda has smallest number of CPUs.

Table 3.1 Performance results on TRUBA

Version	Performance (per day)
1	6.37 ns
2	6.94 ns
3	37.56 ns
4	38.66 ns
5	14.92 ns
6	71.37 ns
7	69.46 ns
8	69.36 ns
9	71.07 ns
10	14.28 ns
11	40.04 ns
12	40.01 ns

PDS energy groups were defined in simulations 1, 3, 5 and 10 (see Methods). For GROMACS 5.1.4, defining different energy groups did not impact performance significantly. Conversely, in the case of GROMACS 2020, when the energy groups were not defined as the system, the simulation performances worsen significantly (as in simulation 5 and 10). This is happening since GROMACS 2020 cannot perform energy calculations on GPU. Upon comparing simulations 5 and 6, we can see that the GPU usage increases the simulation performance by 4.80 times (in akya-cuda cluster). For barbun-cuda, using GPU increases the performance by 2.80 times (Versions 10 and 11). While keeping the CPU/GPU usage the same, the performance difference between akya-cuda and barbun-cuda happens due to the difference in the GPUs and CPUs used.

Interestingly, increasing the number of GPU cards used did not improve the simulation performance. The optimum performance was achieved by using CPU/GPU ratio as 40/1 (Figure 3.1). In conclusion, we obtained the best performance with GROMACS 2020 on akya-cuda cluster with 40/1 CPU/GPU configuration. Therefore, we run our simulations on akya-cuda with these system settings. The related input/output/parameter files are shared over GitHub: https://github.com/CSB-KaracaLab/gmx_performance_on_HPC.

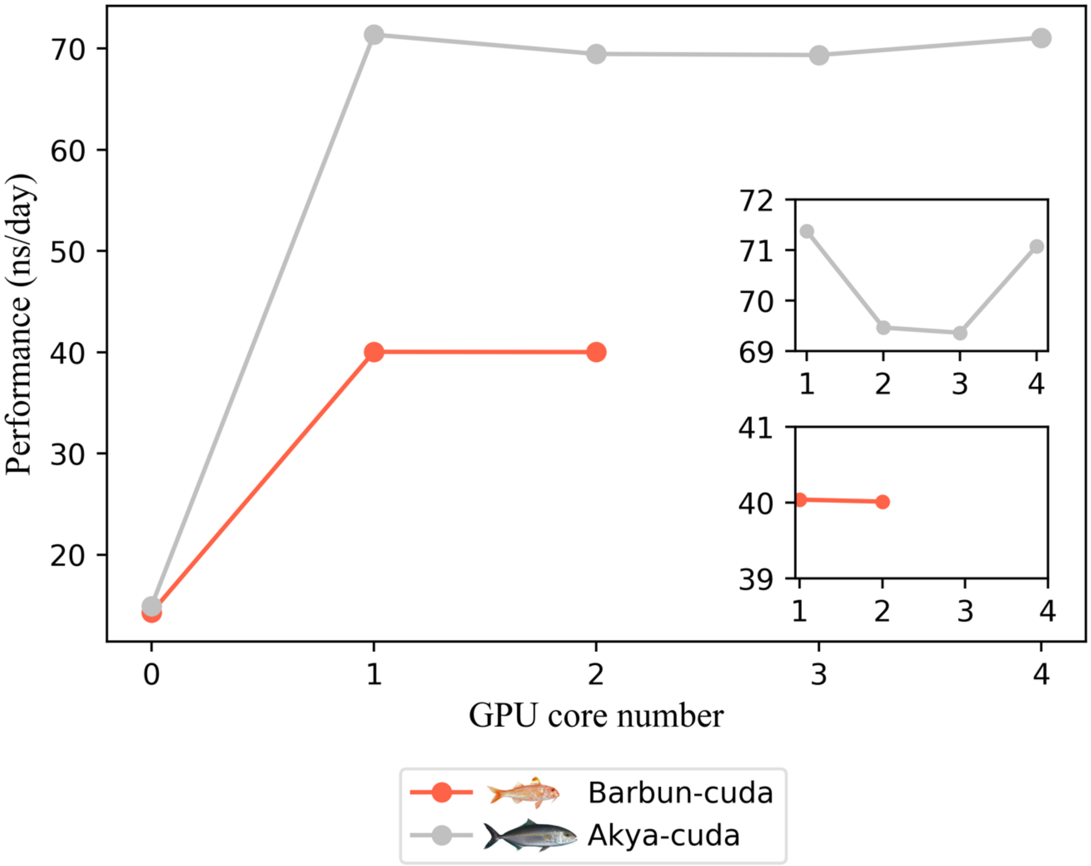


Figure 3.1 Effect of GPU amount on performance in TRUBA clusters while keeping the amount of CPUs constant. Barbun-cuda cluster is represented in orange and gray indicates akya-cuda.

3.2 Methodology Control

We performed 3 parallel simulations for each complex, 6 at total, for 500 ns long. For the quality control of these simulations, we calculated root mean square deviation (RMSD), radius of gyration (R_g), minimum distance to periodic image (mindist) and root mean square fluctuation (RMSF).

3.2.1 RMSD Results

To calculate RMSDs, the frames for a given simulation was compared to the initial structure. As we are working with multi component structures, we calculated RMSDs of Protein-DNA-SAM (PDS), Protein-backbone, and DNA-backbone separately (Figure 3.2).

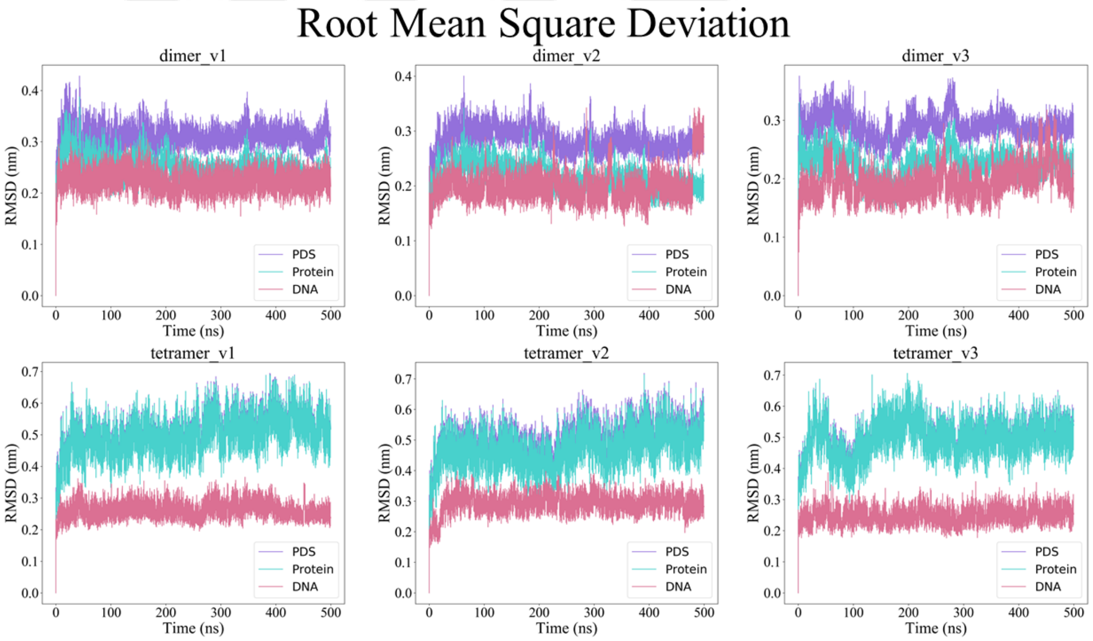


Figure 3.2 Representation of Root Mean Square Deviation Profile of each simulation. Light purple color represents Protein-DNA-SAM (PDS) complexes, turquoise lines represent protein-backbone and violet red lines indicate DNA-backbone structures.

As shown in Figure 3.2, each simulation reaches an equilibrium state after the first 200 ns. For dimer simulation, the equilibrium PDS fluctuates between 0.15 and 0.45 nm. In the case of tetramer, the equilibrium PDS and protein-backbone RMSDs fluctuate between 0.35 nm to 0.70. The following analyses were performed on the last 300 ns of the simulations, after discarding the equilibration period.

3.2.2 R_g Results

R_g describes the shape and compactness of the structures during the simulations. Obtained the R_g values of each simulation is given in Figure 3.3.

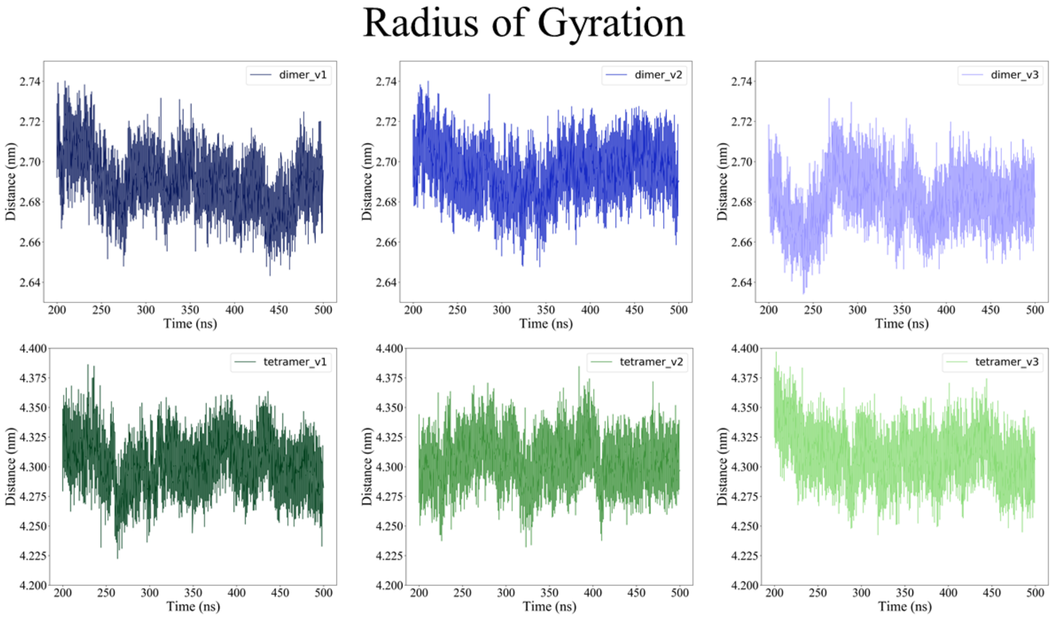


Figure 3.3 Radius of Gyration of the replica simulations. Blue color shades represent dimer simulations, while green shades indicate tetramer simulations.

In tetramer simulations, R_g values are almost the double of the ones observed for the dimer simulations. This is expected as the size of the structure doubles. The compactness of structures in each simulation is stable during the production run.

3.2.3 Mindist Results

The minimum distance to periodic image (mindist) ensures that the molecules in each periodic box do not interact with each other. For the last 300 ns, we analyzed mindist of each simulation (Figure 3.4). As can be seen in Figure 3.4, the minimum distance between periodic images is always larger than 1.20 nm, which is the cut-off for long range electrostatics. Therefore, we can safely claim that our simulations are technically safe and sound.

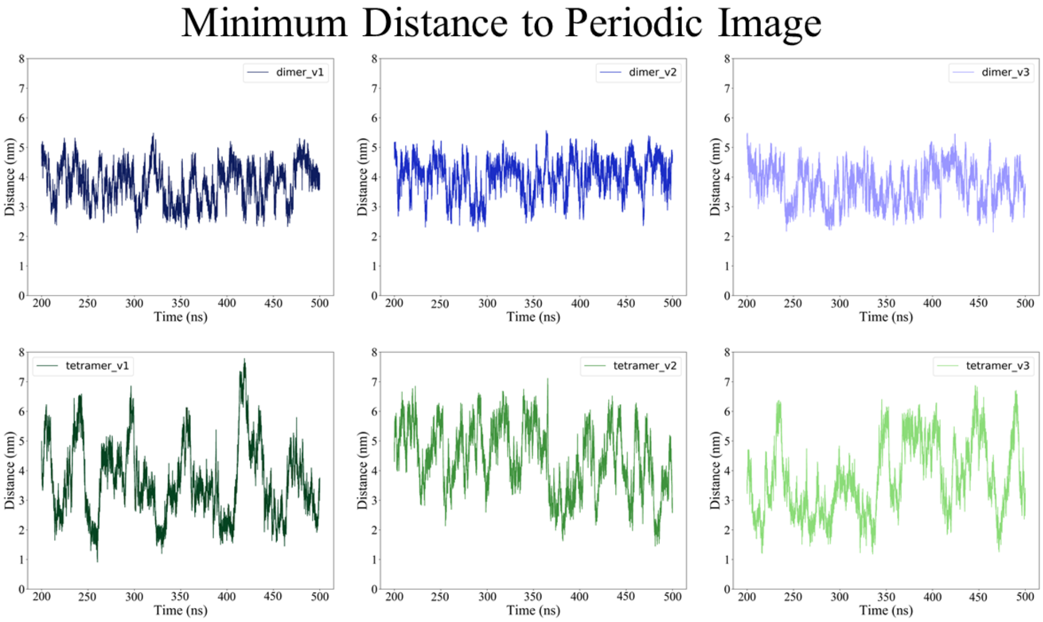


Figure 3.4 Minimum Distance to Periodic Image of All Simulations. Blue lines represent dimer simulations and green lines represent tetramer simulations. The color gets lighter as the replica number increases.

3.2.4 RMSF Results

Cα atoms of protein, DNA and SAM structures are used separately for least square fitting for RMSF calculation. Therefore, these RMSF values contain only residue fluctuations. In Figure 3.5, RMSF of DNA is analyzed by DNA strand with flipped cytosine and complementary strand separately. To that end, we cropped DNA in tetramers into two as they match with DNA in dimer and removed the AATT sequence which is not found in dimers. The highest RMSF values belong to the tails of DNA structures. For the first A nucleotide in complementary DNA and the last T nucleotide in DNA strand with flipped cytosine, tetramers have lower RMSF values than dimers as these nucleotides are not found in at the end of DNA strands (check Figure 2.1). Lowest RMSF values in both dimer and tetramer simulations are seen in GpC*pG sequence which are the neighboring nucleotides of flipped cytosine. During the methylation reaction, flipped cytosine is stably coordinated by several amino acids, thus the lower RMSF is expected. RMSF values belong to complementary strands show various RMSF profiles but DNA strands with flipped cytosine show almost the same RMSF profile for each nucleotide.

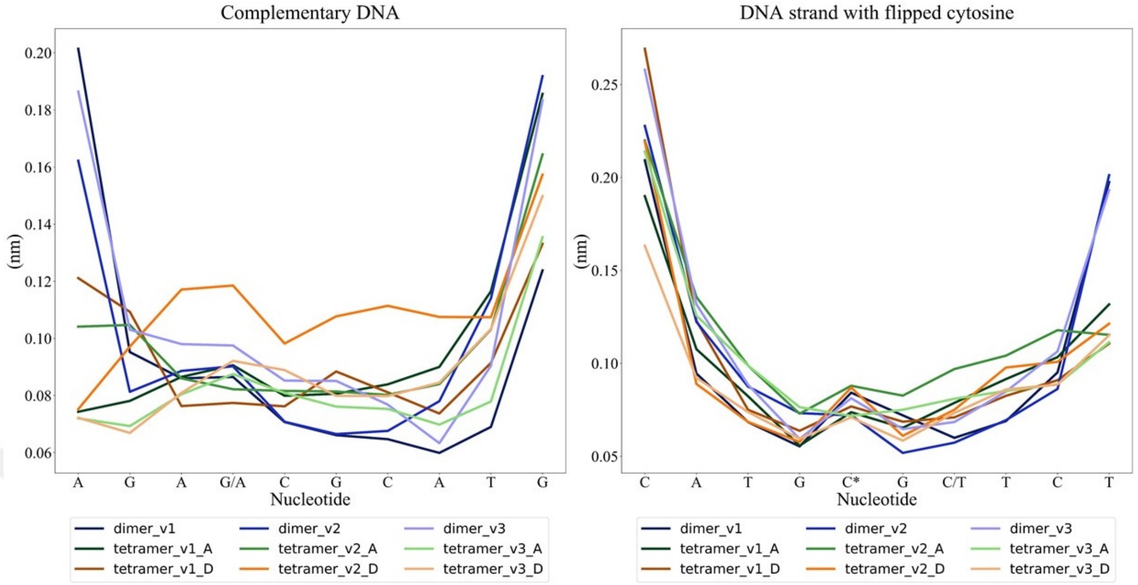


Figure 3.5 Root Mean Square Fluctuation of two strands of DNA. Dimer simulations shown with blue colors, DNA strands interacting with 3A-A chain of tetramer simulation is represented with green colors and DNA strands interacting with 3A-D chain of tetramers is shown with orange colors. Flipped cytosine is emphasized with C*. G/A and C/T indicates nucleotide difference in between dimer and tetramer.

RMSF of DNMT3A enzymes is represented Figure 3.6 and RMSF graph of DNMT3As separated by chains is given in Appendix C. 3A-A chain of dimer simulations seem to have higher RMSF than tetramers for more residues. For instance, RMSF values in between 650th and 680th residues is lower for tetramers than dimers as they belong to residues in hydrophilic interface. For dimers, residues in between 650 and 680 move more freely as they belong to outer parts. Formation of hydrophilic interface in tetramers limit the movement of these residues as they participate in hydrophilic interactions.

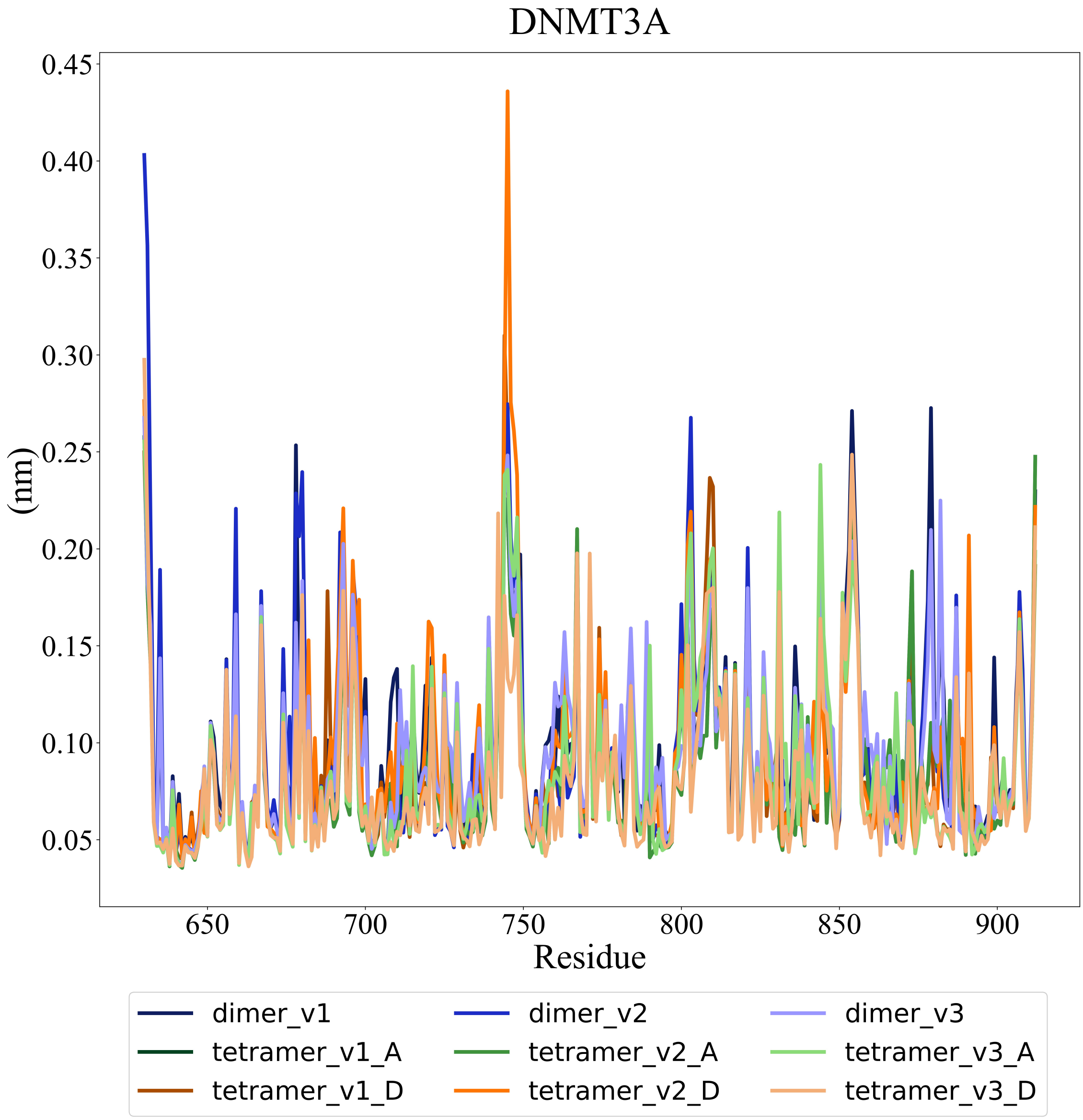


Figure 3.6 Root Mean Square Fluctuation of DNMT3A for each simulation.

RMSF values of DNMT3L structures are given in Figure 3.7. All simulations have similar RMSF profile among the residues except for residues after 300th residue and in between 215th and 220th residues. These residues show higher RMSF values for different simulations which can be explained by their location as they belong to the outer regions of DNMT3L and do not participate in any interaction with DNMT3A directly. Other regions that interact with DNMT3A show more similar RMSF profile.

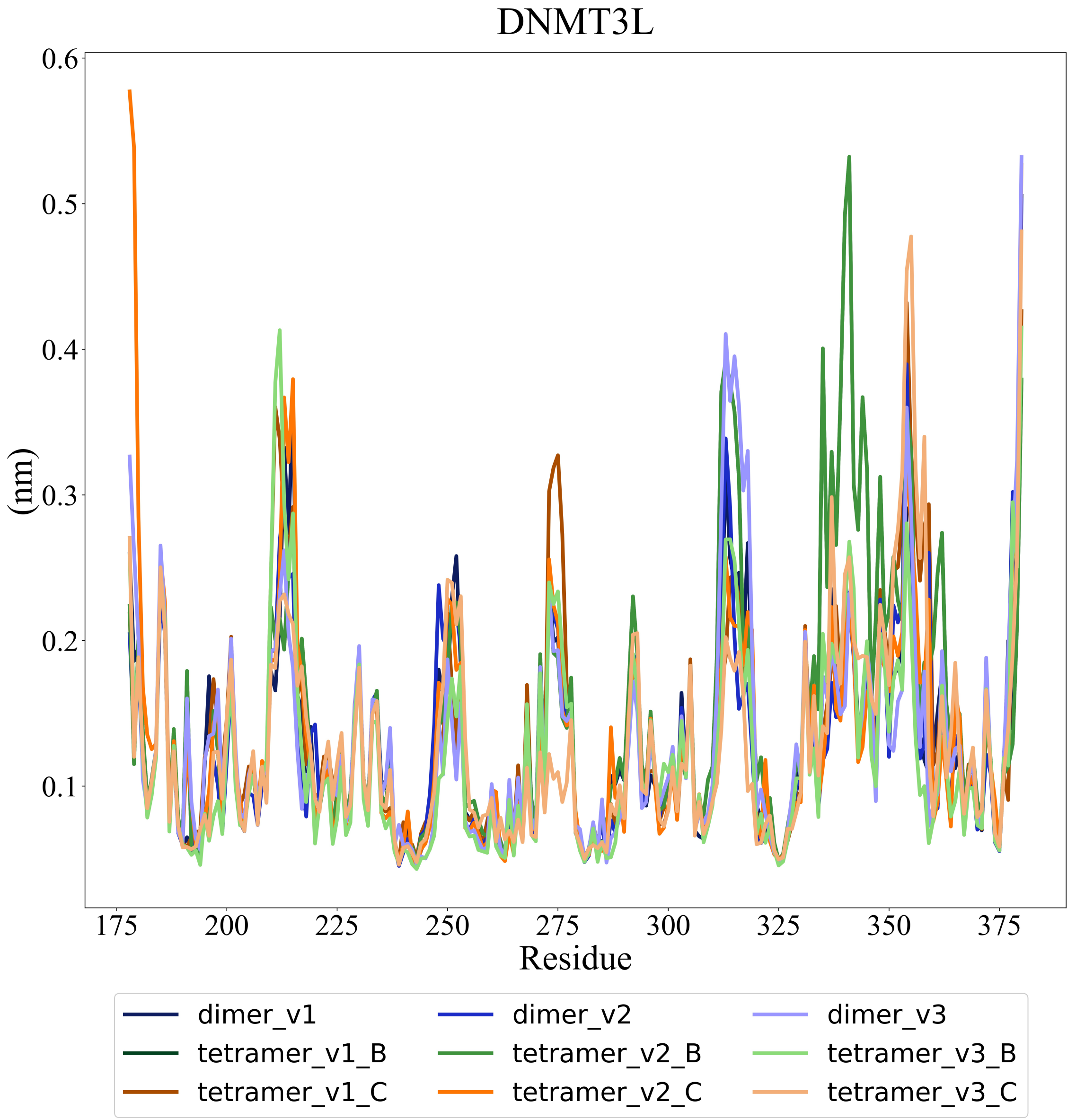


Figure 3.7 Root Mean Square Fluctuation of DNMT3L.

The average of RMSF values are calculated for DNMT3A and DNMT3L separately where 3A:3L-AB and 3A:3L-CD corresponds to subunits of tetramer consists A-B and C-D chains respectively (Table 3.2). For DNMT3A, the highest RMSF average belong to dimers which is expected as the movement will decrease for residues in hydrophilic interface. Although, all DNMT3L proteins show higher fluctuation in average than DNMT3As, there is no explicit difference between dimers and tetramers for DNMT3L.

Table 3.2 Average of RMSF values

Version	DNMT3A	DNMT3L
dimer_v1	0.10	0.13
dimer_v2	0.10	0.13
dimer_v3	0.10	0.13
tetramer_v1 3A:3L-AB	0.08	0.15
tetramer_v2 3A:3L-AB	0.09	0.12
tetramer_v3 3A:3L-AB	0.09	0.12
tetramer_v1 3A:3L-CD	0.09	0.14
tetramer_v2 3A:3L-CD	0.09	0.13
tetramer_v3 3A:3L-CD	0.08	0.13

3.3 PCA analysis to reveal allosteric effect of DNMT3L

For PCA analysis, the monomers are organized within the frames to follow 3L:3A order for dimers and 3L:3A:3A:3L order for tetramers. The first 10 eigenvectors and eigenvalues were deduced from the covariance matrix. Cross correlation maps of all simulation versions were normalized within -1 and 1 range (Appendix D). To analyze dominant correlations of all replica simulations, we plotted the mean of all three replica trajectories together within the range of -1 and 1 (Figure 3.8). The Cross correlation (CCor) maps show regions of each complex, which exert correlated motion. In the dimer CCor map, the first 203 residues belong to DNMT3L and the rest to DNMT3A, structures are seen to be not fully correlated as the purple color is softer. Although each protein shows low correlation within themselves, amount of correlating regions in between DNMT3L and DNMT3A is quite large. Moreover, correlation in between domains of DNMT3A in dimer is lesser. In tetramer graph, after the first 486 residues D chain of DNMT3A and C chain of DNMT3L proteins are seen. Formation of tetramer structure leads to increase in amount of correlated regions within proteins. Also, some regions in protein are seem to reach full correlation. Surprisingly, the two most distant proteins, two chains of DNMT3L, appear to have correlation with each other. Moreover, each DNMT3L chain correlates with not only neighboring 3A but some regions of other 3A.

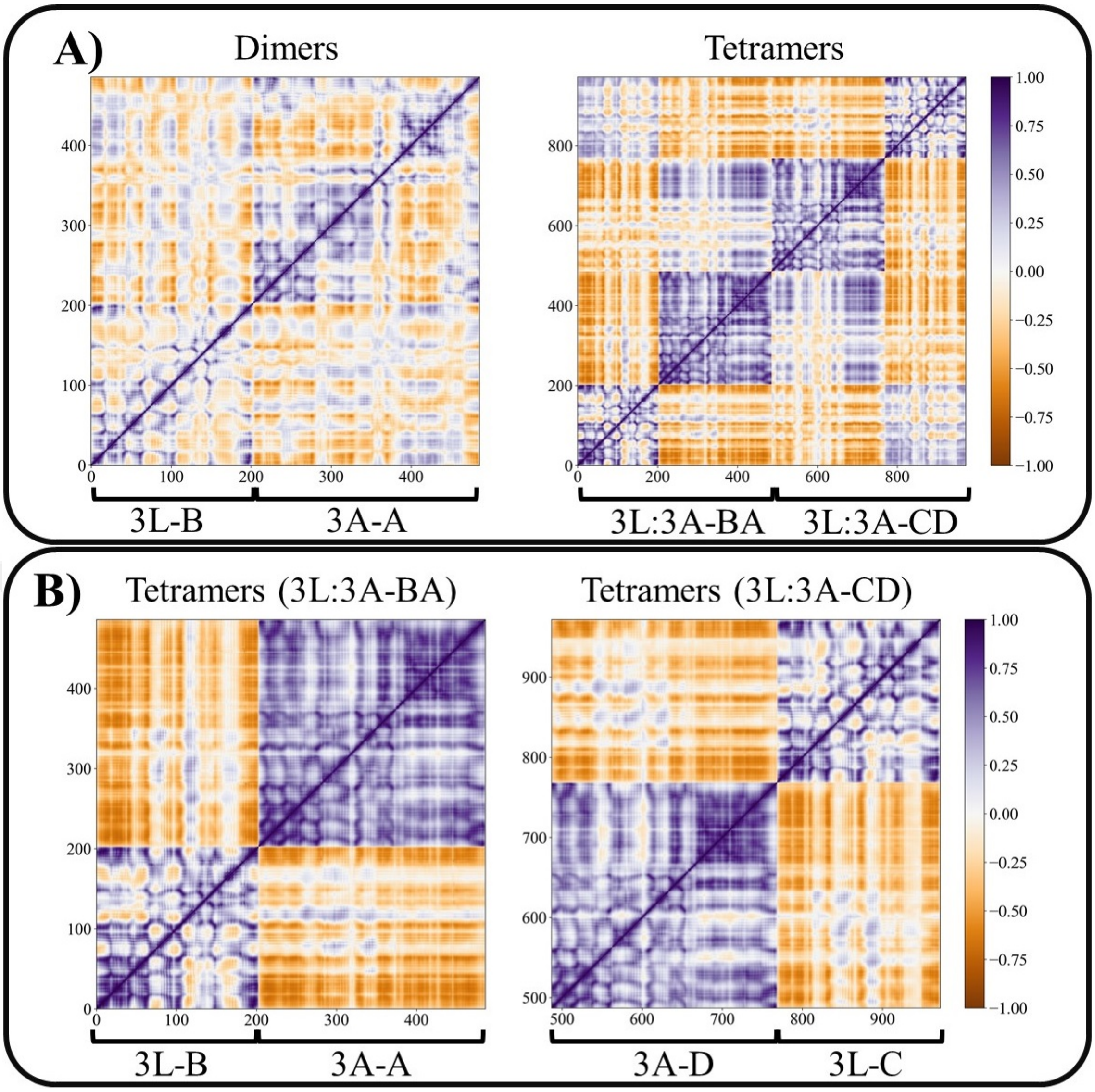


Figure 3.8 Cross correlation (CCor) maps. A) The CCor map of the dimer and tetramer. B) The CCor map of 3A:3L-AB and 3A:3L-CD chains of tetramers. Color legend: +1 indicates positive correlation in purple color and -1 represents anti-correlation in orange.

To compare the same regions in both dimer and tetramer, we isolated 3A:3L-AB and 3A:3L-CD chains from tetramer (Figure 3.8B). Both 3A:3L-AB and 3A:3L-CD chains show higher correlations within proteins than dimers. Additionally, correlated and anti-correlated parts are easy to distinguish in tetramers. Although we expected to see same correlated regions in between 3A:3L-AB and 3A:3L-CD chains as they are identical, we observed slightly different parts. Therefore, we emphasized the correlated regions specific to each complex with different colored boxes in Figure 3.9 in detail.

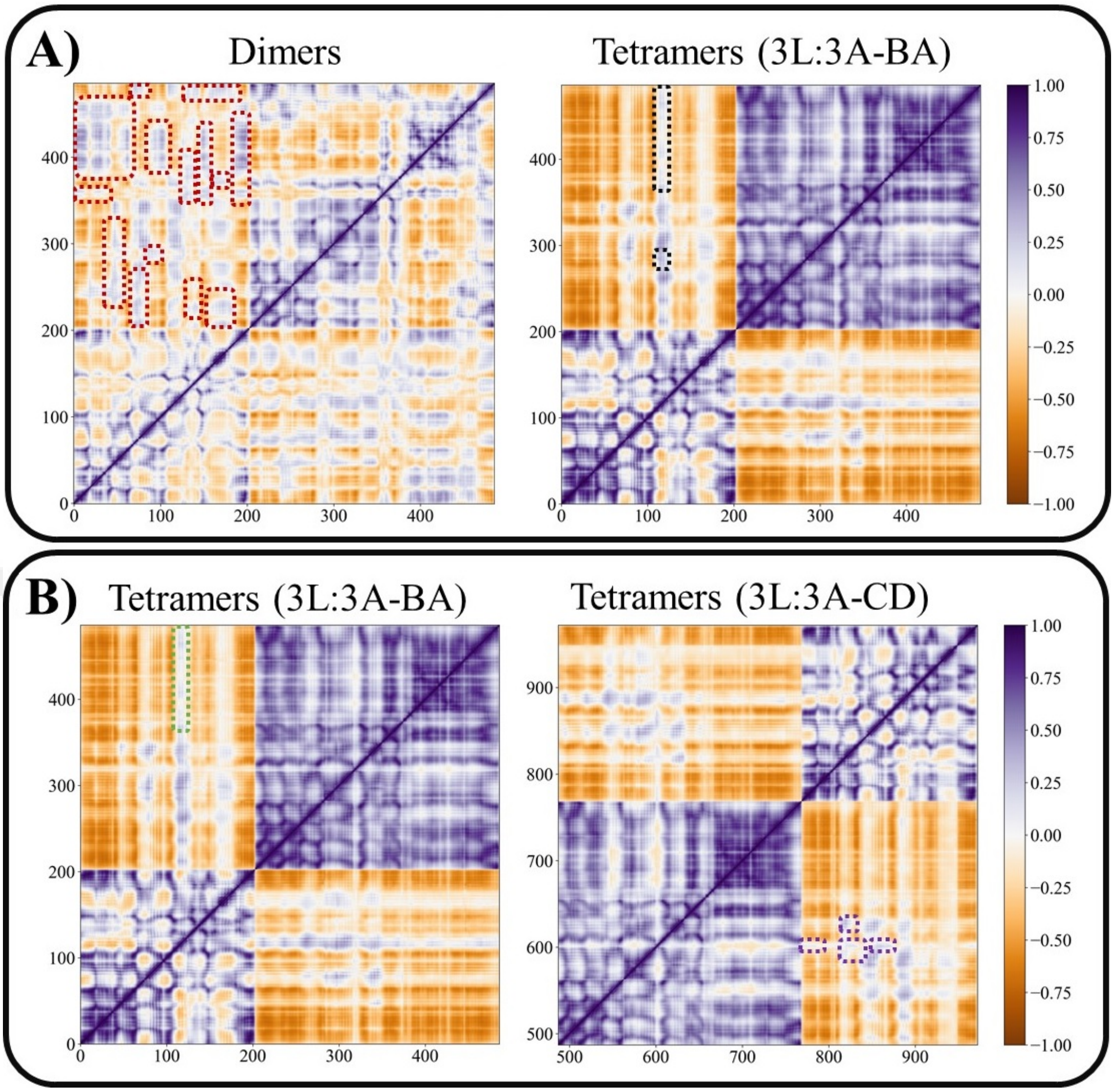


Figure 3.9 Complex specific correlated parts. A) Difference between dimers and 3A:3L-AB chains of tetramers B) Difference between 3A:3L-AB and 3A:3L-CD chains of tetramers

For further, we listed correlated regions in detail (Table 3.3) with PCA numbering (see Appendix A and B). Locations of essential motifs on CCor maps can be found in Figure D.2. In common correlated regions, 5 amino acids correspond to two essential motifs in DNMT3A. 283rd, 284th and 285th amino acids (in PCA numbering) are named as Motif IV which binds target cytosine. 260th and 261th amino acids are belonged to Motif III that binds SAM. Correlated regions specific to dimer include three more essential motifs additional to Motif IV. 304th residue binds SAM and called as Motif V. Motif VI corresponds to 330th, 331st, 332nd residues. Also, 219th residue, Motif I, is associated with substrate binding activity.

Table 3.3 Correlating regions in each complex between proteins

No	DNMT3A	DNMT3L	Specifity
1	337-350	45-51	In both complexes
2	256-270, 280-316, 331-352	66-89	In both complexes
3	205-215, 219-238, 255-281, 294-351	108-125	In both complexes
4	376-466	1-16, 19-43	In dimer only
5	411-441	56-66	In dimer only
6	451-466	53-66	In dimer only
7	388-407	59-66	In dimer only
8	472-486	66-81, 107-120, 153-182	In dimer only
9	384-438	89-105	In dimer only
10	205-281	65-81	In dimer only
11	352-367, 372-392, 399-405	127-144	In dimer only
12	205-332	43-53	In dimer only
13	219-235	130-145	In dimer only
14	352-365, 378-465	178-202	In dimer only
15	205-213, 219-238	156-185	In dimer only
16	247-258	167-184	In dimer only
17	368-381	159-179	In dimer only
18	351-367, 377-443	145-158	In dimer only
19	281-293	108-125	In tetramer only
20	352-391, 403-423, 433-486	112-123	3A:3L-AB only
21	770-787	598-605	3A:3L-CD only
22	811-825	617-635	3A:3L-CD only
23	819-832	599-607	3A:3L-CD only
24	852-874	597-606	3A:3L-CD only

In tetramer, there is another correlated region with Motif IV specific to complex. The other complex specific correlated region in tetramer includes residues in hydrophilic interface.

The findings in Table 3.3 are visualized in Figure 3.10. For tetramer representation, we selected 3A:3L-AB and 3A:3L-CD chains with cropped 10 nucleotide long DNA chains from tetramer complex to be comparable with dimer structure.

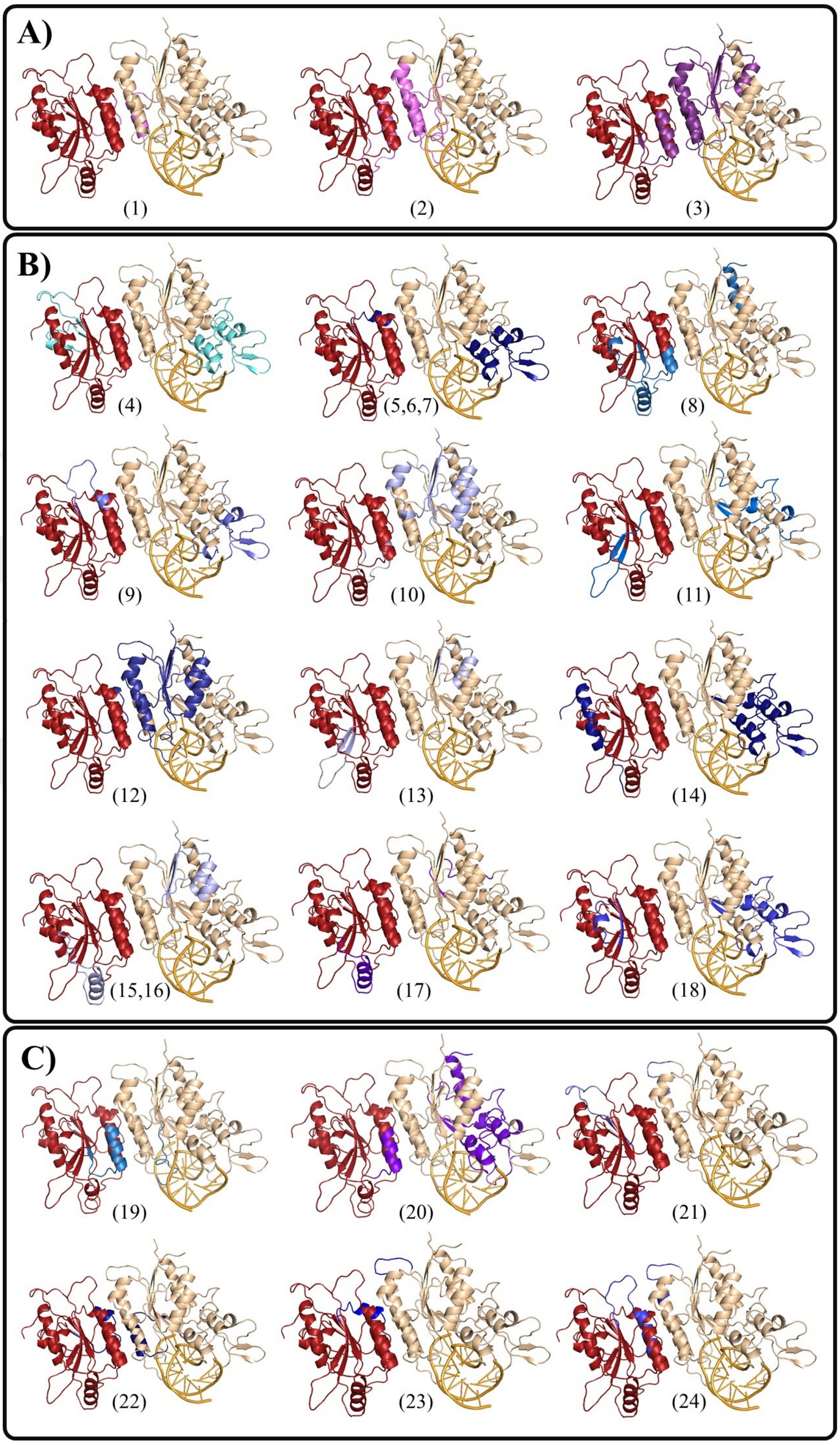


Figure 3.10 Representation of correlating regions. A) Commons, B) Dimer specifics C) Tetramer specifics

For further, we analyzed the PCA results in aspect of change in motion. 1458 eigenvectors and eigenvalues are calculated for dimers as they consist of 486 Cα atoms. Tetramer structures have 972 Cα atoms, so 2916 eigenvalues and eigenvectors are calculated for each version. All eigenvectors are listed according to their eigenvalues in a way that the amount of motion covered by each eigenvalue decrease by GROMACS covar tool. The first 10 eigenvalues are capable of capturing relevant motions in proteins (Groot et al., 2001; Isaak et al.,2018). For dimer versions, first ten principal components cover 56.82%, 51.63% and 56.54% of the total movement in proteins, respectively (Figure 3.11). In tetramers, first ten principal components capture 68.76%, 68.42% and 66.62% of the total movement.

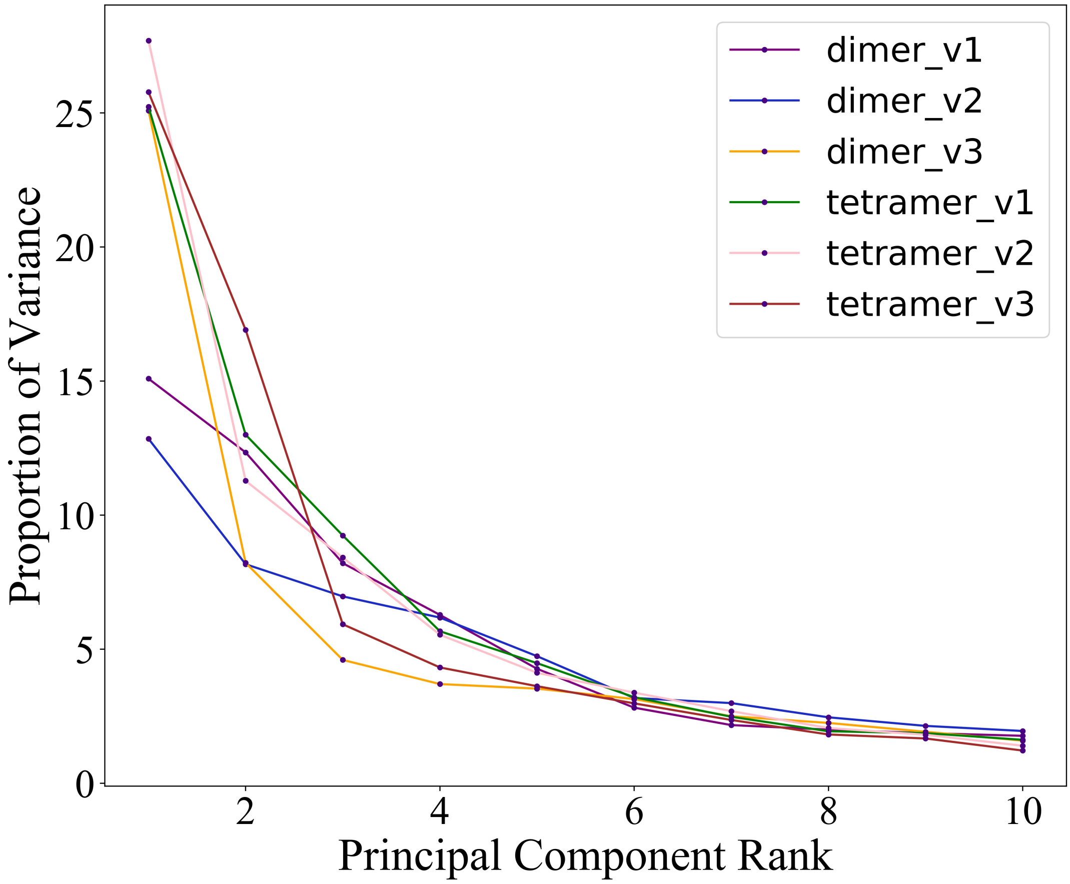


Figure 3.11 Proportion of variance of first ten PCs of each simulation.

In all simulations, first mode dominates the rest of the modes and ratios of the first modes for all simulations are 15.09%, 12.85%, 25.09%, 25.23%, 27.70% and 25.78%, respectively. Proportion of variance of first modes in tetramers are higher than dimers except for dimer_v3.

For further investigation, the first two modes of each simulation is analyzed to observe the change in motion (Figures 3.12 and 3.13).

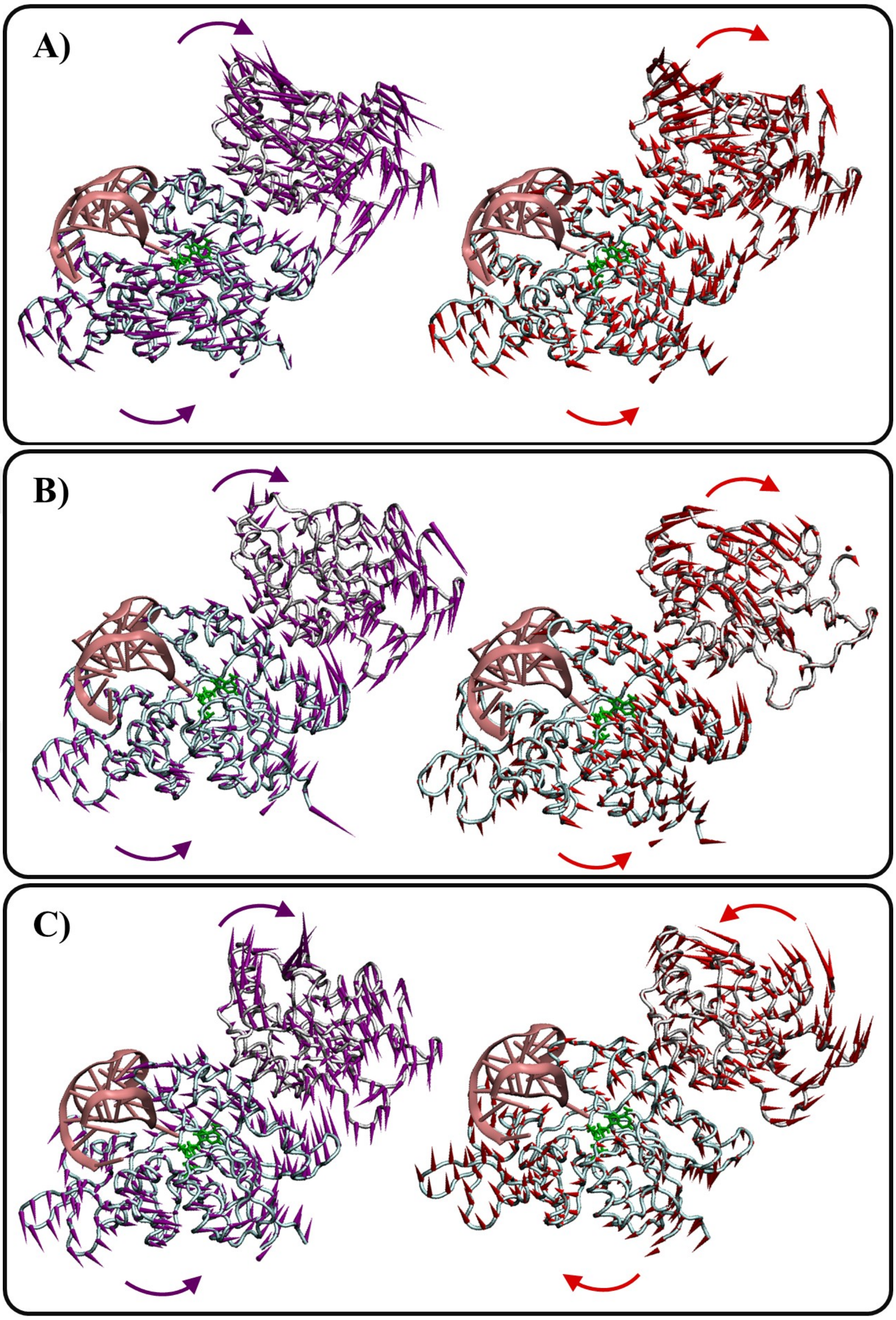


Figure 3.12 First two modes of dimers, the left side of the figure represents first mode and right side shows second mode. DNMT3L structure visualized with gray and DNMT3A colored with cyan where DNA represented with pink. A) dimer_v1. B) dimer_v2. C) dimer_v3

All dimer versions show similar motions for the first two modes except the second mode of dimer_v3 as DNMT3L rotates clockwise and DNMT3A rotates in opposite direction. For tetramers, the first mode of tetramer_v1 shows opposite rotations than others.

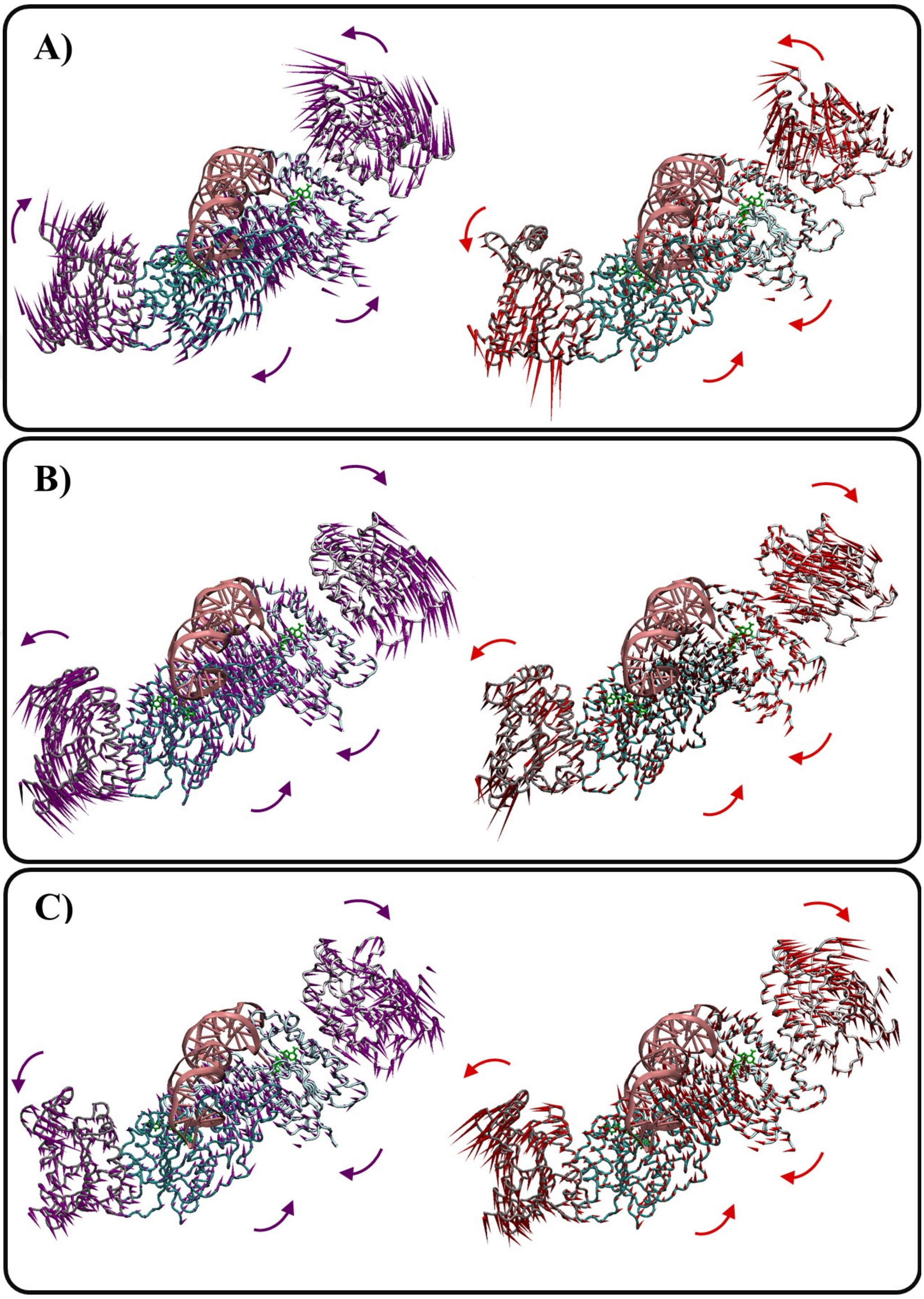


Figure 3.13 First two modes of tetramers, coloring choices for dimers are applied. A) tetramer_v1. B) tetramer_v2. C) tetramer_v3

For comparison, in all dimer simulations DNMT3As and DNMT3Ls rotate opposite direction as one of them moves clockwise and the other one moves counterclockwise, but in tetramers neighboring DNMT3As and DNMT3Ls rotates in the same direction.

Furthermore, we projected the first and second modes of each simulation on other simulations for the same complexes (Figure 3.14). Each point represents a different conformation along the simulation and overlapping points represent similar conformations. All dimer simulations sample similar conformations during the simulation (Figure 3.14A). For tetramers, tetramer_v3 simulation show different sampling than other tetramers.

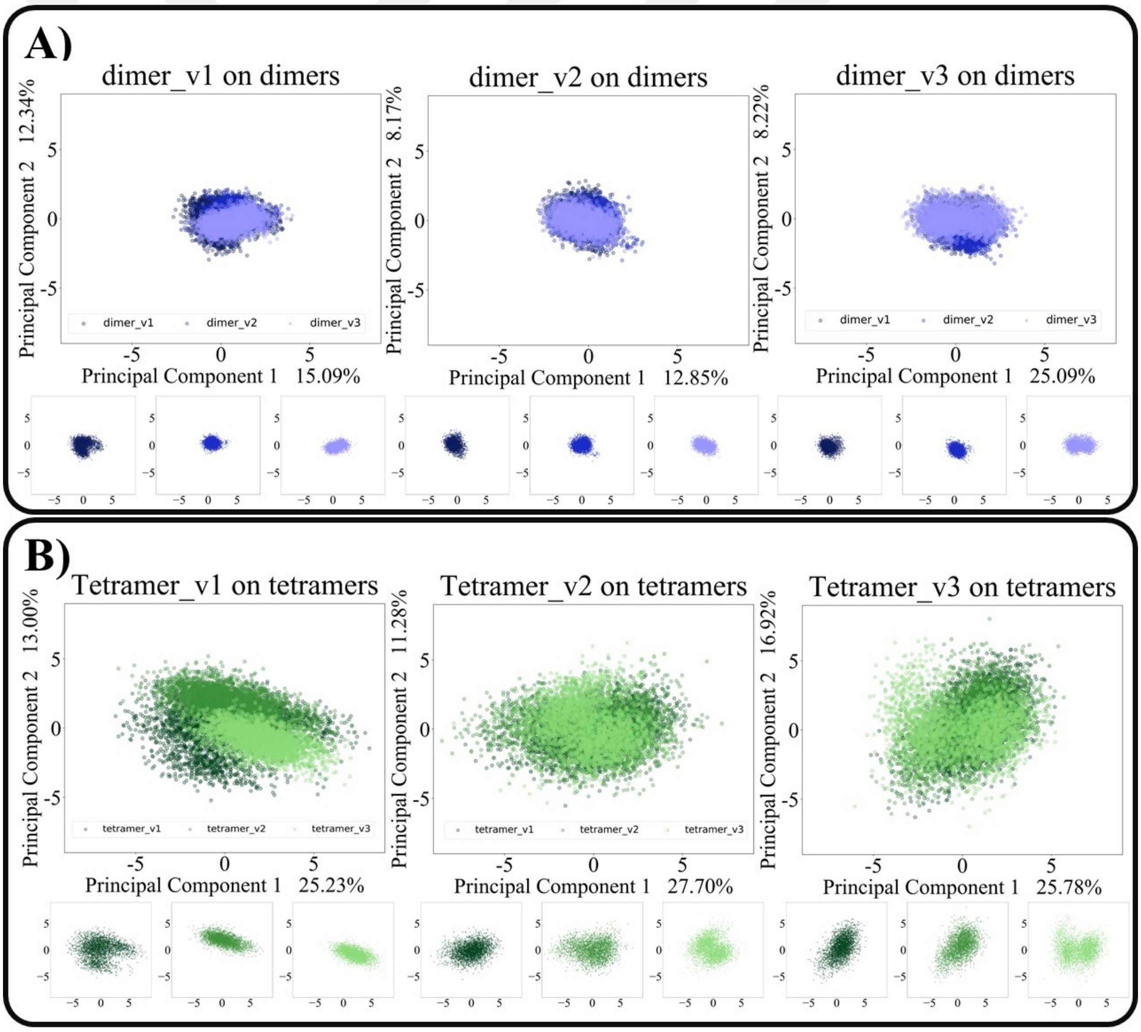


Figure 3.14 Projection of first and second modes of each simulation on other simulations. A) Dimers, each version is indicated with different tone of blue. B) Tetramers, each version is represented with different tone of green.

3.4 Interfacea to examine Protein-DNA interactions

Protein-DNA interactions are categorized as specific and non-specific interactions. If the interaction occurs between protein and sugar phosphate backbone of the DNA, then it is referred to nucleotide-non-specific. An interaction is classified as nucleotide-specific, if it is formed between any part of an amino acid and bases of DNA.

Interfacea was run on the trajectories to illuminate the interaction profile differences. Here, we only concentrated on protein-DNA interactions. For each frame, we deduced individual and the total number of protein-DNA interactions observed throughout the simulation. We then pooled all the observations to obtain a statistically meaningful set. The average interaction number per frame is given in Table 3.4.

Table 3.4 Average of interaction in each frame

Version	Hydrogen Bond	Hydrophobic Interaction	Salt Bridge
dimer	16.98	22.62	4.39
tetramer	35.58	45.71	10.44
tetramer 3A-A:DNA	16.87	20.06	5.25
tetramer 3A-D:DNA	18.71	25.65	5.19

The DNMT3A in dimer structure has more average interaction amount than 3A-A chain of tetramer except for salt bridge. The distribution of interactions is plotted in Figure 3.15. The average and distribution of interactions remains almost the same for hydrogen bond and hydrophobic interaction.

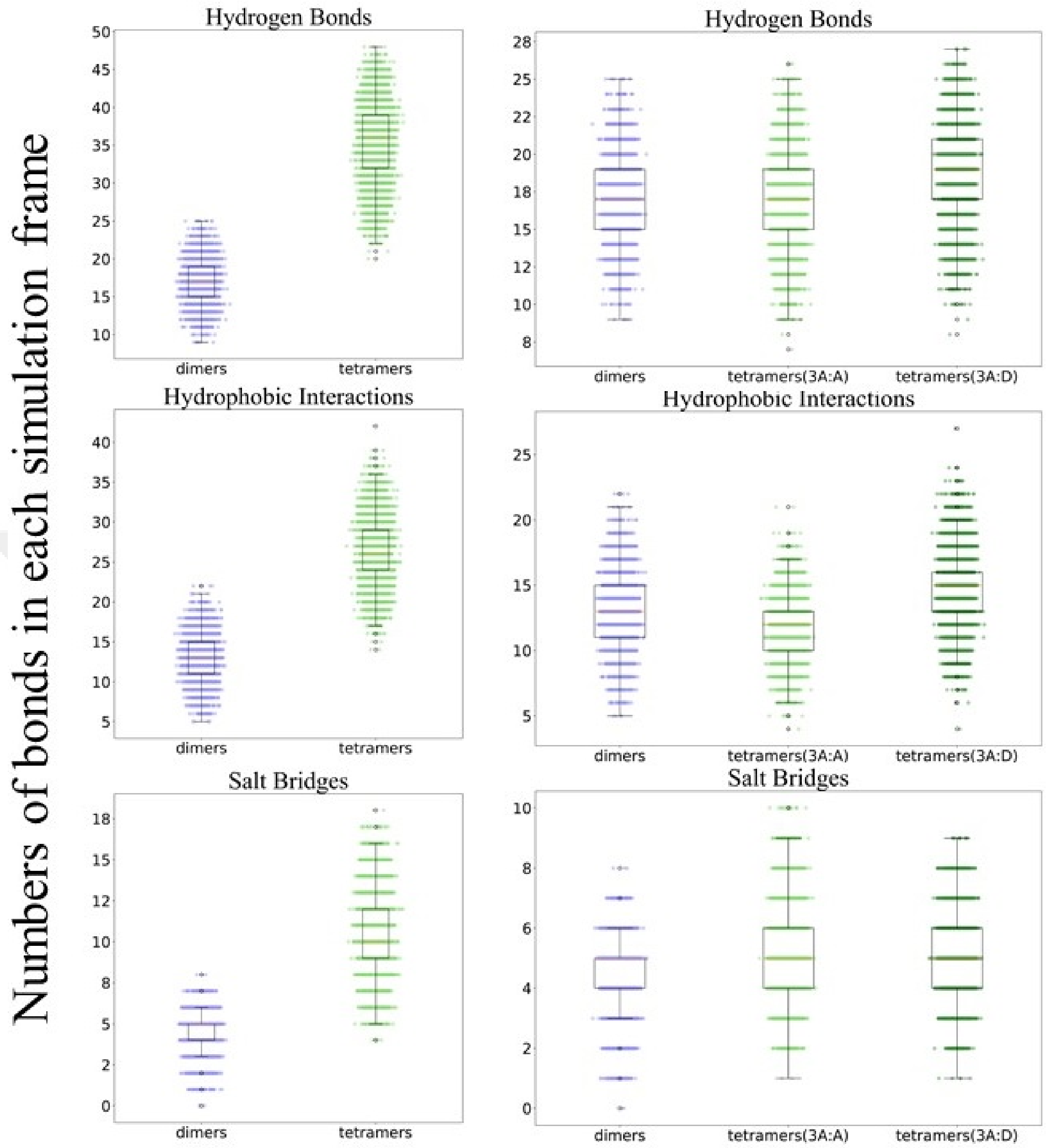


Figure 3.15 Distribution of protein DNA interactions among complexes. Each complex is represented with different color and chain discrimination by tone difference.

Since the most prominent difference between dimer and tetramer structures observed with salt bridge interaction, we showed salt bridges between DNA and protein (Figure 3.16). For each complex, we selected the frame with highest salt bridge number among all versions to represent.

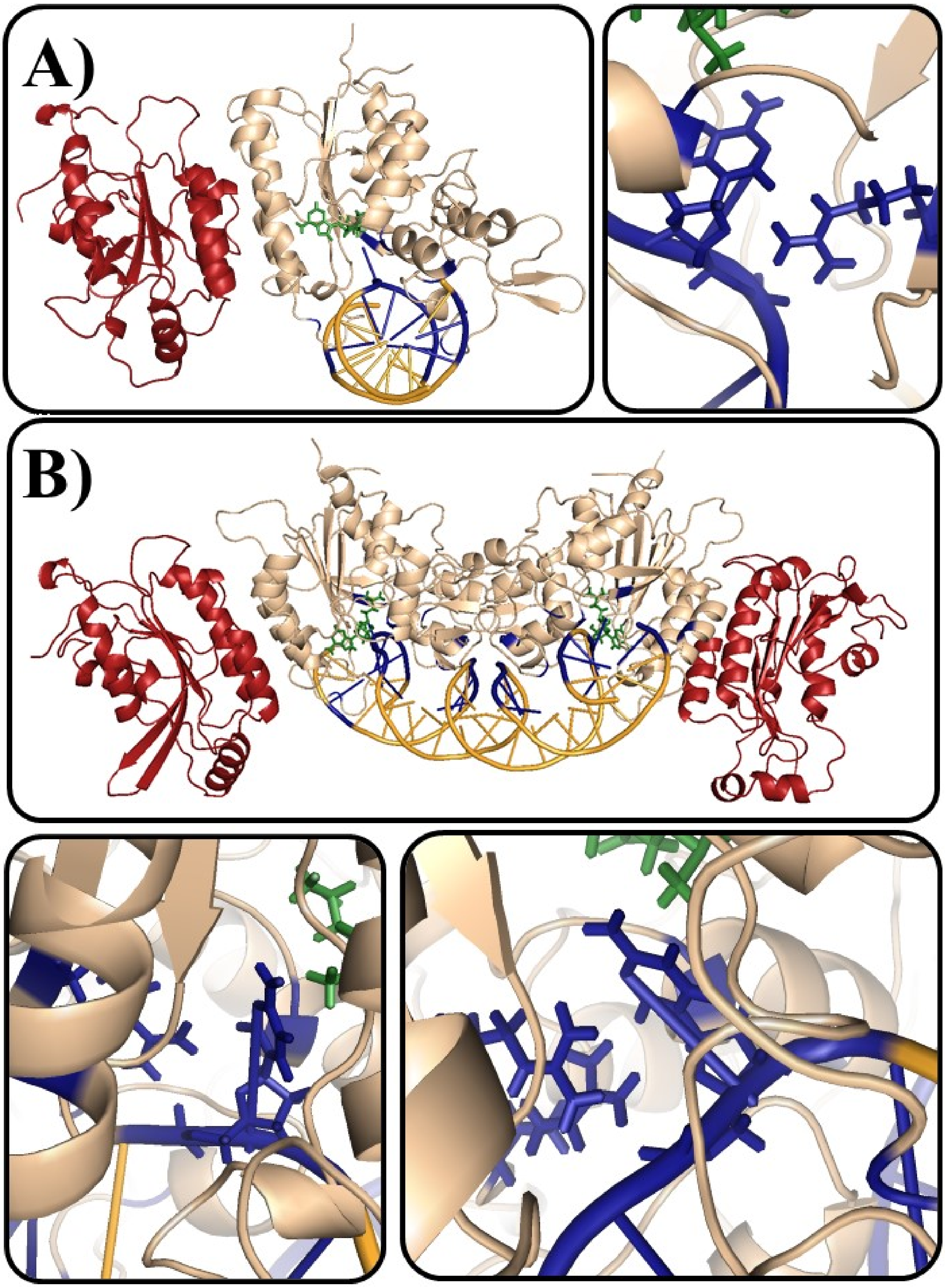


Figure 3.16 Representation of locations of salt bridges between DNA and protein (DNMT3A). A) Salt bridges in dimer with one interaction with flipped cytosine, B) Salt bridges in tetramer with two interactions with flipped cytosine.

In detail, dimer_v3 is selected which has 8 salt bridges at 2722nd frame. Additionally, 2469th frame is selected in tetramer_v1 with 18 salt bridges. In dimer structure, 5 of the salt bridges are in between target cytosine and neighboring nucleotides and only one of them is with flipped cytosine. In tetramer, 10 of salt bridges belong to A chain with 6 interaction with neighboring nucleotides and 2 of them targets cytosine. Other 8 salt bridges are occurred by D chain and only 6 of them interact with neighboring and one of them interacts with target cytosine. Most of the other salt bridges are located in hydrophilic interface in between two active sites of DNA.

3.4.1 Protein-DNA Interactions among CpGpC or CpGpT sequence

Since it is hard to obtain small changes in a complex system, we narrowed our focus to the flipped cytosine and its neighboring nucleotides (CpGpC for dimers and CpGpT for tetramers). The interaction numbers are given in Table 3.5 for only nucleotide non-specific interactions for salt bridges as any of the complexes do not show nucleotide specific salt bridge interaction.

Table 3.5 Percentage of salt bridge interactions in each frame for flipped cytosine (shown with *) and its neighboring nucleotides

Complex	Interacts with	Type	# of interaction	Percentage
dimer	cytosine*	non-specific	4418	44.20 %
dimer	guanine	non-specific	5534	55.37 %
dimer	cytosine	non-specific	43	0.43 %
tetramer 3A:3L-AB	cytosine*	non-specific	8928	77.41 %
tetramer 3A:3L-AB	guanine	non-specific	2603	22.57 %
tetramer 3A:3L-AB	thymine	non-specific	2	2 %
tetramer 3A:3L-CD	cytosine*	non-specific	4010	37.79 %
tetramer 3A:3L-CD	guanine	non-specific	6186	58.29 %
tetramer 3A:3L-CD	thymine	non-specific	416	3.92 %

Interactions with third nucleotide can be ignored as they seen in limited number of frames. 3A:3L-CD chains of tetramers show similar ratios to dimers for both flipped cytosine and guanine nucleotide as opposed to 3A:3L-AB chains. Interestingly, ratio of the dimer and 3A:3L-CD chains are lesser for flipped cytosine but higher for guanine nucleotide but the total number of interactions for these two nucleotides remain close in each complex which led us to think that lost interactions with flipped cytosine could be compensated with guanine nucleotide for dimer and 3A:3L-CD chains. To identify this difference between 3A:3L-AB and 3A:3L-CD chains on structure, we selected tetramer_v3 simulation as it has more salt bridge interactions than other versions. In 2978th frame, 3A:3L-AB chains include two salt bridge interaction with flipped cytosine where 3A:3L-CD chains have only one interaction. Among dimer complexes, we chose 2966th frame with one interaction. Tetramer structure is divided into two pieces to be comparable with dimer. In Figure 3.17, all the chosen frames and their interactions with flipped cytosine is shown in detail.

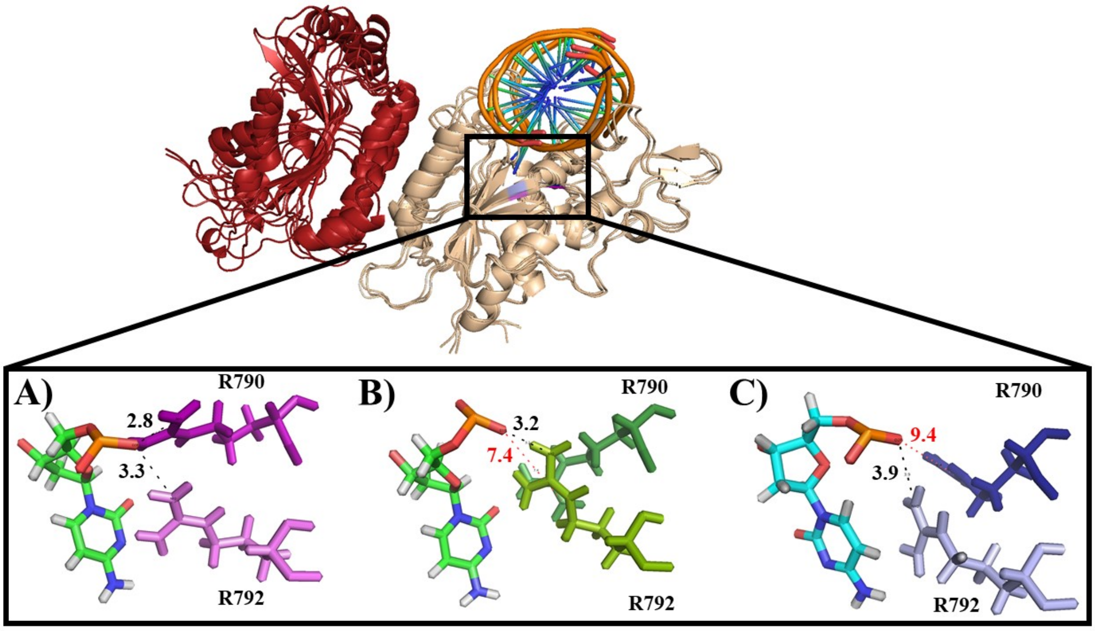


Figure 3.17 Saltbridge interactions in chosen frames for dimer and tetramer complexes. A) 3A:3L-AB chains of tetramer, B) 3A:3L-CD chains of tetramer and C) Dimer complex. Formed interactions are emphasized with black and red indicates opposite.

R790 interacts with flipped cytosine in only 3A:3L-AB chains due to conformational change of the amino acid and this interaction is the reason of the increase in the ratio of 3A-A chains since any of the other complexes do not involve any interaction in between R790 amino acid and flipped cytosine in any frame. All the interactions occurred with sugar phosphate backbone of flipped cytosine, specifically OP1 atom.

Distribution of hydrogen bond interactions among flipped cytosine and its neighboring nucleotides is given in Table 3.6 for both specific and non-specific interactions. Guanine nucleotide shows higher ratio for nucleotide non-specific nucleotide interactions than flipped cytosine in each complex. Third nucleotide differs in dimers as cytosine and appears to be interacted more than thymine nucleotide during the simulations. Since thymine nucleotide makes lesser hydrogen with its corresponding nucleotide than cytosine, decrease in the interaction number could be an effect of increase in flexibility of third nucleotide. As total number of frames for each complex is 9003 after addition of each version, it should be also noted that nucleotide non-specific interactions with the third nucleotide does not exist all the time during the simulation.

Table 3.6 Percentage of hydrogen bond interactions in each frame for flipped cytosine (shown with *) and its neighboring nucleotides

Complex	**Interacts with**	**Type**	**# of interaction**	**Percentage**
dimer	cytosine*	non-specific	9160	24.53 %
dimer	guanine	non-specific	23163	62.02 %
dimer	cytosine	non-specific	5026	13.46 %
tetramer 3A:3L-AB	cytosine*	non-specific	9823	32.55 %
tetramer 3A:3L-AB	guanine	non-specific	17503	58.00 %
tetramer 3A:3L-AB	thymine	non-specific	2851	9.45 %
tetramer 3A:3L-CD	cytosine*	non-specific	14186	35.77 %
tetramer 3A:3L-CD	guanine	non-specific	22066	55.63 %
tetramer 3A:3L-CD	thymine	non-specific	3411	8.60 %
dimer	cytosine*	specific	21784	97.71 %
dimer	guanine	specific	488	2.19 %
dimer	cytosine	specific	22	0.10 %
tetramer 3A:3L-AB	cytosine*	specific	30877	89.17 %
tetramer 3A:3L-AB	guanine	specific	3745	10.82 %
tetramer 3A:3L-AB	thymine	specific	5	0.01 %
tetramer 3A:3L-CD	cytosine*	specific	28397	97.44 %
tetramer 3A:3L-CD	guanine	specific	745	2.56 %
tetramer 3A:3L-CD	thymine	specific	1	0.00 %

In all complexes, more than one nucleotide non-specific interaction occurs for flipped cytosine and guanine nucleotide as the interaction number is higher than 9003.

Nucleotide specific interactions occur with flipped cytosine in higher ratio than other nucleotides. Each frame has at least 2 nucleotide specific interaction with flipped cytosine for all complexes. The high amount of specific hydrogen bond interaction is expected as flipped cytosine should be tightly regulated during the methylation reaction. Interactions with the second nucleotide, guanine, is increase 5 times for 3A:3L-AB chains and others can be ignored due to appearance rate during the simulation.

Hydrophobic interactions occurring with the three nucleotides is given in Table 3.7 for both specific and nucleotide non-specific interactions. By examination of non-specific hydrophobic interactions, 3A:3L-AB and 3A:3L-CD chains show similar ratios for all three nucleotides decreasing from flipped cytosine to thymine.

Table 3.7 Percentage of hydrophobic interactions in each frame for flipped cytosine (shown with *) and its neighboring nucleotides

Complex	Interacts with	Type	# of interaction	Percentage
dimer	cytosine*	non-specific	15596	37.15 %
dimer	guanine	non-specific	24347	58.00 %
dimer	cytosine	non-specific	2036	4.85 %
tetramer 3A:3L-AB	cytosine*	non-specific	26234	56.95 %
tetramer 3A:3L-AB	guanine	non-specific	19358	42.02 %
tetramer 3A:3L-AB	thymine	non-specific	472	1.02 %
tetramer 3A:3L-CD	cytosine*	non-specific	24221	51.55 %
tetramer 3A:3L-CD	guanine	non-specific	22130	47.10 %
tetramer 3A:3L-CD	thymine	non-specific	635	1.35 %
dimer	cytosine	specific	0	0.00%
tetramer 3A:3L-AB	thymine	specific	4283	100 %
tetramer 3A:3L-CD	thymine	specific	8957	100 %

For dimers, non-specific hydrophobic interactions are higher with guanine nucleotide than flipped cytosine. There is no nucleotide specific hydrophobic interaction for dimers but for tetramers third nucleotide, thymine, involves in nucleotide specific hydrophobic interaction. In detail, amino acids interact with C5 and C7 atoms in thymine. As third nucleotide in dimers, cytosine, does not consist of these atoms, nucleotide specific hydrophobic interactions do not occur. To distinguish nucleotide specific hydrophobic interactions depending on nucleotide difference, we chose two frames that only one of them shows specific hydrophobic interactions (Figure 3.18). 997th frame of tetramer_v1 consists 3 specific hydrophobic interactions where two of them belong to 3A:3L-AB chains. 1002nd does not show any nucleotide specific interactions.

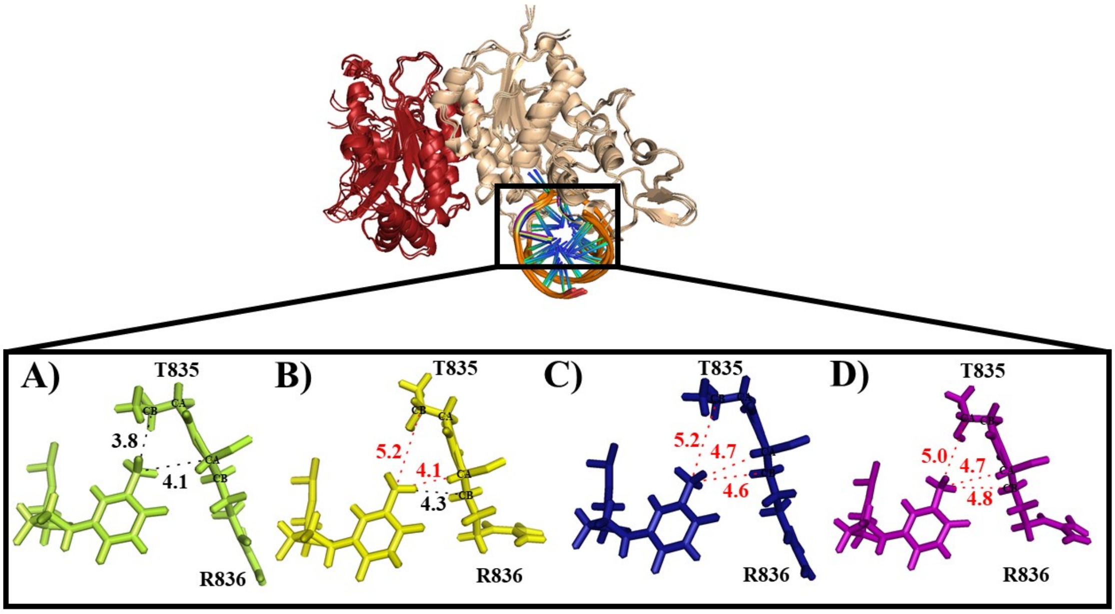


Figure 3.18 Hydrophobic interactions in chosen frames for tetramer complexes. A) 3A:3L-AB chains of tetramer_v1 at 997th frame, B) 3A:3L-CD chains of tetramer_v1 at 997th frame, C) 3A:3L-AB chains of tetramer_v1 at 1002nd frame and D) 3A:3L-CD chains of tetramer_v1 at 1002nd frame. Formed interactions are shown with black and red indicates opposite.

In Figure 3.18A, R836 and T835 residues interact with flipped cytosine in 3A:3L-AB chains. In 3A:3L-CD chains, interaction with T835 residue is lost due to increase in distance.

4. CONCLUSIONS

In this thesis, we investigated the role of DNMT3L on DNMT3A at molecular level with molecular dynamics simulation. As DNMT3L bound to DNMT3A found in two oligomeric states, we analyzed both complexes in terms of formation of hydrophobic and hydrophilic interfaces. With RMSF analysis, we found that DNMT3L is fluctuates more in tetramer structure, but this finding is not enough of reach a conclusion as it could be resulted by being outer side of the complex.

We applied PCA analysis to analyze allosteric effect DNMT3L which does not directly interact with DNA. Correlation of proteins within themselves increases in tetramer structure. Although correlated regions decrease in number in tetramer structure, significance of correlation increase positively. The most interesting finding is high correlation between two DNMT3L proteins even they located in two separate ends without any direct interaction. Moreover, we detected that 3A:3L-AB and 3A:3L-CD chains show different correlations which is not expected. Additionally, RMSF and interaction profile results support this finding. As we know tetramer structure is symmetric, we expected more similar interaction profiles. Simulation time could be insufficient in order to reach symmetry in structure.

As salt bridge interactions play an important role on selectivity in protein DNA interactions, high number of salt bridges in tetramer might be related with increase in specific binding of DNMT3A to DNA but further analyses are required. Since CpGpX sequence differs in dimer and tetramer, we observed change in interaction profile such as formation of nucleotide specific hydrophobic interactions with C5 and C7 atoms of thymine.

As future work, interaction in both dimer and tetramer will be analyzed in detail to elucidate effect of specific interactions on working mechanism of DNMT3A. To understand the effect of third nucleotide in CpGpX sequence, we are planning to run simulations of dimer with CpGpT DNA sequence and tetramer with CpGpC DNA sequence. Moreover, we are aiming to mutate a residue in the hydrophilic interface to detect its effect, specifically R882H mutation related with acute myeloid leukemia.

APPENDIX A: Residue Renumbering of DNMT3A

Table A.1: Residue Numbering of DNMT3A

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
630	8	204	487	K
631	9	205	488	R
632	10	206	489	K
633	11	207	490	P
634	12	208	491	I
635	13	209	492	R
636	14	210	493	V
637	15	211	494	L
638	16	212	495	S
639	17	213	496	L
640	18	214	497	F
641	19	215	498	D
642	20	216	499	G
643	21	217	500	I
644	22	218	501	A
645	23	219	502	T
646	24	220	503	G
647	25	221	504	L
648	26	222	505	L
649	27	223	506	V
650	28	224	507	L

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
651	29	225	508	K
652	30	226	509	D
653	31	227	510	L
654	32	228	511	G
655	33	229	512	I
656	34	230	513	Q
657	35	231	514	V
658	36	232	515	D
659	37	233	516	R
660	38	234	517	Y
661	39	235	518	I
662	40	236	519	A
663	41	237	520	S
664	42	238	521	E
665	43	239	522	V
666	44	240	523	C
667	45	241	524	E
668	46	242	525	D
669	47	243	526	S
670	48	244	527	I
671	49	245	528	T
672	50	246	529	V
673	51	247	530	G
674	52	248	531	M
675	53	249	532	V
676	54	250	533	R
677	55	251	534	H

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
678	56	252	535	Q
679	57	253	536	G
680	58	254	537	K
681	59	255	538	I
682	65	256	539	M
683	66	257	540	Y
684	67	258	541	V
685	68	259	542	G
686	69	260	543	D
687	70	261	544	V
688	71	262	545	R
689	72	263	546	S
690	73	264	547	V
691	74	265	548	T
692	75	266	549	Q
693	76	267	550	K
694	77	268	551	H
695	78	269	552	I
696	79	270	553	Q
697	80	271	554	E
698	81	272	555	W
699	82	273	556	G
700	83	274	557	P
701	84	275	558	F
702	85	276	559	D
703	86	277	560	L
704	87	278	561	V

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
705	88	279	562	I
706	89	280	563	G
707	90	281	564	G
708	91	282	565	S
709	92	283	566	P
710	93	284	567	C
711	94	285	568	N
712	95	286	569	D
713	96	287	570	L
714	97	288	571	S
715	98	289	572	I
716	99	290	573	V
717	100	291	574	N
718	101	292	575	P
719	102	293	576	A
720	103	294	577	R
721	104	295	578	K
722	105	296	579	G
723	106	297	580	L
724	107	298	581	Y
725	108	299	582	E
726	109	300	583	G
727	110	301	584	T
728	111	302	585	G
729	112	303	586	R
730	113	304	587	L
731	114	305	588	F

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
732	115	306	589	F
733	116	307	590	E
734	117	308	591	F
735	118	309	592	Y
736	119	310	593	R
737	120	311	594	L
738	121	312	595	L
739	122	313	596	H
740	123	314	597	D
741	124	315	598	A
742	125	316	599	R
743	126	317	600	P
744	127	318	601	K
745	128	319	602	E
746	129	320	603	G
747	130	321	604	D
748	131	322	605	D
749	132	323	606	R
750	133	324	607	P
751	134	325	608	F
752	135	326	609	F
753	136	327	610	W
754	137	328	611	L
755	138	329	612	F
756	139	330	613	E
757	140	331	614	N
758	141	332	615	V

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
759	142	333	616	V
760	143	334	617	A
761	144	335	618	M
762	145	336	619	G
763	146	337	620	V
764	147	338	621	S
765	148	339	622	D
766	149	340	623	K
767	150	341	624	R
768	151	342	625	D
769	152	343	626	I
770	153	344	627	S
771	154	345	628	R
772	155	346	629	F
773	156	347	630	L
774	157	348	631	E
775	158	349	632	S
776	159	350	633	N
777	160	351	634	P
778	161	352	635	V
779	162	353	636	M
780	163	354	637	I
781	164	355	638	D
782	165	356	639	A
783	166	357	640	K
784	167	358	641	E
785	168	359	642	V

Continued on Next Page…

Table A.1 – Continued

PDB resi	**MD resi**	**PCA resi A**	**PCA resi D**	**Residue name**
786	169	360	643	S
787	170	361	644	A
788	181	362	645	A
789	182	363	646	H
790	183	364	647	R
791	184	365	648	A
792	185	366	649	R
793	186	367	650	Y
794	187	368	651	F
795	198	369	652	W
796	199	370	653	G
797	200	371	654	N
798	201	372	655	L
799	202	373	656	P
800	203	374	657	G
801	204	375	658	M
802	205	376	659	N
803	206	377	660	R
804	207	378	661	P
805	208	379	662	L
806	211	380	663	A
807	212	381	664	S
808	213	382	665	T
809	214	383	666	V
810	215	384	667	N
811	216	385	668	D
812	217	386	669	K

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
813	218	387	670	L
814	219	388	671	E
815	220	389	672	L
816	221	390	673	Q
817	222	391	674	E
818	223	392	675	C
819	224	393	676	L
820	225	394	677	E
821	226	395	678	H
822	227	396	679	G
823	228	397	680	R
824	229	398	681	I
825	230	399	682	A
826	231	400	683	K
827	261	401	684	F
828	262	402	685	S
829	263	403	686	K
830	264	404	687	V
831	265	405	688	R
832	266	406	689	T
833	267	407	690	I
834	268	408	691	T
835	269	409	692	T
836	270	410	693	R
837	271	411	694	S
838	272	412	695	N
839	273	413	696	S

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
840	274	414	697	I
841	275	415	698	K
842	276	416	699	Q
843	277	417	700	G
844	278	418	701	K
845	280	419	702	D
846	281	420	703	Q
847	282	421	704	H
848	283	422	705	F
849	284	423	706	P
850	285	424	707	V
851	286	425	708	F
852	287	426	709	M
853	288	427	710	N
854	289	428	711	E
855	290	429	712	K
856	291	430	713	E
857	292	431	714	D
858	293	432	715	I
859	294	433	716	L
860	295	434	717	W
861	296	435	718	C
862	297	436	719	T
863	298	437	720	E
864	299	438	721	M
865	300	439	722	E
866	301	440	723	R

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
867	302	441	724	V
868	303	442	725	F
869	304	443	726	G
870	305	444	727	F
871	306	445	728	P
872	307	446	729	V
873	308	447	730	H
874	309	448	731	Y
875	310	449	732	T
876	311	450	733	D
877	312	451	734	V
878	313	452	735	S
879	314	453	736	N
880	315	454	737	M
881	316	455	738	S
882	317	456	739	R
883	318	457	740	L
884	319	458	741	A
885	320	459	742	R
886	321	460	743	Q
887	322	461	744	R
888	323	462	745	L
889	324	463	746	L
890	325	464	747	G
891	326	465	748	R
892	327	466	749	S
893	328	467	750	W

Continued on Next Page…

Table A.1 – Continued

PDB resi	MD resi	PCA resi A	PCA resi D	Residue name
894	329	468	751	S
895	330	469	752	V
896	331	470	753	P
897	332	471	754	V
898	333	472	755	I
899	334	473	756	R
900	335	474	757	H
901	336	475	758	L
902	337	476	759	F
903	338	477	760	A
904	345	478	761	P
905	346	479	762	L
906	347	480	763	K
907	348	481	764	E
908	349	482	765	Y
909	350	483	766	F
910	351	484	767	A
911	352	485	768	C
912	353	486	769	V

APPENDIX B: Residue Renumbering of DNMT3B

Table B.1: Residue Numbering of DNMT3B

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
178	1178	1	770	M
179	1179	2	771	F
180	1180	3	772	E
181	1181	4	773	T
182	1182	5	774	V
183	1183	6	775	P
184	1184	7	776	V
185	1185	8	777	W
186	1186	9	778	R
187	1187	10	779	R
188	1188	11	780	Q
189	1189	12	781	P
190	1190	13	782	V
191	1191	14	783	R
192	1192	15	784	V
193	1193	16	785	L
194	1194	17	786	S
195	1195	18	787	L
196	1196	19	788	F
197	1197	20	789	E
198	1198	21	790	D

Continued on Next Page…

Table B.1 – Continued

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
199	1199	22	791	I
200	1200	23	792	K
201	1201	24	793	K
202	1202	25	794	E
203	1203	26	795	L
204	1204	27	796	T
205	1205	28	797	S
206	1206	29	798	L
207	1207	30	799	G
208	1208	31	800	F
209	1209	32	801	L
210	1210	33	802	E
211	1211	34	803	S
212	1212	35	804	G
213	1213	36	805	S
214	1214	37	806	D
215	1215	38	807	P
216	1216	39	808	G
217	1217	40	809	Q
218	1218	41	810	L
219	1219	42	811	K
220	1220	43	812	H
221	1221	44	813	V
222	1222	45	814	V
223	1223	46	815	D
224	1224	47	816	V
225	1225	48	817	T

Continued on Next Page…

Table B.1 – Continued

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
226	1226	49	818	D
227	1227	50	819	T
228	1228	51	820	V
229	1229	52	821	R
230	1230	53	822	K
231	1231	54	823	D
232	1232	55	824	V
233	1233	56	825	E
234	1234	57	826	E
235	1235	58	827	W
236	1236	59	828	G
237	1237	60	829	P
238	1238	61	830	F
239	1239	62	831	D
240	1240	63	832	L
241	1241	64	833	V
242	1242	65	834	Y
243	1243	66	835	G
244	1244	67	836	A
245	1245	68	837	T
246	1246	69	838	P
247	1247	70	839	P
248	1248	71	840	L
249	1249	72	841	G
250	1250	73	842	H
251	1251	74	843	T
252	1252	75	844	C

Continued on Next Page…

Table B.1 – Continued

PDB resi	**MD resi**	**PCA resi B**	**PCA resi C**	**Residue name**
253	1253	76	845	D
254	1254	77	846	R
255	1255	78	847	P
256	1256	79	848	P
257	1257	80	849	S
258	1258	81	850	W
259	1259	82	851	Y
260	1260	83	852	L
261	1261	84	853	F
262	1262	85	854	Q
263	1263	86	855	F
264	1264	87	856	H
265	1265	88	857	R
266	1266	89	858	L
267	1267	90	859	L
268	1268	91	860	Q
269	1269	92	861	Y
270	1270	93	862	A
271	1271	94	863	R
272	1272	95	864	P
273	1273	96	865	K
274	1274	97	866	P
275	1275	98	867	G
276	1276	99	868	S
277	1277	100	869	P
278	1278	101	870	R
279	1279	102	871	P

Continued on Next Page…

Table B.1 – Continued

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
280	1280	103	872	F
281	1281	104	873	F
282	1282	105	874	W
283	1283	106	875	M
284	1284	107	876	F
285	1285	108	877	V
286	1286	109	878	D
287	1287	110	879	N
288	1288	111	880	L
289	1289	112	881	V
290	1290	113	882	L
291	1291	114	883	N
292	1292	115	884	K
293	1293	116	885	E
294	1294	117	886	D
295	1295	118	887	L
296	1296	119	888	D
297	1297	120	889	V
298	1298	121	890	A
299	1299	122	891	S
300	1300	123	892	R
301	1301	124	893	F
302	1302	125	894	L
303	1303	126	895	E
304	1304	127	896	M
305	1305	128	897	E
306	1306	129	898	P

Continued on Next Page…

Table B.1 – Continued

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
307	1307	130	899	V
308	1308	131	900	T
309	1309	132	901	I
310	1310	133	902	P
311	1311	134	903	D
312	1312	135	904	V
313	1313	136	905	H
314	1314	137	906	G
315	1315	138	907	G
316	1316	139	908	S
317	1317	140	909	L
318	1318	141	910	Q
319	1319	142	911	N
320	1320	143	912	A
321	1321	144	913	V
322	1322	145	914	R
323	1323	146	915	V
324	1324	147	916	W
325	1325	148	917	S
326	1326	149	918	N
327	1327	150	919	I
328	1328	151	920	P
329	1329	152	921	A
330	1330	153	922	I
331	1331	154	923	R
332	1332	155	924	S
333	1333	156	925	R

Continued on Next Page…

Table B.1 – Continued

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
334	1334	157	926	H
335	1335	158	927	W
336	1336	159	928	A
337	1337	160	929	L
338	1338	161	930	V
339	1339	162	931	S
340	1340	163	932	E
341	1341	164	933	E
342	1342	165	934	E
343	1343	166	935	L
344	1344	167	936	S
345	1345	168	937	L
346	1346	169	938	L
347	1347	170	939	A
348	1348	171	940	Q
349	1349	172	941	N
350	1350	173	942	K
351	1351	174	943	Q
352	1352	175	944	S
353	1353	176	945	S
354	1354	177	946	K
355	1355	178	947	L
356	1356	179	948	A
357	1357	180	949	A
358	1358	181	950	K
359	1359	182	951	W
360	1360	183	952	P

Continued on Next Page…

Table B.1 – Continued

PDB resi	MD resi	PCA resi B	PCA resi C	Residue name
361	1361	184	953	T
362	1362	185	954	K
363	1363	186	955	L
364	1364	187	956	V
365	1365	188	957	K
366	1366	189	958	N
367	1367	190	959	C
368	1368	191	960	F
369	1369	192	961	L
370	1370	193	962	P
371	1371	194	963	L
372	1372	195	964	R
373	1373	196	965	E
374	1374	197	966	Y
375	1375	198	967	F
376	1376	199	968	K
377	1377	200	969	Y
378	1378	201	970	F
379	1379	202	971	S
380	1380	203	972	T

APPENDIX C: RMSF graphs of DNMT3A and DNMT3L separated by chains

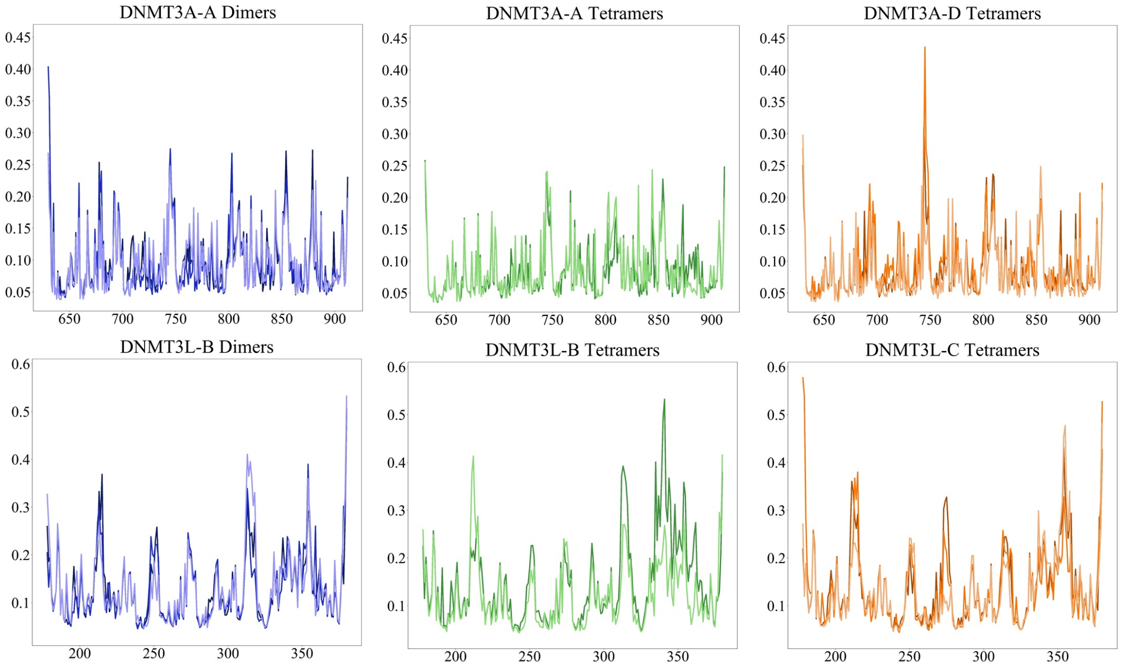


Figure C.1 RMSF graphs of DNMT3A and DNMT3L separated by chains

APPENDIX D: CCor maps of each replica simulations

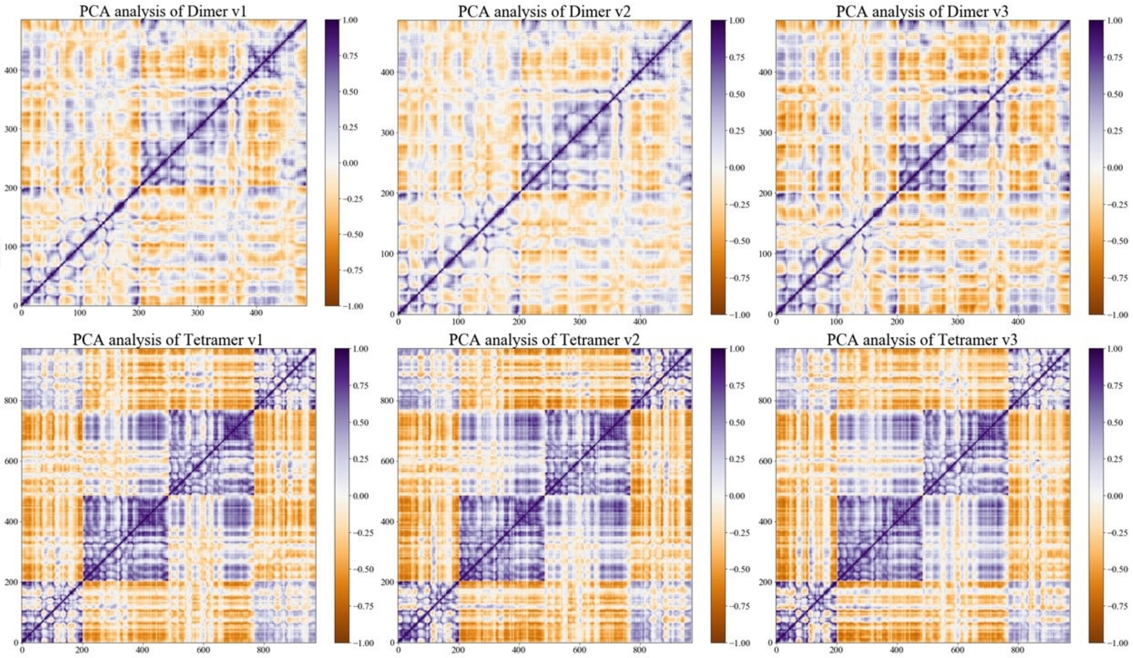
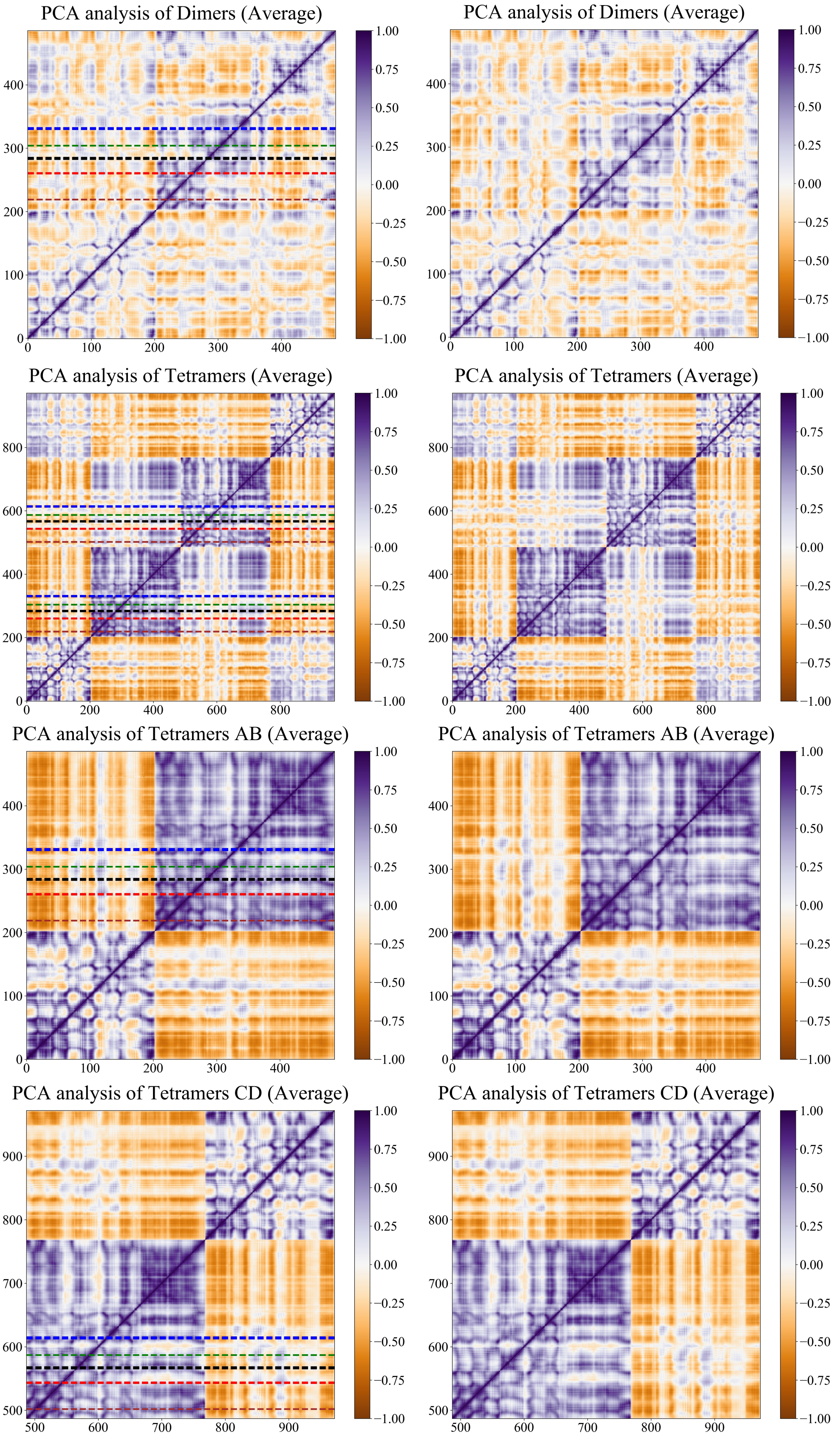


Figure D.1 CCor maps of each replica simulation


PCA analysis of Dimers (Average)

PCA analysis of Dimers (Average)

PCA analysis of Tetramers (Average)

PCA analysis of Tetramers (Average)

PCA analysis of Tetramers AB (Average)

PCA analysis of Tetramers AB (Average)

PCA analysis of Tetramers CD (Average)

PCA analysis of Tetramers CD (Average)

Figure D.2 CCor maps of average simulations with motifs locations. All motifs (I, III, IV, V and VI are represented with brown, red, black, green and blue, respectively.

REFERENCES

Abraham, M. J. et al. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 1–2. doi: 10.1016/j.softx.2015.06.001.

Alder, B. J. and Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. The Journal of Chemical Physics, 31(2). doi: 10.1063/1.1730376.

Amadei, A., Linssen, A. B. M. and Berendsen, H. J. C. (1993). “Essential dynamics of proteins,” Proteins: Structure, Function, and Bioinformatics, 17(4). doi: 10.1002/prot.340170408.

Anteneh, H., Fang, J. and Song, J. (2020). Structural basis for impairment of DNA methylation by the DNMT3A R882H mutation. Nature Communications, 11(1), p. 2294. doi: 10.1038/s41467-020-16213-9.

Ashapkin, V. v, Kutueva, L. I. and Vanyushin, B. F. (2016). Dnmt2 is the most evolutionary conserved and enigmatic cytosine DNA methyltransferase in eukaryotes. Russian Journal of Genetics. doi: 10.1134/S1022795416030029.

Au, P. Y. B., Eaton, A. and Dyment, D. A. (2020). Chapter 23 - Genetic mechanisms of neurodevelopmental disorders. in Gallagher, A. et al. (eds) Handbook of Clinical Neurology. Elsevier, pp. 307–326. doi: https://doi.org/10.1016/B978-0-444-64150-2.00024-1.

Bakan, A. et al. (2014). Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics, 30(18). doi: 10.1093/bioinformatics/btu336.

Bakan, A., Meireles, L. M. and Bahar, I. (2011). ProDy: Protein dynamics inferred from theory and experiments. Bioinformatics, 27(11). doi: 10.1093/bioinformatics/btr168.

Batista, P. R. et al. (2006). Molecular dynamics simulations applied to the study of subtypes of HIV-1 protease common to Brazil, Africa, and Asia. in Cell Biochemistry and Biophysics. doi: 10.1385/CBB:44:3:395.

Casadesús, J. and Low, D. (2006). Epigenetic Gene Regulation in the Bacterial World. Microbiology and Molecular Biology Reviews, 70(3). doi: 10.1128/mmbr.00016-06.

Chédin, F. (2011). The DNMT3 family of mammalian de novo DNA methyltransferases. Progress in Molecular Biology and Translational Science. doi: 10.1016/B978-0-12-387685-0.00007-X.

Chen, Z. X. and Riggs, A. D. (2011). DNA methylation and demethylation in mammal. Journal of Biological Chemistry. doi: 10.1074/jbc.R110.205286.

Dalfrà, M. G. et al. (2020). "Genetics and Epigenetics: New Insight on Gestational Diabetes Mellitus," Frontiers in endocrinology, 11, p. 602477. doi: 10.3389/fendo.2020.602477.

Dogan, L. (2020). Molecular Determinant of DNA Methylation Mechanisms in Mammals. Master's Thesis. Dokuz Eylul University.

Dominguez, C., Boelens, R. and Bonvin, A. M. J. J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society, 125(7). doi: 10.1021/ja026939x.

Gowher, H. and Jeltsch, A. (2018). Mammalian DNA methyltransferases: New discoveries and open questions. Biochemical Society Transactions. doi: 10.1042/BST20170574.

Groot, B. L. de et al. (2001). Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. Journal of Molecular Biology, 309(1). doi: 10.1006/jmbi.2001.4655.

Guo, X. et al. (2015). Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. Nature, 517(7536), pp. 640–644. doi: 10.1038/nature13899.

Holz-Schietinger, C. et al. (2011). Oligomerization of DNMT3A controls the mechanism of de novo DNA methylation. Journal of Biological Chemistry, 286(48). doi: 10.1074/jbc.M111.284687.

Isaak, D. J. et al. (2018). Principal components of thermal regimes in mountain river networks. Hydrology and Earth System Sciences, 22(12). doi: 10.5194/hess-22-6225-2018.

Ivani, I. et al. (2015). Parmbsc1: A refined force field for DNA simulations. Nature Methods, 13(1). doi: 10.1038/nmeth.3658.

Jeltsch, A. (2006). On the enzymatic properties of dnmt1: Specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme.

Epigenetics, 1(2). doi: 10.4161/epi.1.2.2767.

Jia, D. et al. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature, 449(7159). doi: 10.1038/nature06146.

Karaca, E. et al. (2010). Building macromolecular assemblies by information-driven docking: Introducing the HADDOCK multi-body docking server. Mol. Cell. Proteomics, 9, pp. 1784-1794. doi: 10.1074/mcp.M000051-MCP201.

Kulis, M. and Esteller, M. (2010). DNA Methylation and Cancer. Advances in Genetics. doi: 10.1016/B978-0-12-380866-0.60002-2.

Law, J. A. and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics. doi: 10.1038/nrg2719.

Ooi, S. K. T. et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature, 448(7154). doi: 10.1038/nature05987.

Páll, S. et al. (2015). Tackling exascale software challenges in molecular dynamics simulations with GROMACS. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi: 10.1007/978-3-319-15976-8_1.

Pearlman, D. A. et al. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Computer Physics Communications, 91(1–3). doi: 10.1016/0010-4655(95)00041-D.

Phillips, J. C. et al. (2005). Scalable molecular dynamics with NAMD. Journal of Computational Chemistry. doi: 10.1002/jcc.20289.

Rahman, A. (1964). Correlations in the motion of atoms in liquid argon. Physical Review, 136(2A). doi: 10.1103/PhysRev.136.A405.

Robertson, K. D. (2005). DNA methylation and human disease. Nature Reviews Genetics. doi: 10.1038/nrg1655.

Sánchez-Romero, M. A. and Casadesús, J. (2020). The bacterial epigenome. Nature Reviews Microbiology. doi: 10.1038/s41579-019-0286-2.

Silva, A. W. S. da and Vranken, W. F. (2012). ACPYPE - AnteChamber PYthon Parser interfacE. BMC Research Notes, 5. doi: 10.1186/1756-0500-5-367.

Sittel, F., Jain, A. and Stock, G. (2014). Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. Journal of Chemical Physics, 141(1). doi: 10.1063/1.4885338.

Takeshita, K. et al. (2011). Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). Proceedings of the National Academy of Sciences, 108(22), pp. 9055-9059. doi: 10.1073/PNAS.1019629108.

Tost, J. (2010). DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. Molecular Biotechnology. doi: 10.1007/s12033-009-9216-2.

Vries, S. J. de, Dijk, M. van and Bonvin, A. M. J. J. (2010). The HADDOCK web server for data-driven biomolecular docking. Nature Protocols, 5(5). doi: 10.1038/nprot.2010.32.

Waddington, C. H. (2011). The epigenotype. Endeavour 1942; 1:18-20, Reprinted in International journal of epidemiology, 41(1).

Zeng, X. et al. (2020). Genome-Wide Characterization of Host Transcriptional and Epigenetic Alterations During HIV Infection of T Lymphocytes. Frontiers in Immunology, 11. doi: 10.3389/fimmu.2020.02131.

Zhang, Z. M. et al. (2018). Structural basis for DNMT3A-mediated de novo DNA methylation. Nature, 554(7692). doi: 10.1038/nature25477.

Zhu, J. K. (2009). Active DNA demethylation mediated by DNA glycosylases. Annual Review of Genetics. doi: 10.1146/annurev-genet-102108-134205.

Zundert, G. C. P. van et al. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. Journal of Molecular Biology, 428(4). doi: 10.1016/j.jmb.2015.09.014.